NLP4CMC 2015

2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media

Proceedings of the Workshop

September 29, 2015
University of Duisburg-Essen, Campus Essen

Organized by the special interest group
"Social Media / Computer-Mediated Communication"
of the German Society for Computational Linguistics & Language Technology (GSCL)
http://gscl.org/ak-ibk.html

Natural Language Processing for Computer-Mediated Communication / Social Media: a Challenging Task

Over the past decade, there has been a growing interest in collecting, processing and analyzing data from genres of social media and computer-mediated communication (CMC): As part of large corpora which have been automatically crawled from the web, CMC data are often regarded as an unloved “bycatch” which is difficult to handle with NLP tools that have been optimized for processing edited text; on the other hand, the existence of CMC data in web corpora is relevant for all research and application contexts which require data sets that represent the full diversity of genres and linguistic variation on the web. For corpus-based variational linguistics, CMC corpora are an important resource for closing the "CMC gap" both in corpora of contemporary written language and in corpora of spoken language: Since CMC and social media make up an important part of contemporary everyday communication, investigations into language change and linguistic variation need to be able to include CMC and social media data into their empirical analyses.

Nevertheless, the development of approaches and tools for processing the linguistic and structural peculiarities of CMC genres and for building CMC corpora is lacking behind the interest of dealing with these types of data in the field of language technology, corpus-based linguistics and web mining.

The goal of the NLP4CMC workshops is to provide a platform for the presentation of results and the discussion of ongoing work in adapting NLP tools for processing CMC data and in using NLP solutions for building and annotating social media corpora. The main focus of the workshops is on German data, but submissions on NLP approaches, annotation experiments and CMC corpus projects for data of other European languages are also welcome.

The 1st NLP4CMC workshop was held in September 2014 at KONVENS at the University of Hildesheim. This volume presents proceedings from the 2nd NLP4CMC workshop which has been held in September 2015 at the annual conference of the German Society for Language Technology and Computational Linguistics (GSCL) at the University of Duisburg-Essen.

We thank all colleagues who have contributed to the workshop with their talks and discussions.

Dortmund and Duisburg, September 2015
Michael Beißwenger
Torsten Zesch

Organizers:

Michael Beißwenger, Technische Universität Dortmund
Torsten Zesch, Universität Duisburg-Essen

Program Committee:

Sabine Bartsch (TU Darmstadt)
Thomas Bartz (TU Dortmund)
Thierry Chanier (Université Blaise Pascal, Clermont-Ferrand)
Isabella Chiari (Università "La Sapienza", Rome)
Stefanie Dipper (Ruhr-Universität Bochum)
Stefan Evert (Universität Erlangen)
Iris Hendrickx (Radboud University Nijmegen)
Verena Henrich (Universität Tübingen)
Tobias Horsmann (Universität Duisburg-Essen)
Lothar Lemnitzer (BBAW, Berlin)
Anke Lüdeling (Humboldt-Universität Berlin)
Harald Lüngen (IDS, Mannheim)
Preslav Nakov (Qatar QCRI)
Günter Neumann (DFKI, Saarbrücken)
Nelleke Oostdijk (Radboud University Nijmegen)
Ines Rehbein (Universität Potsdam)
Roman Schneider (IDS, Mannheim)
Egon W. Stemle (EURAC, Bozen)
Angelika Storrer (Universität Mannheim)
Kay-Michael Würzner (BBAW, Berlin)

Table of Contents

Workshop Program

29.9.2015

10:25–10:30	***Opening***
10:30–12:30	**Building and annotating CMC corpora**
10:30–11:00	*The DiDi Corpus of South Tyrolean CMC Data* Jennifer-Carmen Frey, Aivars Glaznieks and Egon Stemle
11:00–11:30	*Collection, Description, and Visualization of the German Reddit Corpus* Adrien Barbaresi
11:30–12:00	*Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus* Michael Beißwenger, Eric Ehrhardt, Andrea Horbach, Harald Lüngen, Diana Steffen and Angelika Storrer
12:00–12:30	*Building and Annotating a Corpus of German-Language Newsgroups* Jasmin Schröck and Harald Lüngen
12:30–14:00	***Lunch Break***
14:00–15:30	**Using NLP for analysing CMC corpora**
14:00–14:30	*Using discursive information to disentangle French language chat* Matthieu Riou, Nicolas Hernandez and Soufian Salim
14:30–15:00	*Text-based Geolocation of German Tweets* Johannes Gontrum and Tatjana Scheffler
15:00–15:30	*Modes of Communication in Social Media for Emergency Management* Sabine Gründer-Fahrer and Antje Schlaf
15:30–16:00	***Coffee Break***

29.9.2015 (continued)

16:00–17:00 Adapting the NLP toolkit to CMC genres

16:00–16:30 *Unsupervised Induction of Part-of-Speech Information for OOV Words in German Internet Forum Posts*
Jakob Prange, Stefan Thater and Andrea Horbach

16:30–17:00 *Bootstrapped Extraction of Index Terms from Normalized User-Generated Content*
Piroska Lendvai and Thierry Declerck

The DiDi Corpus of South Tyrolean CMC Data

Jennifer-Carmen Frey **Aivars Glaznieks** **Egon W. Stemle**
Institute for Specialised Communication and Multilingualism
European Academy Bozen/Bolzano
Viale Druso 1, 39100 Bolzano
{jennifer.frey,aivars.glaznieks,egon.stemle}@eurac.edu

Abstract

This paper presents the DiDi Corpus, a corpus of South Tyrolean Data of Computer-mediated Communication (CMC). The corpus comprises around 650,000 tokens from Facebook wall posts, comments on wall posts and private messages, as well as socio-demographic data of participants. All data was automatically annotated with language information (de, it, en and others), and manually normalised and anonymised. Furthermore, semi-automatic token level annotations include part-of-speech and CMC phenomena (e.g. emoticons, emojis, and iteration of graphemes and punctuation). The anonymised corpus without the private messages is freely available for researchers; the complete and anonymised corpus is available after signing a non-disclosure agreement.

1 Introduction

The aim of the DiDi project was to build a text corpus to document the current language use of German native speakers from the multilingual province of South Tyrol. We collected a CMC corpus consisting of Facebook wall posts, comments on wall posts and private messages, as well as socio-demographic data of the writers (cf. Glaznieks and Stemle (2014) for more details). Thus, the corpus combines socio-demographic data of the investigated Facebook users such as their language biography, internet usage habits, general parameters such as age, gender and education level with texts on their Facebook profiles. This facilitates sociolinguistic analyses, which has been a secondary objective of the project. To investigate the languages and language varieties used and relate them to socio-demographic parameters (particularly focussing on the users' age and internet experience), a number of annotations have been added to the data. We annotated the predominant language and language variety used, and special CMC phenomena, added a standard transcription to non-standard words, applied part-of-speech tagging, and lemmatisation, and anonymised the corpus considering ethical and legal privacy issues.

2 Corpus

The DiDi corpus has an overall size of around 650,000 tokens gathered from 136 South Tyrolean Facebook users who participated in the DiDi project. It consists of 11,102 Facebook wall posts, 6,507 comments on wall posts and 22,218 private messages of the participants. All messages were written by the participants during the year 2013 (section 3). Although downloading messages from friends and other Facebook users on participant-initiated posts was possible[1], this data must not be used for privacy issues. Consequently, all extraneous data of this kind was removed from the corpus except for the number of replies to messages, the language and the time stamp. This was deemed appropriate in terms of privacy and will likely be relevant for conversational and discourse-centred linguistic analyses of the data, i.e. the corpus does not allow for textual analyses of conversational interaction. As every participant could offer either the private messages and/or the texts on the wall, we were given access to 130 wall profiles and 56 private inbox profiles; 50 participants granted access to both types of data.

3 Data Collection

According to our project design, we aimed for 3 types of data from at least 100 South Tyrolean Facebook users (all with German as L1 and equally spread over various age groups):

1. user consent and privacy agreement,

2. Facebook texts (wall and/or private messages) from the year 2013,
3. socio-demographic data of users' language biography and internet usage habits.

To acquire these types of data from every user in the most structured (for researchers) and simple (for participants) way, we developed a web application that provided an interface to recruit users, inform them about the project's aims and methodologies, allow them to subscribe and explicitly agree to the usage of their data, fill in the online questionnaire and give them the possibility to grant us access to their Facebook data via the Facebook API. Our web application set-up enabled us to download and merge all the necessary data while saving it into our internal document-oriented NoSQL database[2]. See (Frey et al., 2014) for an in-depth description of the process and its technical details.

The user recruitment was mainly accomplished by circulating the web application's URL using chain sampling within Facebook. Additionally, the link was posted in various South Tyrolean Facebook groups and other social media communities to draw further attention to the project. In order to reach more potential users, particularly in older age groups, was targeted Facebook advertising in which the link and some text were posted directly to the walls of South Tyrolean users matching the specific user group.

4 Corpus Annotation

All subsequently mentioned annotation tasks were carried out by three annotators according to a set of annotation guidelines. The tasks were carried out within our processing pipeline: annotators use their favourite spreadsheet program, e.g. Microsoft Excel or LibreOffice, and the pipeline converts between the spreadsheet representation and the structured representation of the database (and vice versa). The spreadsheet representation is a vertical file with one token per line, and individual (blocks of) columns represent annotation layers, which have to be edited according to our annotation guidelines[3]. For example, to merge multiple tokens into one (because they were misspelled) edit the appropriate column and write the proper normalisation in one field and the special token `____' in the column's next line(s).

Problems of individual tasks were compared, and differences were discussed until a consensus was reached. If necessary, the annotation guidelines were updated and previous annotation work (sometimes) redone.

5 Corpus Processing

After the original data provided by the Facebook API and the data from the user questionnaire were downloaded and stored, the data went through various natural language processing (NLP) and annotation steps.

NLP of social media texts is still an unsolved problem. Social media corpora contain many short and noisy texts and the content is usually strongly contextualised; therefore, the corpora differ from each other in many ways and are very domain dependent. NLP algorithms are traditionally trained on news-based corpora and these differences affect their performance. (See, for example, Preotiuc-Pietro et al. (2012) and Baldwin (2012).)

For social media texts from South Tyrol, i.e. for our domain, Glaznieks and Stemle (2014) analysed tokeniser and part-of-speech (POS) tagger performance on non-normalised Facebook data of dialect writers, and they evaluated the added value of various levels of normalisation on the source data. In this *pre-test*, they showed that the poor base-line performance of non-normalised data for this domain can be considerably improved by normalisation.

5.1 Tokenisation

After testing a number of tokenisers for social media texts (most of which are tuned to a specific domain), we decided to use the Python version of the Twitter tokenizer `ark-twokenze-py`[4] as it showed the best results with our non-public Facebook data and could already deal with most of the CMC-related difficulties such as emoticons, hyperlinks and individual abbreviations. However, some problems highlighted by the *pre-test* such as incorrect splitting of various time and date formats or words written with special characters to express users' individual, artistic style were still tokenised poorly and therefore manually corrected.

5.2 Normalisation and tagging of privacy issues

As a result of the *pre-test* we invested most of the manual annotation work in normalising the texts. Keeping in mind the project goals, we only normalised German texts of L1 German speakers in the corpus by using word-by-word transcriptions for each word that was not spelled in standard German be it because of diverge writing or the use of a dialect variety (see Ruef and Ueberwasser (2013) for more details). We used Duden online[5] as a reference to define the target standard spelling of words.

In this annotation task, compound words that were not written as one token in the original were merged together, and words that should have been split into multiple tokens according to standard German were inserted as separate tokens.

While adding this normalisation information, privacy issues were also indicated for later referencing and processing (section 5.5).

5.3 POS-tagging and lemmatisation

`TreeTagger`[6] for German (Schmid, 1995) and the Stuttgart-Tübingen-TagSet (`STTS`)[7] was used for POS-tagging and lemmatisation. The initial results on the previously normalised data were later improved by additional annotation work as the annotations allowed to assign fixed POS tags to previously error-prone tokens (cf. sections 5.4, 5.5, 5.6).

5.4 Handling of dialect lexemes and out-of-vocabulary tokens

During the manual normalisation (section 5.2), the annotators already indicated dialect words that had no equivalent and therefore no spelling in standard German. This information was then used to compile a list of dialect lexemes that are unique for South Tyrolean German. While the list was also linguistically interesting, the primary goal was to unify different spellings of the lexemes and provide an additional lexicon containing part-of-speech information for the POS-tagging and for other subsequent automatic procedures (e.g. classification of used variety). Furthermore, most of the common out-of-vocabulary (OOV) words were listed. Dialect lexemes, foreign language insertions, emoticons and abbreviations that occurred in large amounts were identified, classified as one of those categories and automatically annotated and processed afterwards. This also helped to further improve the POS-tagging of the corpus as in most cases fixed POS tags could be assigned to them (e.g. foreign language insertions received the POS tag *FM* of the `STTS`).

5.5 Anonymisation

The previously indicated privacy issues (section 5.2) were categorised as follows: personal names, group names, geographical names and adjectival references, institution names, hyperlinks, e-mail addresses, phone numbers, and a miscellaneous category containing other private information like numbers of bank accounts, and servers, postal codes, etc.

The original entities were then substituted with information-based type identifiers (PersNE, GruppeNE, GeoNE, GeoADJA, InstNE, link, mail, tel, XXX) that showed the anonymised category (cf. Panckhurst (2013)), keeping any inflectional affixes (e.g. “PersNEs Privatsphäreneinstellung”) and word formations (e.g. “InstNE-Zeltlager”). In addition, the categories determined the POS information on the POS layer (e.g. the POS tag *NE* was assigned to all tokens anonymised as *PersNE* of the `STTS`). This method allows for better readability of data that often consists of several private details, whereas a pure overwriting often leads to nonsense texts. Furthermore, it facilitates automatic analyses on the used private entities since categories and POS tags are defined and reliable.

5.6 Linguistic annotation

A number of annotations such as the predominant language of a text, the variety of German and predefined CMC phenomena were created in order to answer the project’s research questions (cf. Glaznieks and Stemle (2014)). Only those CMC phenomena (e.g. Bartz et al. (2013), Schlobinski and Siever (2013)) that are clearly distinguishable from dialect writing were used. For this reason, we annotated phenomena such as emoticons, emojis, @mentions, CMC-specific acronyms and abbreviations, iterations of graphemes, punctuations and emoticons, asterisk expressions (action words), hyperlinks, and hashtags as CMC phenomena. Other features that either originate from emulation of spoken language (e.g. assimilation, clitics, etc.) or represent deviations from standard German orthography (e.g. case insensitivity) were not categorised as CMC phenomena in order to avoid confusion with particularities induced by writing in dialect. So far, such phenomena were only normalised with standard German equivalents but not annotated with a specific tag. More details on the annotation procedure and results are given in section 6.

6 Corpus Data

There are two types of data in the corpus: (a) sociodemographic data for each participant who also shared Facebook texts, and (b) texts with their linguistic annotations. Both are described in the following.

6.1 User meta data

The meta data of each user provides the necessary demographic data to carry out sociolinguistic analyses with the given language data.

6.1.1 Data gathered by questionnaire

Within the web application, we asked for sociodemographic data of the participants that was mainly centred on the users' language and internet usage biography. Additionally, some standard parameters such as gender, age, level of education and current employment were gathered. Table 1 shows some of the questionnaire data from the DiDi corpus.

Meta data	**Texts L1 German**	**Texts Total**
female	18,615	20,273
male	16,545	19,554
14-19 years	5,807	5,807
20-29 years	5,225	5,289
30-39 years	7,215	7,514
40-49 years	5,258	8,377
50-59 years	9,519	10,016
60 years and older	2,136	2,824
university degree	8,728	11,972
matura	13,893	14,781
no matura	7,362	7,362
no data	5,177	5,712
employed	12,083	15,410
self-employed	9,653	9,806
in education	10,302	10,373
unemployed	2,946	2,946
no data	176	1,292
total	35,160	39,827

Table 1: Distribution of texts by user groups

6.1.2 Data gathered via Facebook

As the Facebook API provides a number of data fields for participating users such as gender, language preferences, etc., we merged these fields in the corpus as far as ethical and moral appropriateness was given.

6.2 Language data

Whereas the data provided by the Facebook API for each user was not exhaustive and mainly not publishable due to privacy issues, the language data was already enriched by various annotations. Some of the most important in terms of linguistic analyses could be named as: timestamp for creation and editing of text, privacy settings for the text, reactions in form of likes, comments, and shares of that text, attachments such as photos, videos or hyperlinks, recipients of private messages and the application used for publishing the text (e.g. *Facebook for Android/iPhone*, *Twitter*).
Annotations that were added to the data for the purpose of the linguistic analysis can be split into text level annotations and token level annotations.

6.2.1 Text level annotations

A number of additional annotations where made to enrich the data gathered from Facebook.

Language The language was annotated on text level using `langid.py` language identification tool (Lui and Baldwin, 2012). The basic automatic annotation was refined manually by validating and correcting every language annotation that was

- under a threshold of 0.8 confidence [8]
- shorter than 30 characters [9]
- identified as neither German, English, Italian, French, Portuguese nor Spanish [10].

Table 2 shows the language classifications for the gathered texts.

Language	**Texts**
German	23,258
Italian	8,216
English	4,344
Spanish	197
French	60
Portuguese	50
other [11]	236
not classifiable [12]	3,466
total	39,827

Table 2: Outline of languages in DiDi corpus

Variety of German The normalisation of the corpus data showed that our participants, when writing in German, used the regional dialect in transcribed form to a large extent, however there were also texts that represented a standard-oriented variety of German. To analyse the differences and proportions of the used varieties we classified the German-tagged texts into 3 categories (see table 3 for details):

1. considered as South Tyrolean dialect,
2. considered as standard-oriented variety of German,
3. not classified.

For the categorisation, we used a rule-based approach based on previously compiled lists of untranslatable dialect lexemes and most common dialect-standard transcriptions as well as information on the quantity and quality of the token's divergence to the standard transcription (see table 3). All texts shorter then 30 characters were not classified for reasons of ambiguity. In addition, the subgroup of not classified texts represented text which included a mixture between standard and dialect varieties that did not allow for a valid classification to either category.

Variety	**Texts**
Standard-oriented German	10,227
South Tyrolean dialect	9,570
Not classified	3,461
Total German texts	23,258

Table 3: Outline of the varieties used in German texts

6.2.2 Token level annotations

The corpus contains the following token level annotations. *Original token:* tokenised automatically and manually corrected. *List of normalisation tokens:* standard transcription of misspelled or dialectal words. *Part-of-speech tag:* on normalised standard transcriptions. *Lemma:* on normalised standard transcriptions. *Foreign language insertions:* according to list of most common OOV tokens classified as foreign language vocabulary. *Untranslatable dialect lexemes:* according to list of dialect lexemes compiled during manual annotation and post-processing of OOV tokens. *CMC phenomena:* list of CMC phenomena rendered relevant for the linguistic analysis of the project's research questions:

- Emoticons
- Emojis
- @Mentions
- Most common CMC acronyms and abbreviations (*cmq*, *thx*, *glg*, ...)
- Iteration of graphemes, punctuation or emoticons
- Asterisk expressions
- Hyperlinks
- Hashtags

7 Conclusion and Future Work

The DiDi corpus provides an insight into private, or at least non-public, informal written language use of people in a multilingual environment. The corpus combines the peculiarities of computer-mediated communication with the sociodemographic data of the writers in question and allows for a detailed investigation of current communicational strategies and language usage. A profound evaluation of the DiDi corpus is needed to ensure the quality of further investigations. Nevertheless, the corpus already offers a vast range of research opportunities not only for linguists interested in CMC, multilingual language use, the use of regional varieties, etc., but also for researchers interested in the technical processing of such textual content.

Further information regarding downloading the corpus data and querying it via ANNIS[13] is available at `http://www.eurac.edu/didi`.

Acknowledgements

The project was financed by the Autonome Provinz Bozen Südtirol, Abteilung Bildungsförderung, Universität und Forschung, Landesgesetz vom 13. Dezember 2006, Nr. 14 „Forschung und Innovation"

[1] Replies to comments, for example, are interwoven into the original content in such a way that it is impossible *not* to download them.

[2] `http://mongodb.org`

[3] See `http://www.eurac.edu/didi` for details.

[4] `https://github.com/myleott/ark-twokenize-py`

[5] `http://www.duden.de`

[6] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[7] `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`

[8] According to the confidence value stated by langid.py

[9] We defined this as a minimal length as shorter texts were too ambiguous to obtain a reliable classification with automatic tools.

[10] These languages were the most common results from the language identification tool, are partly taught in schools, or have been stated as native languages by participants and were therefore expected to show up in the corpus, whereas

References

[Baldwin2012] Timothy Baldwin. 2012. Social media: Friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 58–59, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

[Bartz et al.2013] Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2013. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28(1):157–198.

[Frey et al.2014] Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2014. Collecting language data of non-public social media profiles. In Gertrud Faaß and Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 11–15, Hildesheim, Germany, October. Universitatsverlag Hildesheim, Germany.

[Glaznieks and Stemle2014] Aivars Glaznieks and Egon W. Stemle. 2014. Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *JLCL*, 29(2):31–57.

[Lui and Baldwin2012] Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

[Panckhurst2013] Rachel Panckhurst. 2013. A Large SMS Corpus in French: From Design and Collation to Anonymisation, Transcoding and Analysis. *Procedia - Social and Behavioral Sciences*, 95:96 – 104.

[Preotiuc-Pietro et al.2012] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on real-time analysis and mining of social streams*.

[Ruef and Ueberwasser2013] Beni Ruef and Simone Ueberwasser. 2013. The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. *Non-standard Data Sources in Corpus-based Research*, pages 61–68.

[Schlobinski and Siever2013] Peter Schlobinski and Torsten Siever. 2013. Microblogs global: Deutsch. In Torsten Siever and Peter Schlobinski, editors, *Microblogs global. Eine internationale Studie zu Twitter & Co. aus der Perspektive von zehn Sprachen und elf Ländern*, pages 41–74. Peter Lang.

[Schmid1995] Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

other languages such as Turkish, Danish or Chinese where less probable and therefore manually checked.

[11] The group *other* was used for all manually classified texts that did not belong to any of the previously stated languages.

[12] Category for texts containing solely non-verbal graphs (e.g. emoticons, links, etc.) or ambiguous or multi-language expressions that can not be classified as a single language (e.g. interjections, international greetings or other internationally used words as "super" or "bravo")

[13] `http://annis-tools.org/`

Collection, Description, and Visualization of the German Reddit Corpus

Adrien Barbaresi
Austrian Academy of Sciences (ÖAW)
Berlin-Brandenburg Academy of Sciences (BBAW)
adrien.barbaresi@oeaw.ac.at

Abstract

Reddit is a major social bookmarking and microblogging platform. An extensive dataset of Reddit comments has recently been made publicly available. I use a two-tiered filter to single out comments in German in order to build a linguistic corpus which is then tokenized and annotated. This article offers first insights of both nature and quality of data at the lexical level. Additionally, a visualization makes it possible to grasp the possible geographical distribution of German users of the platform.

1 Introduction

One of the main issues when dealing with web corpora, be it general-purpose corpora or specific ones, consists in the discovery of relevant web documents for linguistic studies. There are for example few projects dealing with computer-mediated communication in German, and it is quite rare to find ready-made resources. The DeRiK project for instance features ongoing work with the purpose to build a reference corpus dedicated to computer-mediated communication (Beißwenger et al., 2013). Previous work towards the constitution of a German blog corpus under CC license implied a significant effort (Barbaresi and Würzner, 2014).

In this respect, it has been particularly surprising to hear from the release of a complete dataset of comments published on Reddit, a major social network. This article describes the steps taken in order to get a first glimpse of German data in the corpus as well as to describe what makes CMC-data in general and Reddit data in particular so different.

One hope is that the Reddit corpus can be used to find relevant examples of previously undocumented language uses for lexicography and dictionary building projects, e.g. the DWDS lexicography project (Geyken, 2007), and/or to test linguistic annotation chains for robustness.

2 Description of the dataset

2.1 Reddit

Reddit is a social bookmarking and microblogging platform owned by the American mass media company Condé Nast. It ranks at first place worldwide in the news category according to the site metrics aggregator Alexa[1], which makes it a typical Internet phenomenon. The short description of the website according to Alexa is as follows: "User-generated news links. Votes promote stories to the front page." Indeed, the entries are organized into areas of interest called "reddits" or "subreddits", which are curated by the users ("redditors") themselves. Since the moderation processes are mature, and since the channels (or subreddits) have to be hand-picked, they ensure a certain stability. From a linguistic point of view, one may say that users account for the linguistic homogeneity if not relevance of their channel.

There is an API for Reddit, allowing automated retrieval of comments. However, search depth is limited: it is often not possible to go back in time further than the 500th oldest post, which severely restricts the number of links one may crawl (Barbaresi, 2013). Continuous crawling is then necessary in order to gather all the possible comments on all the subreddits.

2.2 Original release

The work described in the article directly follows from the recent release of the "Reddit comment corpus": Reddit user *Stuck_In_The_Matrix* (Jason Baumgartner) made the dataset publicly available on the platform archive.org[2] at the beginning of July 2015. In its original release statement on Reddit, Baumgartner claims to have gathered every publicly available Reddit comment, which amounts

[1] http://www.alexa.com/topsites/category/Top/News

[2] https://archive.org/details/2015_reddit_comments_corpus

to 1.65 billion JSON objects.[3] 350,000 comments out of 1.65 billion were unavailable due to Reddit API issues.

While its compiler chose to name it a “corpus”, the whole could rather be called a dataset. In fact, apart from ensuring the most complete collection process possible, no specific steps were taken to allow for a control of the contents in the sense of the linguistic tradition (Barbaresi, 2015).

2.3 Filtering steps

I use a two-tiered filter in order to deliver a hopefully well-balanced performance between speed and accuracy. The combined strategy proved efficient in preliminary tests as well as in previous studies (Lui and Baldwin, 2014). The first filter consists in a dictionary-based approach taking benefit from spell-checking algorithms. It discriminates between comments using thresholds expressed as a percentage of tokens which do not pass the spell check. The second filter is a full-fledged language detection software, which outputs the most probable language according to its model.

First, spell checking algorithms benefit from years of optimization concerning both speed and accuracy. The library used, *enchant*, allows the use of a variety of spell-checking backends, like aspell, hunspell or ispell, with one or several locales.[4] English being the most prominent language on Reddit, each token is tested for errors in both English and German. A comment which induces a relatively high amount of errors for English (more than 30%) but a relatively low one for German (less than 70%) is considered to be interesting enough to proceed to the second step. In other studies (Barbaresi, 2013), I have used a threshold of 0.5; while I did not witness significant changes on Reddit data, I still chose a more defensive setting in order to ensure corpus relevance.

Second, a language identification tool is used to maximize the precision of the language recognition. `langid.py` (Lui and Baldwin, 2012) is open-source[5], it incorporates a model which has been pre-trained on a variety of web documents (Clue Web and Wikipedia inter alia). It has already been used to classify social media text on a large scale (Baldwin et al., 2013) and it is fast enough to be able to classify data from the order of magnitude of the Reddit comments.

The filtering described here is reproducible and can be attempted using other parameters, instructions and code to do so are available.[6]

3 Analysis of the corpus

3.1 Linguistic features

The corpus has been tokenized by WASTE (Jurish and Würzner, 2013) and lemmatized by MOOT (Jurish, 2003). It contains a total of 97,505 comments, 89,681 sentences, 566,362 tokens, and 3,352,472 characters. It is clear that Reddit is almost exclusively an English-speaking platform, however there are eminent German channels and due to the sheer size of the original dataset one could have expected a larger corpus. Maybe the precision of the filters could be lowered in favor of a better recall.

The mean token and sentence length (respectively 5.92 characters and 6.32 tokens) are in line with the expectations concerning computer-mediated communication, and it clearly anchors the corpus on this side of the spectrum. The relatively large vocabulary size in terms of types with 64,314 different forms (27% of which are hapax legomena) calls for further analyses. Qualitatively speaking, the ironic tone Reddit is known for could also prove to be interesting.

POS-tag	Frequency
NN	17.6%
NE	14.4%
ADV	8.6%
VVFIN	6.8%
PPER	5.9%
ADJD	4.7%
ART	4.3%
VAFIN	4.1%
XY	3.8%
ADJA	3.5%

Table 1: Most frequent POS-tags and relative frequency on token level, without spaces and punctuation

The breakdown into different part-of-speech tags shown in table 1 gives insights on the actual contents. The proportion of a number of tags from the STTS tagset is in line with other general or CMC-corpora (Barbaresi, 2014), however the number of tokens tagged as proper nouns (NE) is particularly

[3] https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

[4] http://www.abisource.com/projects/enchant/

[5] https://github.com/saffsd/langid.py

[6] https://github.com/adbar/german-reddit

high (14.4%), which exemplifies the perplexity of the tool itself, for example because the redditors refer to trending and possibly short-lived notions and celebrities, or because of a high proportion of short, elliptic comments which fail to provide enough morpho-syntactic context. The relatively high but acceptable proportion of foreign words on token level (4%) both confirms this hypothesis and validates the language classification performed during corpus building.

Smiley	Frequency	Smiley	Frequency
:)	3207	:-(	39
;)	1667	-.-	33
:(	590	:’(	31
:-)	299	:))	29
ˆˆ	242	:]	25
;-)	238	=(	19
:/	179	:|	18
=)	96		

Table 2: Most frequent smileys and their frequency

Thanks to the special training of the tokenizer on CMC-data, the smileys can be expected to be treated as whole tokens, which makes a focused analysis possible. The major part of frequent smileys listed in table 2 is commonly used, although there are idioms such as “=)” which may be more frequent on this platform. Emojis do not seem to be frequently used in German comments.

3.2 Sociolinguistic factors

Information about the subreddit of each comment is part of the JSON metadata, which makes the extraction of subreddits straightforward. As can be seen in table 3, the most frequent ones include channels where expression in German is expected (*de*, *rocketbeans*, *kreiswichs*) and other where German is not necessarily spoken but could be appropriate (*germany*, *Austria*). Other channels, which are known to be among the most popular ones but whose link to German is not clear, may include enough quotes or occasional discussions (e.g. *AskReddit*) to explain their presence among the most frequent ones.

The nicknames are also part of the metadata returned by the API and as such they can be considered to be reliable information. A total of 51,155 different nicknames can be found throughout the German subset. 5,343 are marked as deleted, i.e. not active at the time of download.

The most frequent author names in table 4 show that although there is a slight trend towards typical German nouns or syllables, they are not the majority. The crowd seems to be relatively evenly distributed, there is no single nickname outweighing all the others. That said, it can be common practice to change nicknames regularly, which also accounts for the relatively high number of deleted accounts found.

Channel	Frequency
de	14018
AskReddit	8163
rocketbeans	4899
funny	3272
kreiswichs	2848
pics	2813
soccer	2571
Austria	1684
WTF	1592
leagueoflegends	1569
reddit.com	1379
todayilearned	1224
germany	1137
gaming	1133
videos	1124

Table 3: Most frequent channels (subreddits) in the corpus

Nickname	Frequency
Wumselito	262
Aschebescher	238
Clit_Commander	221
oldandgreat	210
GuantanaMo	200
fLekkZ	187
Obraka	180
tin_dog	155
4-jan	151
Omnilatent	141

Table 4: Most frequent nicknames in the corpus and their frequency

4 Visualization of extracted place names

4.1 Method

The Reddit comments are not geotagged. Thus, a proxy has to be found in order to get a glance at their socio-geographical distribution. To do so, place names are extracted and projected on a maps, which allows for a better description of the collected data.

First, the German version of the Wiktionary, a user-curated dictionary launched by the Wikimedia foundation, is used in order to get lexical information about common nouns, which allows for a fine-grained discriminating analysis.

Second, geographical information about the places names has been compiled from the Geonames database[7], which is e.g. used by the Openstreetmap project[8], and whose Creative Commons Attribution license will allow for a release of research data in the near future. All databases for current European countries have been retrieved and preprocessed certain place types have been selected. In fact, toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner and Lieberman, 2011), but knowledge-based methods using fine-grained data have already been used with encouraging results (Hu et al., 2014).

The tokenized corpus has been filtered as described above and matched with the database. This operation includes finite-state automatons at two distinct stages: first to discover potential multi-word place names, and second to select the most probable coordinates in the case of homonyms, based on type, relative distance, and population.

4.2 Results

The maps in figure 1 has been generated by the design environment TileMill[9] and customized using the stylesheet language CartoCSS. Both maps were created using the same data, on the left the scale is smaller, while on the right place names for frequent entries have been added.

The linguistic corpora at the basis of the maps are a construct, and so are the maps themselves: although they seem immediately interpretable, the quality of data, the specialization of the processing tools, and quality assessment all have a major impact on the outcome.

The place names seem to be quite evenly distributed in the German language area relatively to the most populated cities and thus the expectations. There are a few interesting exceptions: Berlin is usually more precisely named (e.g. "Berlin-Kreuzberg" instead of just "Berlin"), which gives the capital the shape of a constellation on the map. A clustering phase could be necessary in order to be able to compare it to the other main cities.

All in all, cities the western part of Germany seems to be more frequently mentioned, particularly when they are home to a well-known soccer team (such as Mönchengladbach). In Austria, Vienna clearly outweighs the rest of the country, which could be explained by a higher international visibility as well as a higher density of early-adopters of Reddit.

5 Conclusion

In this article, I have shown how a corpus focusing on German can be built using the publicly available Reddit comment dataset. In order to get a first impression of the corpus, I collected quantitative information and offered a visualization of structured data, more precisely place names which have to be extracted from the comments since they are not geotagged.

The structural properties of the corpus are in line with the expectations concerning CMC (Barbaresi and Würzner, 2014): short sentences, a relatively high number of different lemmata, and a whole vocabulary of smileys. The different nicknames involved seem to be rather evenly distributed, so are the different place names mentioned in the comments, which is good news in terms of diversity.

Since the license restrictions concerning the dataset are unclear, the corpus is only available upon request. Nonetheless, the German subset can be reconstructed and updated from scratch using code released under open source license.[10]

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Adrien Barbaresi and Kay-Michael Würzner. 2014. For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop proceedings*, pages 2–10. Hildesheim University Press.

Adrien Barbaresi. 2013. Crawling microblogging services to gather language-classified URLs. Workflow and case study. In *Proceedings of the 51th Annual Meeting of the ACL, Student Research Workshop*, pages 9–15.

[7] http://www.geonames.org

[8] https://www.openstreetmap.org

[9] https://www.mapbox.com/tilemill/

[10] https://github.com/adbar/german-reddit

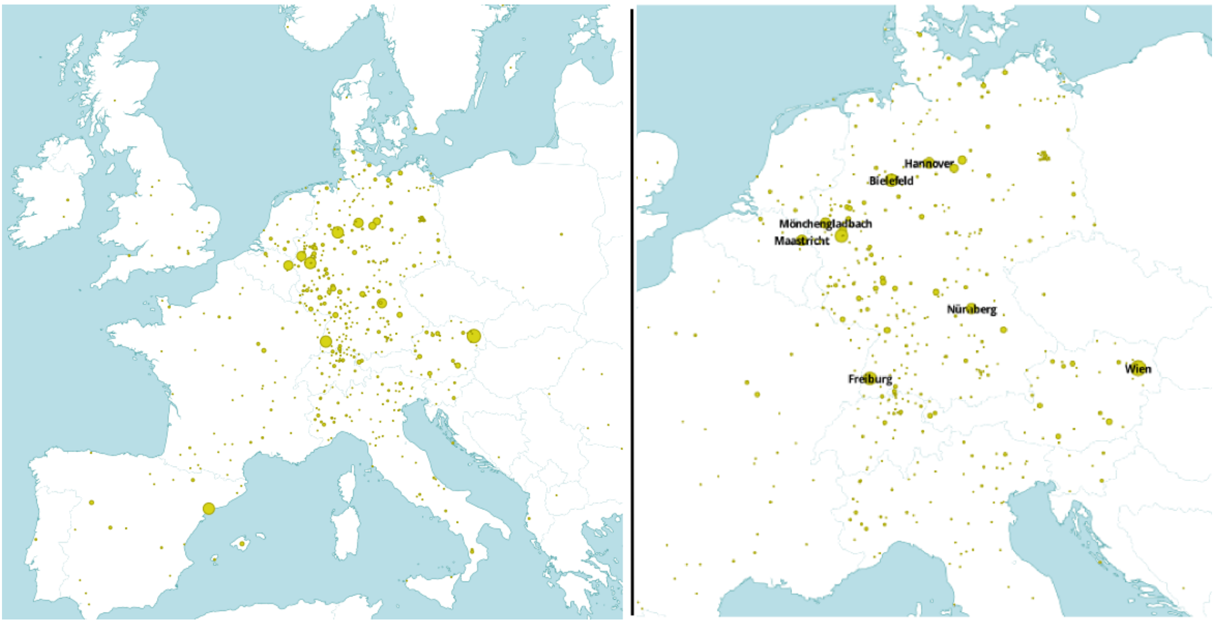


Figure 1: Projection of extracted place names on maps

Adrien Barbaresi. 2014. Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources. In Roland Schäfer and Felix Bildhauer, editors, *Proceedings of the 9th Web as Corpus Workshop*, pages 1–8.

Adrien Barbaresi. 2015. *Ad hoc and general-purpose web corpus construction*. Ph.D. thesis, ENS Lyon.

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. In Christiane Fellbaum, editor, *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*, pages 23–41. Continuum Press.

Yingjie Hu, Krzysztof Janowicz, and Sathya Prasad. 2014. Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia. In *Proceedings of the 8th Workshop on Geographic Information Retrieval*, pages 8–16. ACM.

Bryan Jurish and Kay-Michael Würzner. 2013. Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, 28(2):61–83.

Bryan Jurish. 2003. A Hybrid Approach to Part-of-Speech Tagging. Final report, Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften.

Jochen L Leidner and Michael D Lieberman. 2011. Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, 3(2):5–11.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea.

Marco Lui and Timothy Baldwin. 2014. Accurate Language Identification of Twitter Messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25.

Adding Value to CMC Corpora: CLARINification and Part-of-Speech Annotation of the Dortmund Chat Corpus

Michael Beißwenger[1], Eric Ehrhardt[2], Andrea Horbach[3], Harald Lüngen[4],
Diana Steffen[3], Angelika Storrer[2]

[1] TU Dortmund University, Department of German Language and Literature, D–44221 Dortmund
[2] Mannheim University, Department of German Philology, D–68131 Mannheim
[3] Saarland University, Department of Computational Linguistics and Phonetics, D–66041 Saarbrücken
[4] Institute for the German Language, Department of Central Research: Corpus Linguistics, D–68131 Mannheim

michael.beisswenger@tu-dortmund.de, frehrhar@mail.uni-mannheim.de,
andrea@coli.uni-saarland.de, luengen@ids-mannheim.de,
dsteffen@coli.uni-saarland.de, astorrer@mail.uni-mannheim.de

1 Motivation and Project Framework

ChatCorpus2CLARIN is a curation project of the discipline-specific working group “German Philology” (F-AG 1) within the joint infrastructure project CLARIN-D. In this project, an existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus (cf. 2.1), and samples of other CMC resources will be restructured to conform to current standards for the representation of corpora in the Digital Humanities context. The main goal of this work is to pave the way for the inclusion of linguistically annotated CMC resources in CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology. To this end, the project will (1) transform the metadata and the annotations of the chat corpus into a TEI-compliant format, (2) enrich the data by further linguistic annotations, and (3) integrate the resulting resource into the CLARIN-D Corpus Infrastructures at the Institute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW):

(1) **TEI representation:** For representing the corpus in TEI, the schema drafts and models developed in the TEI special interest group “Computer-mediated communication” are being used. This group is working on a proposal of a TEI standard for CMC genres (Beißwenger et al. 2012, Chanier et al. 2014, Margaretha & Lüngen 2014). In its previous version, the chat corpus has been annotated using a home-grown XML format that describes the main structural features of chat logfiles and user postings as well as selected linguistic phenomena of language use on the internet (emoticons, action words, addressing terms, nicknames). All of these annotations will be transformed into a TEI representation and enriched by additional structural annotations and metadata.

(2) **Additional linguistic annotations:** Except the annotation of selected CMC phenomena, the corpus in its current version does not contain any linguistic annotations. In order to enhance the possibilities for linguistic querying, a layer of part of speech (PoS) annotations will be added to the data. PoS tags using an extended version of the Stuttgart-Tübingen Tagset (STTS, Schiller et al. 1999) have already been added to the corpus using the tools of the project “Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen” (henceforth “Schreibgebrauch”, also see http://www.schreibgebrauch.de) developed at Saarland University.

(3) **Integration into CLARIN-D:** The integration of the resource in the CLARIN-D infrastructures comprises its hosting at the CLARIN-D centres BBAW and IDS and its ingestion in the centres' respective repositories for long-term data archiving. It also comprises developing a CMDI representation of metadata for the resource which will be harvestable via OAI-PMH and accessible from the CLARIN VLO (Virtual Language Observatory). The resource will be addressable via PIDs, it will be searchable in the CLARIN-D Federated Content Search and will also be accessible via web services. The conditions of licensing the corpus resource for scientific use will be defined on the basis of a legal expert opinion that is currently being sought. Depending on the outcome of this expert opinion, the Chat Corpus might be licensed with the CLARIN-D end-user license type PUB (“publicly available”, cf. Oksanen et al. 2010), ACA-NC (academic, non-commercial use, ibid.), or under an alternative license type like the proposed QAO-NC (use via a query engine that retrieves text passages or KWIC lines the size of citations for users registered in CLARIN-D, cf. Kupietz & Lüngen, 2014).

Our contribution to the NLP4CMC workshop focuses on the subtask of PoS tagging. It describes the goals and work packages of the curation project, the resources, the tagging workflow, and first experiences from the post-processing phase.

2 Resources

2.1 The *Dortmund Chat Corpus*

The Dortmund Chat Corpus (Beißwenger 2013) has been collected at TU Dortmund University. The goal of the corpus project was to create a useful resource for researching the peculiarities and linguistic variation in written computer-mediated communication. The corpus comprises 478 logfile documents with 140 240 user postings or 1M words of German chat discourse representing the use of chat software in different application contexts (social chats, advisory chats, chats in the context of learning and teaching, moderated chats in the media context). The corpus has been annotated using an XML format ('ChatXML') that describes (1) the basic structure and properties of chat logfiles and postings, (2) selected "netspeak" phenomena such as emoticons, interaction words, addressing terms, nicknames and acronyms, (3) selected metadata about the chat users. Since 2005, the corpus has been available at http://www.chatkorpus.tu-dortmund.de as an XML version for download and offline querying and as an HTML version for online browsing. It has been widely used as a resource for studying and teaching the peculiarities of German CMC discourse.

2.2 A Tagset for German CMC: 'STTS 2.0'

STTS 2.0 has been created in the context of the DFG scientific network *Empirikom* (http://www.empirikom.net) and of a CLARIN-D initiative and series of workshops (Stuttgart 2012, Tübingen 2013, Hildesheim 2013) for extending the canonical version of STTS (Schiller et al. 1999) for genres which have not been in the scope of the creators of STTS so far (cf. the volume by Zinsmeister et al. 2014). While STTS (1999) focuses mainly on parts of speech in genres of edited text (e.g. newspaper articles, novels), STTS 2.0 builds on the categories of STTS (1999) and extends it with categories and tags for two types of items which have to be taken into consideration when tagging CMC and social media discourse: (1) tags for phenomena which are specific to CMC / social media discourse (emoticons, action words, addressing terms, hash tags, URLs, email addresses), and (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers (e.g., modal particles, discourse markers, colloquial contractions). These extensions are useful for corpus-based research of both CMC and spoken conversation. A common tag set for phenomena of type (2) will also facilitate the comparison of written CMC with transcripts of spoken conversation.

STTS 2.0 exists in two versions:

- a version described in Bartz et al. (2014) as an intermediate result from and contribution to the discussions in the context of the CLARIN-D STTS initiative 2012/2013. This version has been adopted and slightly modified for adapting a PoS tagger within the project "Schreibgebrauch" at Saarland University in Saarbrücken in 2014/15 (Horbach et al. 2014, henceforth *STTS 2.0-BETA*, cf. 2.2.1);
- a version that builds on Bartz et al. (2014) and includes the results from further discussions in the CLARIN-D STTS initiative and in the Empirikom network and which has been made compatible with the modified STTS defined by Westpfahl & Schmidt (2013) and Westpfahl (2014) for tagging the "Research and Teaching Corpus of Spoken German" (FOLK, http://agd.ids-mannheim.de/folk.shtml) at the IDS Mannheim (Beißwenger et al. 2015, henceforth *STTS 2.0-ALPHA*, cf. 2.2.2).

2.2.1 Tagset Used in the Automatic Annotation Pipeline ('STTS 2.0-BETA')

In order to facilitate and speed up human corpus annotation, we use an automatic tool chain from the "Schreibgebrauch" project to pre-annotate the Dortmund Chat Corpus. The tagging component uses a slightly modified version of the tagsets described in Bartz et al. (2014) and Beißwenger et al. (2015) (cf. Horbach et al. 2014). In particular, the tagset differs in the following points:

- The tagset does not differentiate between ASCII and graphic emoticons.
- The tag for interaction words is split into action word indicators (i.e. the * surrounding the actual interaction word), and the interaction word itself, leading e.g. to tagging results like */AWIND Kaffee/NN trink/AW */AWIND.
- There are no particular tags for various kinds of particles or discourse markers, but they are annotated following the original STTS as adverb or conjunction.
- Extra tags are used to mark words that have been erroneously separated or merged, such as "anzu melden" instead of "anzumelden".

2.2.2 Tagset Used as the Target Tagset in the *ChatCorpus2CLARIN* Project ('STTS 2.0-ALPHA')

STTS 2.0-ALPHA is a slightly revised version of the tagset described in Bartz et al. (2014). It has been described in the guideline document Beißwenger et al. (2015) and will be used as the reference tagset in the Empirikom Shared Task for Automatic Linguistic Annotation of German CMC (https://sites.google.com/site/empirist2015/). It is compatible with the modified STTS that will be used for tagging the FOLK corpus at the IDS (Westpfahl & Schmidt 2013, Westpfahl 2014).

Tab. 1 (see appendix) provides an overview of the tags and categories defined in STTS 2.0-ALPHA. The categories defined for CMC-specific items as well as the extensions for frequent types of colloquial contractions are true extensions to STTS (1999). The categories defined for phenomena which are typical of spontaneously spoken language restructure parts of the categories of STTS (1999). Nevertheless, all modifications and extensions defined in STTS 2.0-

ALPHA result in a category set which is still downwardly compatible with STTS (1999) and therefore allows for interoperability with corpora that have been tagged with STTS (1999) (e.g. DWDS, the "Digital Dictionary of the German Language", http://www.dwds.de).

3 Tagging

The pipeline for pre-annotating the Dortmund Chat Corpus uses tools for sentence segmentation and tokenisation, PoS tagging and lemmatisation. For sentence segmentation and tokenisation we used the open source tokeniser jTok (https://github.com/DFKI-MLT/JTok). It can be adapted to different text types since it uses editable regular expressions to define tokens.

For both PoS tagging and lemmatisation we use the TreeTagger. We employ tagging models from Horbach et al 2014, which have been adapted towards CMC data. In this work, the standard TIGER training data set (Brants et. al. 2004) of about 50 000 newspaper sentences has been extended with relatively small amounts of manually annotated CMC data. they annotated about 12 000 tokens for each of the three CMC genres of forum posts, chat and twitter data with STTS 2.0-BETA tags. The chat subcorpus is taken from the Dortmund Chat Corpus. One third of each dataset has been added to TIGER (boosted 5 times in order to give additional weight to the new material) as training data, while the other two thirds have been held out for testing. These gold annotations can be obtained for research purposes directly from the "Schreibgebrauch" project.

Using a tagger model trained with this enriched training set, performance on the chat part of the test portion of the above mentioned gold-standard annotations could be increased from 71.4% (using an out-of-the-box model trained on TIGER only) to 83.5%. As no lemmatisers adapted towards CMC are available (and our annotations did not comprise lemma information), we used the standard TreeTagger lemmatiser trained on TIGER.

4 Outlook: Post-processing

Parts of the automatically PoS-tagged chat corpus will be manually post-processed, i.e. adapted and amended on the basis of the STTS 2.0-ALPHA tagset as described in section 2.2.2. Post-processing will also concern the levels tokenisation, (orthographic) normalisation, and lemmatisation. The goal of this effort is to create a resource of correct reference annotations for chat data which may be used (a) to demonstrate how a precise tokenisation, PoS annotation, lemmatisation and normalisation of (parts of) a chat corpus will support linguistic users in defining sophisticated corpus queries for their linguistic research questions, (b) as a data set for (re-)training and evaluating NLP tools for the various above-mentioned linguistic processing steps for CMC-specific linguistic items and "non-standard" phenomena in written CMC and social media discourse. Furthermore, the results of the post-processing shall serve as a basis for developing better tokenisation and lemmatisation guidelines for CMC.

Manual post-processing will be carried out by a team of students, using the normalisation editor OrthoNormal in FOLKER ("FOLK-Tools", Schmidt 2012), which has originally been developed and applied for the manual normalisation and correction of POS-tagged spoken language transcripts in the FOLK corpus at the IDS (Westpfahl & Schmidt 2013). A more recent version of FOLKER (preview version 1.2) provided by Thomas Schmidt (IDS) offers a new import and export interface for PoS-tagged ChatXML. Fig. 1 (see appendix) shows a screenshot of editing these data in OrthoNormal. At the NLP4CMC workshop we will present first results of comparing a sample of the automatic PoS annotation using STTS2.0-BETA with an "expert" annotation using STTS2.0-ALPHA and discuss the results by the hand of examples.

References

Books/Papers

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41 (1), 161-164. Extended version: http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1), 157-198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI) 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline Document, Dortmund 2015. https://sites.google.com/site/empirist2015/home/annotation-guidelines

Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; Knig, Esther; Lezius, Wolfgang; Rohrer, Christian; Smith, George; Uszkoreit, Hans (2004): TIGER: Linguistic interpretation of a german corpus. Journal of Language and Computation, Special Issue, 2(4), 597-620.

Chanier, Thierry; Poudat, Celine; Sagot, Benoit; Antoniadis, Georges; Wigham, Ciara; Hriba, Linda; Longhi, Julien; Seddah, Djamé (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Journal of Language Technology and Computational Linguistics JLCL 29 (2), 1-30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf

Horbach, Andrea; Steffen, Diana; Thater, Stefan; Pinkal, Manfred (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of KONVENS 2014, 171-177.

Horbach, Andrea; Thater, Stefan; Steffen, Diana; Fischer, Peter M.; Witt, Andreas; Pinkal, Manfred (2015): Internet Corpora: A Challenge for Linguistic Processing. In: Datenbank-Spektrum 15 (1), 41-47.

Kübler, Sandra; Baucom, Eric (2011): Fast domain adaptation for part of speech tagging for dialogues. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, RANLP, 41-48.

Kupietz, Marc; Lüngen, Harald (2014): Recent developments in DEREKO. In: Nicoletta Calzolari et al. (eds): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Journal of Language Technology and Computational Linguistics (JLCL) 29 (2), 59-82. http://www.jlcl.org/2014_Heft2/3Margaretha Luengen.pdf

Oksanen, Ville; Lindén, Krister; Westerlund, Hanna (2010): Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. In: Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Malta.

TEI Consortium (2015): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available online at: http://www.tei-c.org/Guidelines/P5/

Schiller, Anne; Teufel, Simone; Stöckert, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmidt, Thomas (2012): EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In: Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop, 47-50.

Westpfahl, Swantje; Schmidt, Thomas (2013): POS für(s) FOLK – Part of Speech-Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics 28 (1), 139-156. http://www.jlcl.org/2013_Heft1/6Westpfahl.pdf

Westpfahl, Swantje (2014): STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In: Lori Levin und Manfred Stede (eds.): Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 1–10. http://www.aclweb.org/anthology/W14-4901.

Zinsmeister, Heike; Heid, Ulrich; Beck, Kathrin Beck (Eds., 2014): Das STTS-Tagset für Wortartentagging - Stand und Perspektiven. Special issue of the Journal for Language Technology and Computational Linguistics. http://www.jlcl.org

Internet Sources

„Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen“: http://www.schreibgebrauch.de/

CLARIN-D (“Common Language Resources and Technology Infrastructure”) – German section: http://www.clarin-d.de/en/

Dortmund Chat Corpus (“Dortmunder Chat-Korpus"): http://www.chatkorpus.tu-dortmund.de/

DWDS („Digitales Wörterbuch der deutschen Sprache“): http://www.dwds.de

Empirikom (DFG scientific network “Empirische Erforschung internetbasierter Kommunikation”): http://www.empirikom.net

EmpiriST2015 Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication: https://sites.google.com/site/empirist2015/

FOLK (“Forschungs- und Lehrkorpus Gesprochenes Deutsch”): http://agd.ids-mannheim.de/folk.shtml

FOLKER (Transcription editor for FOLK), preview version 1.2 with functionalities for editing data from the Dortmund chat corpus with OrthoNormal: http://agd.ids-mannheim.de/folker.shtml

JTok (rule-based tokeniser): http://heartofgold.opendfki.de/browser/trunk/jtok

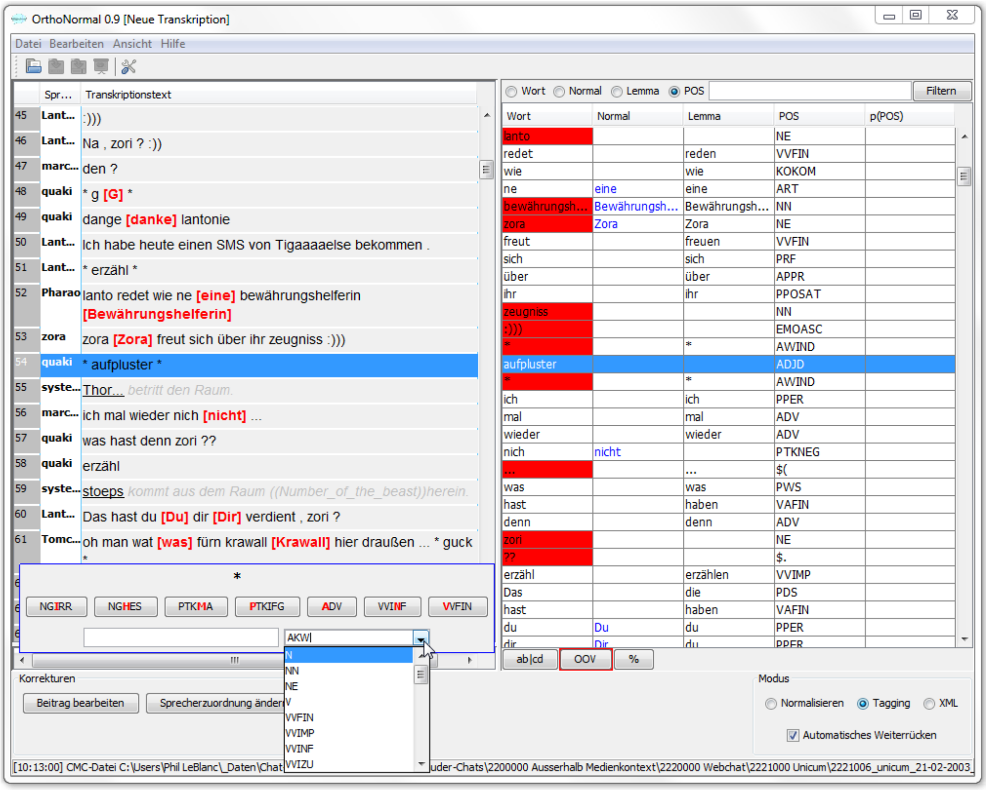
TEI special interest group „Computer-mediated communication“: http://www.tei-c.org/Activities/SIG/CMC/

Virtual Language Observatory (VLO): https://vlo.clarin.eu/

Appendix

PoS tag	Category	Examples
I. Tags for phenomena which are specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	Graphic emoticon	😄 👽 😜
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://www.tu-dortmund.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART is already existing in STTS 1999)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone
PTK IFG	'Intensitätspartikeln', 'Fokuspartikeln', 'Gradpartikeln'	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ... *with V2 clauses*
ONO	Onomatopoeia	boing, miau, zisch

Tab. 1: Overview of extensions and modifications to STTS (1999) in STTS 2.0-ALPHA (Beißwenger et al. 2015).



	Spr…	Transkriptionstext
45	Lant...	:)))
46	Lant...	Na , zori ? :))
47	marc...	den ?
48	quaki	* g [G] *
49	quaki	dange [danke] lantonie
50	Lant...	Ich habe heute einen SMS von Tigaaaaelse bekommen .
51	Lant...	* erzähl *
52	Pharao	lanto redet wie ne [eine] bewährungshelferin [Bewährungshelferin]
53	zora	zora [Zora] freut sich über ihr zeugniss :)))
54	quaki	* aufpluster *
55	syste...	Thor... betritt den Raum.
56	marc...	ich mal wieder nich [nicht] ...
57	quaki	was hast denn zori ??
58	quaki	erzähl
59	syste...	stoeps kommt aus dem Raum ((Number_of_the_beast))herein.
60	Lant...	Das hast du [Du] dir [Dir] verdient , zori ?
61	Tomc...	oh man wat [was] fürn krawall [Krawall] hier draußen ... * guck *

*

AKW|

○ Wort ○ Normal ○ Lemma ● POS

Wort	Normal	Lemma	POS	p(POS)
lanto			NE	
redet		reden	VVFIN	
wie		wie	KOKOM	
ne	eine	eine	ART	
bewährungsh...	Bewährungsh...	Bewährungsh...	NN	
zora	Zora	Zora	NE	
freut		freuen	VVFIN	
sich		sich	PRF	
über		über	APPR	
ihr		ihr	PPOSAT	
zeugniss			NN	
:)))			EMOASC	
*		*	AWIND	
aufpluster			ADJD	
*		*	AWIND	
ich		ich	PPER	
mal		mal	ADV	
wieder		wieder	ADV	
nich	nicht		PTKNEG	
...		...	$(	
was		was	PWS	
hast		haben	VAFIN	
denn		denn	ADV	
zori			NE	
??			$.	
erzähl		erzählen	VVIMP	
Das		die	PDS	
hast		haben	VAFIN	
du	Du	du	PPER	
dir	Dir	du	PPER	

○ Normalisieren ● Tagging ○ XML

☑ Automatisches Weiterrücken

[10:13:00] CMC-Datei C:\Users\Phil LeBlanc_Daten\Chat…uder-Chats\2200000 Ausserhalb Medienkontext\2220000 Webchat\2221000 Unicum\2221006_unicum_21-02-2003_

Fig. 1: Editing PoS-tagged ChatXML with OrthoNormal.

Building and Annotating a Corpus of German-Language Newsgroups

Jasmin Schröck
Institut für Deutsche Sprache
Mannheim
schroeck@direktion.ids-mannheim.de

Harald Lüngen
Institut für Deutsche Sprache
Mannheim
luengen@ids-mannheim.de

Abstract

Usenet is a large online resource containing user-generated messages (news articles) organised in discussion groups (newsgroups) which deal with a wide variety of different topics. We describe the download, conversion, and annotation of a comprehensive German news corpus for integration in DeReKo, the German Reference Corpus hosted at the Institut für Deutsche Sprache in Mannheim.

1 Introduction

Usenet news are an instance of the genre of computer-mediated communication (CMC) which is of interest in many current research questions (cf. Beißwenger & Storrer 2008). Recent initiatives for the creation of CMC corpora have co-operated firstly in the DFG research network empirikom (Beißwenger 2012), and since 2013 within the TEI Special Interest Group on CMC, amongst other things to create a TEI-based standard for the encoding and annotation of CMC corpora for use in empirical linguistics research. Several CMC corpora based on versions of the encoding scheme provided by the TEI CMC SIG have been compiled so far, e.g. (German) chat and wikipedia discussions (Beißwenger et al. 2012; Margaretha & Lüngen 2014) and (French) corpora of various CMC subgenres in the project CoMeRe (Chanier et al. 2014). Currently the Dortmund Chatkorpus (Beißwenger 2013) is being prepared along the lines of the TEI CMC SIG for integration in CLARIN research infrastructures. Consequently, the aim of the work described in this paper was to close another gap by creating an edited Usenet corpus containing all newsgroups from the de. hierarchy and annotating relevant CMC phenomena according to the principles proposed by the TEI CMC SIG. The news corpus has been marked up for metadata and text structure according to I5, which is the TEI customization (Lüngen & Sperberg-McQueen 2012) used for the encoding of texts in DeReKo and which incorporates features of the TEI CMC SIG.

2 Usenet

Usenet originated in 1979 and is based on the NNTP internet protocol (Horton & Adams 1987). The features of news messages include rich formatted metadata (the NNTP header) with fields for the sender, the posting date, the subject, the reply history of a message and other types of information. Header fields are obligatory or optional.

In the message body, many textual features also found in emails or letters prevail, such as salutations (openers and closers), postscripts, or signatures. Another characteristic feature is the highly recursive usage of quotations from previous articles, often introduced by an automatically generated line containing the e-mail address and name of the author and the posting date of the quoted article. Finally, the language used in news messages contains many familiar netspeak phenomena such as the use of emoticons, interjections, and inflectives (cf. Feldweg et al. 1995; Gausling 2005).

A newsgroup works similar to a web discussion forum, one difference being that all newsgroups are organised in a universal, topic-based hierarchy. Newsgroups are stored world-wide on so-called news servers, and everyone is free to set up such a server. All news servers are regularly synchronised with each other so that they offer the same amount of news messages in each newsgroup sooner or later. Similarly, everyone is free to connect to a news server using news client software to subscribe to newsgroups to read

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 17–22, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

messages and to post one's own news messages to the server.

Usenet communication has had its heydays in the 1990s, consequently it is a pre-Web 2.0 form of CMC. But Usenet lives on, as a dedicated community has constantly been using it.

3 Related Work

A previous German Usenet corpus initiative was undertaken in the ELWIS project (*Corpus-based development of lexical knowledge bases*), where a corpus of contemporary German was compiled, beginning in 1992. All messages of the year 1993, containing altogether 433,000 articles in 647 newsgroups, served as a base for the investigation of the language use in newsgroups (cf. Feldweg et al. 1995). More recently, the WestburyLab at the University of Alberta collected English-language Usenet data in the project *A reduced redundancy Usenet Corpus* from 2005 until 2011 (Shaoul & Westbury 2013). Their corpus covers 47,860 newsgroups containing more than seven billion words and seems to be the largest news archive ever prepared as a linguistic corpus. Another English-language corpus was created by Matt Mahoney in the project *Usenet as a text corpus* in 2000, containing 53,247 articles from 9,359 newsgroups (Mahoney 2000). Neither of the previous corpora seems to have been marked up using XML/TEI, nor have they been annotated for CMC phenomena.

4 Creation of the Corpus

Following a common strategy in the construction of TEI corpora from text archives (cf. e.g. Fankhauser et al. 2013; Margaretha & Lüngen 2013), we divided the corpus creation into several steps: In a first step, all German-language Usenet data currently available on the newsserver news.individual.de was downloaded and converted into a well-formed XML version of the NNTP format (dubbed nntpXML) in a straightforward way. In a second stage, the nntpXML data was filtered and converted into the TEI-based I5 target format. In the third stage, we applied heuristics for the annotation of CMC-phenomena typical of Usenet articles to the newly created corpus, creating I5 with annotations (see also Figure 1).

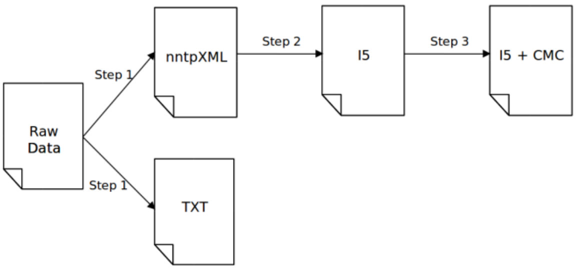


Figure 1: Workflow

4.1 Download and conversion to nntpXML

In the first stage, all currently available Usenet articles of all 379 German-language newsgroups from the de-hierarchy were downloaded from the newsserver news.individual.de (run by FU Berlin, with a retention time of 621 days) on 1 June 2015 using the Python client nntplib. The downloaded data was preprocessed by converting it to Unicode and to well-formed nntpXML, using the Python library lxml. For each newsgroup, a separate file was generated. The original Usenet structure was mostly preserved, only the header lines were ordered in obligatory, optional and others (see Horton & Adams 1987).

4.2 Conversion from nntpXML to I5

The generated nntpXML files were then converted to the TEI format I5 which is used for DeReKo. An I5 corpus file is structured according to the three levels *corpus* (<idsCorpus>, the root element), *document* (<idsDoc>), and *text* (<idsText>). All news articles of one calendar year were stored in a separate <idsDoc> while each article was included in a separate <idsText> document. Note that this corpus structure differs from previous CMC corpora where one thread or logfile containing a set of postings usually corresponds to one corpus text (Beißwenger et al. 2012; Margaretha & Lüngen 2014; Chanier et al. 2014). With news articles (and similarly emails), the messages come neither grouped in a self-contained document (like e.g. a Wikipedia page), nor is news a synchronous type of communication like chat, hence we do not consider threads or logfiles as suitable corpus units for news. Each of the three levels received its own header containing the metadata that were appropriate and could be extracted from the messages. The I5 structure was created using python with lxml, and XSLT stylesheets.

A major task was to identify TEI elements for the encoding of the metadata of the original header of each article. One issue in this area was how to represent the reply history of a message as contained in the NNTP "References" header line. From this header field, news readers like Mozilla Thunderbird derive the threaded view of the messages. Since neither the TEI Guidelines, nor the TEI CMC SIG provides metadata elements for the reply history, we resorted to a recent proposal by the TEI Correspondence SIG (2015). This SIG develops, amongst other things, TEI elements for the encoding of correspondence-specific metadata applying to all kinds of correspondence such as letters, telegrams, diaries, e-mails and blogs. Consequently, we added the elements <correspDesc> and <correspContext> to I5. The latter serves the encoding of information about previous and following messages in a correspondence.[1]

Furthermore, we adopted the suggestion by Beißwenger et al. (2012) to create a list of participants in the newsgroup (<listPerson>) and a timeline (<timeline>), but stored them in separate files as a step in the anonymisation of the data. The list of persons contains the e-mail address for each participant and their name if available. Also, following Beißwenger et al. (2012), the text of an article was wrapped in a <posting> element whose attributes @who and @synch refer to the corresponding ID in the list of persons and the timeline, respectively.

4.3 Annotation of CMC phenomena and quality assessment of the annotations

For an annotation of CMC phenomena as introduced in Section 2, we developed several heuristics and implemented them in an XSLT 2.0 stylesheet, creating a regular expression for each phenomenon and tagging the matching strings with a suitable TEI element. The following CMC phenomena were annotated: quotations, as well as the lines introducing them, links to the World Wide Web, links to other newsgroups, salutations (openers and closers), postscripts, user signatures, and emoticons. Apart from these CMC categories, paragraphs were annotated. Table 1 shows the phenomena and the respective elements used for their annotation.

Table 1: Annotated phenomena, used elements and their source

Phenomenon	**Source**	**Example**
Link to www	TEI	<ref type="www" target="URL"> URL </ref>
Link to newsgroup	TEI	<ref type="newsgroup" target="de.rec.fahrrad"> de.rec.fahrrad </ref>
Opener	TEI	<seg type="opener"> Hallo, </seg>
Closer	TEI	<seg type="closer"> Ciao, NAME </seg>
Postscript	TEI	<seg type="postscript"> P.S. dürften sich eigentlich links der durchgezogenen Linie in dieser Fahrradstraße noch Fahrräder aufhalten? </seg>
Signature	TEI	<trailer> -- Life's a road, not a destination. </trailer>
Emoticon	Beißwenger et al. (2012)	<interactionTerm> <emoticon> :-(</emoticon> </interactionTerm>
Quotation, with or without introductory line	TEI	<cit type="replyCit"> <bibl type= "introQuote"> Am 23.09.2013 12:33, schrieb NAME: </bibl> <quote> <p>Die Zukunft ist da, seilzuglose Rennräder sind möglich geworden. </p> </quote> </cit>

We assessed the quality of the CMC annotations by conducting a small evaluation of each annotated CMC feature on 200 articles from five newsgroups, which according to their topics seemed reasonably diverse: de.etc.sprache.deutsch (the German language),

[1] Apparently, these elements have been added to the official TEI Guidelines in the meantime, cf. TEI Consortium (2015).

de.rec.mampf (food/eating) de.comp.os.ms-windows.misc (windows operating system), at.gesellschaft.politik (society and politics, Austria) and de.soc.senioren (senior citizens) (cf. Schröck 2015). For this test set, correct reference annotations were created manually by an expert familiar with the TEI elements used for the annotation of the CMC categories and their TEI (or SIG) definitions. The reference set eventually contained 3,291, the test set 3,438 annotation instances (TEI elements marking up CMC phenomena). Comparing the test set with the reference, the micro average precision over all eleven annotation categories was found to be 79%, and the micro average recall was 82%. The categories that were identified best by the regular expressions were signature and emoticon. The categories most difficult to identify were postscript and opener. However, these two categories, and also links to newsgroups, didn’t occur very frequently in the test and reference sets, which were relatively small.

Furthermore, the results for openers, closers and introduction lines of quotes, which often contain names, could be improved by using Named Entity Recognition.

The results for all elements and the overall results are shown in Table 2.

Table 2: Number of annotation instances for each element in reference set (# ref) and test set (# test) and micro-averaged precision (P), recall (R) and F-measure (F)

I5 Tag	# ref	# test	P	R	F
<cit>	618	661	.69	.74	.71
<bibl>	359	263	.94	.69	.80
<quote>	606	661	.80	.87	.83
<p>	1186	1358	.79	.91	.85
<ref type= ”www”>	109	106	.88	.85	.86
<seg type= “signature“>	89	86	1	.97	.98
<emoticon>	99	89	.93	.84	.88
<ref type= ”newsgroup”>	13	28	.46	1	.63
<seg type= ”postscript”>	2	16	.13	1	.23
<seg type= ”closer”>	182	140	.81	.63	.71
<seg type= ”opener”>	28	30	.4	.43	.41
	Σ = 3291	Σ = 3438	Ø = .79	Ø = .82	Ø = .80

5 The Corpus

Download was carried out on 1 June 2015 and took 12 hours, using four threads on a linux machine with an AMD Opteron 8439 SE processor with 48 cores at 2.8GHz, and 256G RAM. The news server potentially contained 1,004,157 articles in 379 newsgroups in the *de* hierarchy; however, 62,878 articles were discarded because their X-No-Archive field was set to yes, and another 70,376 because their encoding could not be determined and hence not be converted to UTF-8. The conversion-to-I5 phase took 2:20 hours, and the annotation phase took 54 hours. Four newsgroups contained no messages, and with one newsgroup, the annotation did not terminate. From the remaining 374 groups, 11 messages were discarded because they contained an error in the Date field. The resulting, annotated full corpus in I5 format contains 374 newsgroups comprising 870,892 news articles with 128.78 million word tokens. It takes up 7.2G of disk space. The corpus contains messages posted between 24/9/2013 and 1/6/2015. The biggest newsgroup (929MB) is de.soc.politik.misc containing 117,950 messages (16.8 million word tokens). However, the size of the corpus will be further reduced in the deduplication and cleaning step.

6 Conclusion

The news corpus described in this paper is currently being further anonymised, cleaned, and de-duplicated, mostly according to the principles described in Shaoul & Westbury (2013). The resulting version is scheduled to be included in the upcoming DeReKo release DeReKo-2015-II. However, the question of whether the corpus can be shared with the linguistic community remains to be solved. CMC texts, like all other texts, are subject to copyright, and in principle each author of an article contained in the corpus would have to give his or her consent first. On the other hand we think that a news article is not the same as a text on the W3C, as someone who posts to a newsgroup is aware of the fact (in fact wants) that his/her message will be distributed to many servers all over the world in the first place. We are currently seeking legal advice in this matter in cooperation with the CLARIN-D curational project Chatkorpus2CLARIN.[2] Until further notice, the news corpus will be accessible from the premises of the IDS Mannheim only.

We intend to update the corpus with new news articles regularly. We are also aiming at downloading from a news server with a longer retention time, though as far as we can see, longer retention times are only offered by commercial news servers.

Reference

Beißwenger, Michael (2012): Forschungsnotiz: Das Wissenschaftliche Netzwerk "Empirische Erforschung internetbasierter Kommunikation" (Empirikom). In: *Zeitschrift für germanistische Linguistik* 40 (3), pp. 459-461.

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: *Zeitschrift für germanistische Linguistik* 41 (1), pp. 161-164.

Beißwenger, Michael; Storrer, Angelika (2008): Corpora of Computer-Mediated Communication In: Lüdeling, Anke; Kytö, Merja (eds.): *Corpus Linguistics. An international Handbook.* Vol. 1, Berlin: de Gruyter, pp. 292-308.

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: *Journal of the Text Encoding Initiative* [Online] 3.

Beißwenger, Michael; Lemnitzer, Lothar (2013): Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt "Digitales Wörterbuch der deutschen Sprache" (DWDS). In: *Journal for Language Technology and Computational Linguistics (JLCL)* 28 (2), Special issue on "Webkorpora in Computerlinguistik und Sprachforschung".

Chanier, Thierry; Poudat, Céline; Sagot, Benoît; Antoniadis, Georges; Wigham, Ciara R.; Hriba, Linda; Longhi,_Julien; Seddah, Djamé (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: *Journal of Language Technology and Computational Linguistics (JLCL)* 29 (2), pp. 1-30.

Feldweg, Helmut; Kibinger, Ralf; Thielen, Christine. (1995): Zum Sprachgebrauch in deutschen Newsgruppen. In: Schmitz, U. (ed.): *Neue Medien.* Osnabrücker Beiträge zur Sprachtheorie 50, Oldenburg: Red. OBST, pp. 143-154.

Gausling, Timo (2005): Der Newsgroup-Beitrag - eine kommunikative Gattung? In: *Studentische Arbeitspapiere zu Sprache und Interaktion (SASI)* 4. Series "Arbeitspapiere des Centrum Sprache und Interaktion der Westfälischen Wilhelms-Universität", available online at: http://noam.uni-muenster.de/sasi/Gausling_SASI.pdf (last visited 2015-07-10).

Horbach, Andrea; Thater, Stefan; Steffen, Diana; Fischer, Peter M.; Witt, Andreas; Pinkal, Manfred (2015): Internet Corpora: A Challenge for Linguistic Processing. In: *Datenbank-Spektrum* 15 (1), pp. 41-47.

Horton, M.; Adams, R. (1987): *RFC-1036 Standard for Interchange of USENET Messages.* Available online at: http://tools.ietf.org/html/rfc1036 (last visited 2015-07-10).

Lüngen, Harald; Sperberg-McQueen, C. M. (2012): A TEI P5 Document Grammar for the IDS Text Model. In: *Journal of the Text Encoding Initiative* [Online] 3.

Mahoney, Matt (2000): *Usenet as a text corpus.* Florida Tech, CS Dept., available online at: https://cs.fit.edu/mmahoney/dissertation/corpus.html (last visited 2015-07-13).

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: *Journal of Language Technology and Computational Linguistics (JLCL)* 29 (2), pp. 59-82.

Schröck, Jasmin (2015): *Erstellung eines deutschsprachigen Usenet-Newsgroup-Korpus und Annotation von Phänomenen internetbasierter Kommunikation.* BA-Thesis, University Heidelberg.

Shaoul, Cyrus; Westbury Chris (2013): *A reduced redundancy USENET corpus (2005-2011).* Edmonton, AB: University of Alberta, available online at: http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html (last visited 2015-07-13).

TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Available online at: http://www.tei-c.org/Guidelines/P5/ (last visited 2015-07-10).

TEI Correspondence SIG (2015). Information and examples available online at: http://www.tei-c.org/Activities/SIG/Correspondence/,

[2] http://chatkorpus.tu-dortmund.de/

http://wiki.tei-c.org/index.php/SIG:Correspondence and https://github.com/TEI-Correspondence-SIG/ (last visited 2015-07-10).

Using discursive information to disentangle French language chat

Matthieu Riou   **Soufian Salim**   **Nicolas Hernandez**
LINA UMR 6241 Laboratory
University of Nantes (France)
`firstname.lastname@univ-nantes.fr`

Abstract

In internet chatrooms, multiple conversations may occur simultaneously. The task of identifying to which conversation each message belongs is called disentanglement. In this paper, we first try to adapt the publicly available system of Elsner and Charniak (2010) to a French corpus extracted from the Ubuntu platform. Then, we experiment with the discursive annotation of utterances. We find that disentanglement performances can vary significantly depending on corpus characteristics. We also find that using discursive information, in the form of functional and rhetoric relations between messages, is valuable for this task.

1 Introduction

Interest in live chats has grown as they gained popularity as a channel for computer-mediated communication. While many chat services are designed to only allow dialogue in between two participants, a number of them let multiple participants join in and send messages into a common text stream. It is therefore frequent for multi-party chats to feature several simultaneous conversations. The task of identifying these conversations and the messages that belong to them is called disentanglement. It is a required preprocessing step for many higher-level analysis systems, such as those that rely on contextual knowledge about utterances. For example, dialogue act classifiers typically depend on information about previous utterances in the conversation (Kim et al., 2012). Moreover, any sequential analysis system, such as one based on CRFs, would in fact require chat disentanglement.

In this paper, we first consider the existing disentanglement system proposed by Elsner and Charniak (2010), which is based on lexical analysis, and attempt to adapt it to French language chats from the Ubuntu chatrooms. Then, we experiment with discursive information by annotating relations between messages, and try to see if adding such information to the feature sets improves upon the existing system.

2 Related Work

Elsner and Charniak are predominant in the literature. Most notably, they worked on the construction of an annotated corpus used as a reference for many works, *"Are you talking to me?"* (Elsner and Charniak, 2008). It was used by Wang and Oard (2009), who proposed a method for conversation reconstitution based on message context. Their results are state-of-the-art for Elsner and Charniak's corpus. Mayfield et al. (2012) proposed a learning model for the detection of information-sharing acts at the sentence level, then the aggregation of these sentences into sequences and finally the clustering of resulting sequences into conversations.

This paper is primarily based on the publicly available system presented by Elsner and Charniak (2010)[1]. They adopt a two-step approach to chat disentanglement. The first step is to determine for each pair of messages whether they belong to the same conversation or not. In order to do that, they start by identifying candidate message pairs. These pairs are formed based on whether the two messages were sent in an interval of 129 seconds or less[2]. The intuition behind this heuristic is that the more distant in time two messages are, the less likely it is that they belong to the same conversation. Then, they use a maximum-entropy classifier to determine whether they actually do. The second step is to partition these messages into several clusters to obtain the automatically annotated cor-

[1] `http://www.ling.ohio-state.edu/~melsner/#software`

[2] This particular value was chosen because it is the threshold after which the classifier no longer outperforms the majority baseline.

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 23–27, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

pus. They accomplish this by using a greedy voting algorithm.

Their system is based on lexical similarity between messages as well as chat-specific and discourse features. However, said discourse features remain simple and could easily be improved upon: they only record the presence of cue words (indicating greetings, answers and thanks), whether a message is a question, and whether a message is long (over ten words). Most importantly, the system does not make use of any information about message context at all.

3 Method

We first present how we adapted Elsner and Charniak's system to be used on French language data. Then, we show how we extended the base system to make use of additional discursive information.

3.1 French-language implementation

The main reason why their system requires adapting before it can be used on non-English corpora is that it relies on linguistic resources: a list of stop words, a list of technical words, and several lists of cue words to recognize greetings (3 words: "hey", "hi" and "hello"), answers (5 words: "yes", "yeah", "ok", "no" and "nope") and thanks (3 words: "thank", "thanks" and "thx").

For our adaptation, stop words (the fifty shortest words) were directly extracted from the corpus. The list of technical words was generated from different sources: some were extracted from the corpus (URLs and large numbers), and some were extracted from Ubuntu's French and English language documentations[3][4][5].

Cue words were translated into French and expanded to include more variations. Relevant English terms commonly used in French were kept. As a result we obtained a list of 24 cue words for greetings, 85 for answers and 23 for thanks.

3.2 Extension with relational information

We propose an extension to Elsner and Charniak's system that consist of the addition of new discursive information on top of existing lexical features. One of this paper's goals is to measure how discursive information can improve the performance of a chat disentanglement system.

Here, we focus on relations between messages. We present a simple annotation scheme for inter-utterance relations inspired by the DIT++ taxonomy of dialogue acts (Bunt, 2009). We distinguish between functional dependencies (such as the one between a question and an answer) and rhetorical relations (such as the one between a clarification and the utterance it relates to). Rhetorical relations are further subdivided into three classes: explicit subordination relations, explicit coordination relations and implicit coordination relations. We distinguish explicit and implicit coordination relations in order to represent differently series of utterances that are only indirectly related. For example, two consecutive questions could be merely related by the fact that they both serve to further the advancement of a common task. These situations are frequent in problem-oriented conversations such as those found in the Ubuntu platform.

These four classes allow us to represent the structure of a multi-party dialogue. There are two ways they can then be used to build features for the disentanglement system. For each message pair, we can choose to record only whether they are related in some way, or we can have a separate feature for each kind of relation. It is interesting to note that specifying the relation type can be informative and could help determine whether implicit coordination relations are relevant to the task. Because of their implicit nature we expect them to be very hard to automatically detect, so if they happen to be instrumental for chat disentanglement the overall difficulty of the task might be higher than expected.

4 Experimental framework

We first briefly describe our data, then we report our guidelines and inter-annotator agreement scores for our manual annotation tasks. Finally we present the metrics we use for evaluate the automatic disentanglement.

4.1 Corpus

The corpus was built from logs of the French language Ubuntu's IRC channel dedicated to user support[6]. It is part of an effort for building a multimodal computer-mediated communication corpus

[3] Ubuntu's French language glossary: `http://doc.ubuntu-fr.org/wiki/glossaire`

[4] Ubuntu's French language thesaurus: `http://doc.ubuntu-fr.org/thesaurus`

[5] Ubuntu's English language glossary: `https://help.ubuntu.com/community/Glossary`

[6] irc.freenode.net/ubuntu-fr

in French Hernandez and Salim (2015). The conversations found in the corpus are for the most part task-oriented. It contains 1,229 messages, all of which were manually annotated in terms of conversation and relations (See more details in Section 4.2). The corpus covers 58 different conversations, 12 of which contain only one message, for a total of 46 actual multi-participant conversations (average number of participants: 3.69). These conversations have an average length of 26 messages and a median length of 4 messages. They are interrupted on average 4 times by messages belonging to a different conversation.

4.2 Annotations and agreement

In order to compute inter-annotator agreement, 200 additional messages were annotated in terms of conversation and relation by respectively three and two annotators. Our metric is Cohen's Kappa. The annotators were French native speakers, with background in Linguistic and Natural Language Processing, but had varying levels of annotation experience.

For the conversation annotation task, we give as guidelines the following intuitive definition: Consider as a conversations, the set of utterances

- (whose content are) related to or dependent on a similar context or information need.
- and uttered by the same person or by persons in an interactive situation.

For the relation annotation task, the guidelines were based on the definition given in Section 3.2.

Both annotations tasks were carried out independently on distinct messages.

The results for the conversation annotation task, in table 1, show a very strong agreement between the three annotators.

	A_1	A_2	A_3
A_1	1.0	0.95	0.92
A_2		1.0	0.97
A_3			1.0

Table 1: Agreements for conversation annotation.

For the relation annotation task, we find an agreement of 0.80 when we consider only whether the utterances are related, and of 0.68 when we consider the particular type of each relation. These values corroborate results previously reported in the literature. Identifying a relation's correct type is a difficult task for humans; and we expect the same to be true for machines. Further work will look into the particular situations in which annotators disagree.

4.3 Metrics

We use the same metrics as Elsner and Charniak to evaluate how well different disentanglement methods perform. We use `one-to-one accuracy` to measure the global similarity between the reference and the automatic annotations. It is computed by pairing up conversations from both annotations in a way that maximizes total overlap, and then report it as a percentage. This is useful to estimate whether two annotations are globally matching or not. In contrast, the `local agreement` metric (loc_k) measures the agreement for pairs in a given context of size k. For a given message, each k previous messages are either in the same or a different conversation. The loc_k score is the average agreement of two annotators on these k pairs, averaged over all utterances. It is useful to evaluate local agreement, which is important for the analysis of ongoing conversations.

All scores are computed over 5-fold cross-validation.

5 Experiments and results

First we tried to estimate how relevant the discursive information is for recognizing conversations. Then we evaluate the adaptation of Elsner and Charniak's system for the French language, as well as the addition of discursive features.

5.1 Using discursive relations to recreate conversations algorithmically

We try to determine whether message relations are good indicators of conversational clusters. In order to do that, we project relations into conversations according to the assumption that two related messages belong to the same conversation. We obtain a new set of conversation annotations.

We then compare this new set with the reference annotations. Using the one-to-one metric to measure global similarity we find that this method performs at 0.90 accuracy. Using the loc_3 metric, we find a 0.96 agreement. This shows that discursive relations are highly valuable for the task of chat disentanglement, but also highlights the fact that there can be relations between messages of different conversations.

5.2 Adaptation of Elsner and Charniak's system for the French language

For this experiment we compare the results of our adapted system to Elsner and Charniak's as well as the five following baselines:

- **All different:** each utterance is a separate conversation.
- **All same:** the whole transcript is a single conversation.
- **Blocks of** ***k*****:** each consecutive group of *k* utterances is a conversation.
- **Pause of** ***k*****:** each pause of *k* seconds or more separates two conversations.
- **Speaker:** each speaker's utterances are treated as a monologue.

Results for the adapted system are compared to each baseline in table 2. The third and fourth baselines, "blocks" and "pause", are computed with an optimal k[7].

	one-to-one	loc_3
All different	0.05	0.17
All same	0.25	0.83
Speaker	0.45	0.51
Blocks	0.49	0.83
Pause	**0.71**	0.85
System	0.68	**0.87**
System with relations	0.60	0.84

Table 2: Result comparison with each baseline.

Unlike in the experiments described by Elsner and Charniak, here the system fails to significantly outperform the "pause" baseline. It barely beats it according to the loc_k metric and is best when performance is measured by one-to-one accuracy. However, the system and the best baseline's results are both far higher than those Elsner and Charniak obtained on their corpus. Their results are reported in table 3.

	one-to-one	loc_3
Best baseline	0.35 (Pause)	0.62 (Speaker)
System	**0.41**	**0.73**

Table 3: Results reported by Elsner and Charniak

This discrepancy can be explained by a structural difference between the two corpora. The loc_k metric for the "all same" baseline shows that on a local window, messages usually belong to the same conversation: simply put, our corpus is less entangled than Elsner and Charniak's.

5.3 Addition of relational discursive features

For a different experiment, we add relational features to the classifier. We choose not to consider the specific type of relation, but merely record whether two messages are related. The results are reported in table 2. We find that adding relational features in such a way do not improve the system. This may be explained by the fact that due to the way candidate pairs are selected, the system does not take message relations into account when they are separated by a certain time interval.

6 Conclusion and future work

We adapted Elsner and Charniak's disentanglement system to French and tested it on a chat corpus extracted from the French language Ubuntu platform's main IRC channel. Results were much higher than those reported in the original paper, underscoring the fact that disentangling performances are heavily correlated with how deeply conversations are intertwined in the data. The experiment also showed that a simple heuristic can be as effective as a complex trainable system when conversations are only lightly entangled. Therefore, corpus characteristics should be taken into account in order to choose an appropriate approach.

We also experimented with discursive features in the form of relational information between messages. We found that using such information to algorithmically annotate conversations yielded much more accurate results than the machine learning systems or any baseline. When we tried to use relations as feature to feed the maxent classifier, however, its global performance decreased.

Additional work is required to elaborate a typology allowing for the selection of a corpus' most appropriate disentanglement system. We also plan on performing additionnal experiments making use of the specific types of relations between messages.

[7] Block size is set at 105 for one-to-one accuracy and at 245 for loc_3, and pause time at 240 seconds for both metrics.

Acknowledgments

This material is based upon work supported by the Unique Interministerial Fund (FUI) No. 17. It is part of the ODISAE[8] project. The authors would like to thank the anonymous reviewers for their valuable comments.

References

Harry Bunt. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of the AAMAS 2009 Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts" (EDAML 2009)*, pages 13–24, Budapest, Hungary, 2009.

Micha Elsner and Eugene Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08)*, pages 834–842, Columbus, OH, USA, 2008.

Micha Elsner and Eugene Charniak. Disentangling chat. *Computational Linguistics*, 36(3):389–409, 2010.

Nicolas Hernandez and Soufian Salim. Construction d'un large corpus libre de conversations écrites en ligne synchrones et asynchrones en français à partir de ubuntu-fr. In *The first international research days (IRDs) on Social Media and CMC Corpora for the eHumanities*, Rennes, France, 2015.

Nam Su Kim, Lawrence Cavedon, and Timothy Baldwin. Classifying dialogue acts in multiparty live chats. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 2012)*, pages 463–472, Bali, Indonesia, 2012.

Elijah Mayfield, David Adamson, and Carolyn Penstein Rosé. Hierarchical conversation structure prediction in multi-party chat. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2012)*, pages 60–69, Seoul, South Korea, 2012.

Lidan Wang and Douglas W Oard. Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*, pages 200–208, 2009.

[8] `http://www.odisae.com/`

Text-based Geolocation of German Tweets

Johannes Gontrum **Tatjana Scheffler**
Department of Linguistics
University of Potsdam, Germany
gontrum,tatjana.scheffler@uni-potsdam.de

Abstract

We show a new, data-driven method for geolocating single tweets based on the geographical variance of their tokens. While more than half of German tweets do not contain reliable textual indicators of their location, our method can locate 40% of tweets very accurately, up to a distance of 7km (median) or 93km (mean).

1 Introduction

Twitter data is interesting for many NLP applications because of its abundant metadata. This includes geolocation data (GPS coordinates), indicating where the tweet's author was located at the time of writing. Geolocation information is important for the detection of regional events, the study of dialectal variation (Eisenstein, to appear 2015), and many other possible applications. However, not all users allow the public distribution of their location data, and in some language communities, geolocated tweets are very rare. For example, only about 1% of German tweets contain a location, and these come from an even smaller number of users that allow this feature (Scheffler, 2014).

In this paper we introduce an approach to recover a geolocation of origin for individual tweets using only the text of the tweet. This allows the enrichment of Twitter corpora that do not contain sufficient geo information, even for unseen users or users who never share their location. This is important since many users (e.g. in Germany) use made-up or false locations in their user profile field. We use geo-tagged tweets in order to derive a lexicon of regionally salient words, which can then be used to classify incoming tweets.

2 Related Work

Geolocation of Twitter messages can be based on the user's location as indicated in the profile, or a tweet's GPS location. Text-based geolocation does not take user information into account. Previous approaches however commonly aggregate all of a user's tweets (Cheng et al., 2010; Wing and Baldridge, 2014) or conversations including replies (Chandra et al., 2011) to determine one location. Some researchers have instead attempted to directly derive location-specific words or dialectal variation from geotagged tweets (Eisenstein et al., 2010; Eisenstein, to appear 2015; Gonçalves and Sánchez, 2014), using GPS locations or user profile locations.

(Pavalanathan and Eisenstein, 2015) compared the data sets obtained by user profile and GPS geolocation of tweets, respectively, and show that they differ significantly with respect to demographics and linguistic features. (Graham et al., 2014) show that user profile information is only rarely a reliable indicator of the location of the user, more than half of profiles containing empty location fields, unhelpful locations ("earth") or diverging user profile and GPS information.

In a previous paper (Scheffler et al., 2014), we first attempted to geolocate individual tweets based only on that tweet's text, using predefined "dialect" regions in Germany as our goal. In that work, we also discussed a thesaurus-based approach using an existing list of known dialectal words as seed words. That approach was vastly inferior to a method that automatically induces regionally salient words from geo-tagged tweets. The current paper shows a completely new, data-driven solution to that problem.

3 Approach

It is important to note that there are at least two distinct sources for regionally distinctive language in a tweet: (i) the current location of the author, which leads to the use of local event and place names, and (ii) the dialectal region of origin of the author, which yields regionally salient dialectal expressions. In principle, these two sources are independent of each other (think of an Austrian travelling to Berlin). However, using current methods, neither we nor any of the previous work can systematically distinguish these two types of geographic origin of a tweet. In this work, we assume that for statistical purposes, most users are located close to their region of origin and thus do not address this problem further. However, this may lead to discrepancies in individual cases where a user is either travelling or writes about a distant location.

Further, for evaluation purposes we regard the GPS metadata information provided by Twitter as gold location data for our corpus. This is in line with previous approaches, but potentially biases the algorithm towards case (i) above – the current location of the tweet author. Dialect origin information is a lot harder to obtain, but could potentially be gathered through surveys or in an unsupervised or bootstrapping manner.

Data Our corpus consists of 65 mio. tweets that have been collected through the Twitter API between February and May 2015, by filtering the Twitter stream using a keyword list of common German words (Scheffler, 2014). Language identification was carried out using LangID (Lui and Baldwin, 2012). Further, we extracted only tweets that were geo-tagged and located in Germany, Switzerland or Austria. To remove bots we manually created lists of suspicious user ids and ignored messages containing the words 'nowplaying' or '4sq'. We tokenized the lower-cased tweets, removed numbers, URLs, user-mentions and most special characters. After removing the '#'-character, hashtags remain in the tweets since they can provide useful information about local events. Only 360k tweets (0.55% of all collected documents) fulfilled our criteria. We then randomly extracted 1000 messages each for testing and development.

Background Our method is based on the observation that tokens are not used at all locations with the same frequency. Hence, there must exist a function that describes the probability that a token is used in a tweet at given coordinates. Additionally, we assume that these probabilities are distributed around a specific location at which the probability of the token is the highest.

We have discovered that in contrast to common words that are used uniformly throughout, regional words like city names are used in an area with a diameter of 50-150km by many users. The highest level of information is provided by local terms denoting for example local events or street names that are only used a few times, but at a very narrow location. This distinction can be observed by printing the probability density function (PDF) of the tokens in tweet (1), see Figure 1.

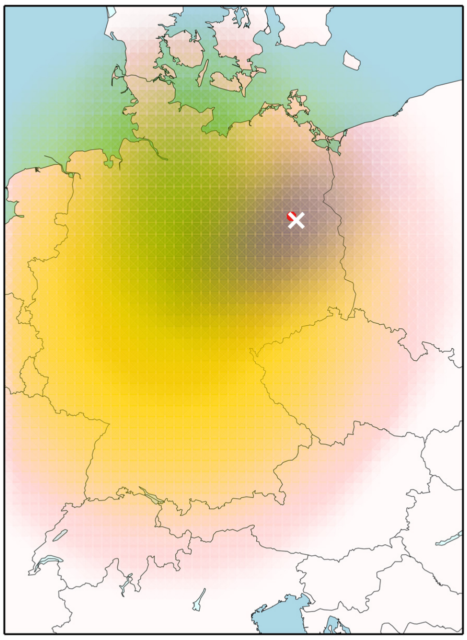


Figure 1: PDF of tokens in tweet (1): regional words *berlin* (blue), *hhwahl* (green); highly local word *nordbahnhof* (red); common words (yellow). Position of the tweet marked by a white cross.

(1) *balken gucken und so hhwahl pa nordbahnhof in berlin*

The common tokens (*balken, gucken, und, so, pa, in*) are drawn in yellow. They are so widely distributed that they cover the whole of Germany and are not providing any local information that could help classify the tweet. The regional word *hhwahl*, denoting an election in Hamburg, is illustrated in green and the density function of the other regional word *berlin* is drawn in blue around the location of the city. Finally, the word *nordbahnhof* has only been observed close to that station in Berlin and is therefore a local word (red). In fact, the tweet was sent within a distance of only 4km.

Classification method The tweet in (1) illustrates the importance of finding a parameter to distinguish common and widespread words from regional and local tokens. Additionally, we need a method to weight the remaining tokens so that highly local words are given more significance than less concrete regional words. We use the variance of the probability distribution of a token as a score that can be used to solve both our problems.

Since the variance describes how widespread the data points are, regional or local words that appear only in a small area will have a low variance, while variance is high for common words or even low-frequent words like typos that are not regionally biased. First, we use the variance as a threshold to remove common words from tweets and calculate the geographical midpoint of the remaining tokens. We found that the median for a token position outperforms the mean especially for infrequent terms, since it marks an actual coordinate where the token was used.

An analysis of our data reveals the importance of low-variance local terms. If a tweet contains one of these highly local tokens, the tweet's position is almost entirely determined by that token's median position and any influence of other tokens would worsen our score. Secondly, we therefore weight the individual tokens by their inverse variance σ^{-1}, so that very local tokens receive an extremely high score and overshadow all other words. If a tweet on the other hand contains exclusively regional words, their inverse variance is not too high so all of them have an influence on the position.

Algorithm The median position and the variance for each token in a tweet is calculated based on the coordinates of all tweets in the training corpus in which they are used. Note that we are converting the longitude and latitude information provided by Twitter to three-dimensional Cartesian coordinates. Since longitude and latitude are projections on a sphere, the calculation of midpoints and distances becomes less error prone this way. Therefore, we are from now on regarding median and variance values as vectors.

Equation (2) shows the calculation of the location of a tweet t with tokens $t_0, ..., t_n$, their variance values $\vec{\sigma_0}, .., \vec{\sigma_1}$ and their median $\vec{m_0}, .., \vec{m_1}$.

$$Loc(t) = \frac{\sum_{i=0}^{n} \vec{\sigma_i}^{-1} * \vec{m_i}}{\sum_{i=0}^{n} \vec{\sigma_i}^{-1}} \qquad (2)$$

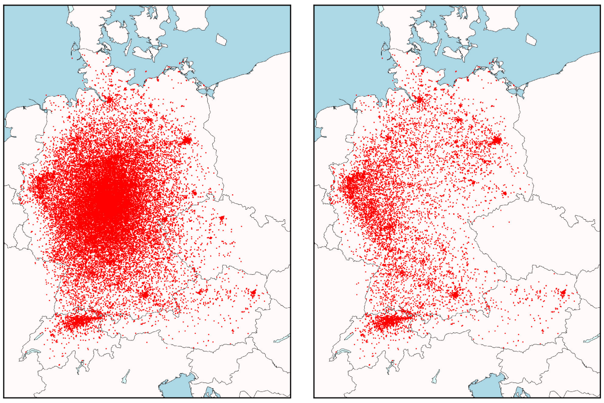


Figure 2: Left: Mean coordinates of all tokens. Right: Only regional tokens under the assumption that 25% of all tokens are regionally salient.

4 Results and Discussion

Filtering Step It is clear that some tweets are unsuitable for geolocation using only their text. This is due to the fact that a majority of tokens are so common that they carry no information about any location whatsoever. As a consequence, the original position of tweets that contain only these irrelevant tokens cannot be recovered from the text alone. To make things worse, any attempt to do so will lead to unjustified confidence in the calculated position and will result in an unreliable algorithm.

Figure 2 shows the mean coordinates for all tokens in the corpus on the left, while in the right graphic only the top 25% of tokens (by lowest variance) remain. The blob in the center of Germany are those meaningless tokens that are removed with a decreasing variance threshold. For this reason, we are deliberately filtering a number of tokens that are lacking reliable information and consequently accept a high amount of unclassifiable tweets for the sake of accuracy.

Experiments The determination of a variance threshold for common words can be seen as an estimate of the ratio of regional tokens in the corpus. For example, a threshold of 30% means that we regard the 30% of the tokens with the lowest variance as regional and remove all other words. Figure 3 displays these scores for different parameter estimates of the percentage of regional words (x-axis). As expected, the geolocation error (measured in distance to the true location) decreases

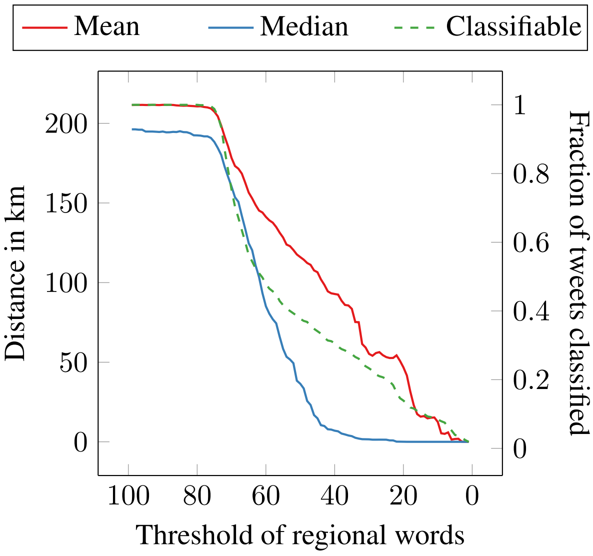


Figure 3: The mean and median distance in *km* between the predicted and the true metadata coordinates of a tweet.

with a stronger threshold, as the amount of unclassifiable tweets grows. We can make out three stages that correspond to our classification of tokens: The first notable improvement of the score happens when the most frequent of the common words are removed at about 70%. In the next stage widespread regional words are gradually removed and at about 30%-40%, most tweets rely exclusively on local words.

Even though the distance median drops below 10km at 43%, the mean distance stays relatively high. We explain this gap by a few tweets whose predicted location is hundreds of kilometers away from their true metadata position. As discussed above, this can happen either when tweets mention distant events or locations, or when people travel away from their dialect regions and use dialectal expressions in tweets. Since we compare the predicted location with the GPS metadata from Twitter (our "gold" data), our method cannot avoid these problems. On the other hand, some tokens are wrongly classified as local or regional due to their infrequent appearance in our small training corpus and therefore the accuracy will increase with a bigger data set. Table 1 shows the geolocation errors as well as the number of classified tweets for different variance parameter thresholds of regional words.

Threshold	Mean	Median	#Tweets
100	212km	196km	1000
75	207km	188km	988
50	116km	36km	377
40	93km	7km	306
30	55km	1.56km	233
20	47km	0.06km	139
10	12km	0.00km	84

Table 1: Results of geolocation algorithm for different variance estimates: "Threshold"=ratio of 'regional' words (by variance), error distances to the true location, and number of classified tweets (N=1000) are given.

We have also analyzed which tokens are classified as local or regional for certain cities, as shown in Table 2. In Berlin and Essen for example, mostly street or district names are revealing, while in Zurich dialectal words are dominating.

Finally, we created a score to compare our results to the ones from a previous paper (Scheffler et al., 2014), where the German speaking area was manually divided into seven regions, and success was measured by the percentage of tweets correctly classified into these regions. To achieve a rough comparison, we used a clustering algorithm on randomised data to create seven regions that cover an equally large area. In (Scheffler et al., 2014) a threshold was used to remove common words and only 20% of all tweets were classified, resulting in 53% correctly classified tweets. When adjusting our method to this threshold, we accurately classify 86% of tweets into the correct region, a large improvement. However, since the previous paper used a different dataset, the results are still not directly comparable.

In summary, this paper introduces a new, language independent, highly accurate approach to geolocating single tweets based on the geographical variance of words in the corpus. The method can be further augmented by user-oriented approaches in order to improve recall.

5 Future Work

The task opens up many avenues for future research. Most importantly, the differentiation of the two essentially distinct sub-tasks – identifying the location and dialect origin of the author – must be addressed, although this will require

Berlin	Zurich	Essen
kadewe	tagi	rheinische
kudamm	uf	hattingen
alexanderplatz	het	herne
friedrichshain	isch	westfalen
brandenburg	scho	ddorf
fernsehturm	au	ruhr
dit	zuerichsee	thyssenkrupp
morjen	gseh	duisburg

Table 2: Notable local tokens with low variance and high frequency in Berlin, Zurich, and Essen.

more complex models. A resource for location words such as OpenStreetMap might help here. Another obvious improvement, also suggested by a reviewer, is the training of the words' significance weights by machine-learning methods (instead of fixing them to the variance). Finally, it is still unclear how much the algorithm overfits to certain frequent and predictable tweeters, like bots. Frequently-tweeting bots may on the one hand hurt performance, since the model falsely associates all its words with the bot's location. On the other hand, this may also help if the test data also includes tweets from the same source. This behavior can be tested by evaluating the system on sufficiently different material (e.g., from a different point in time (Rehbein, p.c.)), and mitigated by developing methods to exclude non-natural tweets during preprocessing.

Acknowledgments

The authors would like to thank Matthias Wegel and Steve Wendler for their involvement in previous versions of this work. We are grateful to Ines Rehbein and the anonymous reviewer for their detailed comments on our work and suggestions for future improvements.

References

S Chandra, L Khan, and F B Muhaya. 2011. Estimating Twitter User Location Using Social Interactions–A Content Based Approach. In *IEEE Third International Conference on Social Computing (SocialCom)*, pages 838–843, October.

Z. Cheng, J. Caverlee, and K. Lee. 2010. You are where you tweet. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, page 759.

J. Eisenstein, B. O'Connor, N. Smith, and E.P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287.

J. Eisenstein. to appear 2015. Identifying regional dialects in online social media. In *Handbook of Dialectology*.

B. Gonçalves and D. Sánchez. 2014. Crowdsourcing dialect characterization through twitter. *PLOS One*, 9(11):1–10.

M. Graham, S. Hale, and D. Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*.

M. Lui and T. Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Republic of Korea.

U. Pavalanathan and J. Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data.

T. Scheffler, J. Gontrum, M. Wegel, and S. Wendler. 2014. Mapping German tweets to geographic regions. In *Proceedings of NLP4CMC workshop at the 12th KONVENS.*

T. Scheffler. 2014. A German Twitter snapshot. In N. Calzolari et al., editor, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

B. Wing and J. Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 336–348.

Modes of Communication in Social Media for Emergency Management

Sabine Gründer-Fahrer
University of Leipzig, InfAI
Augustusplatz 10
04109 Leipzig, Germany
gruender@uni-leipzig.de

Antje Schlaf
University of Leipzig, InfAI
Augustusplatz 10
04109 Leipzig, Germany
antje.schlaf@informatik.uni-leipzig.de

Abstract

The paper examines how social media were used during the flood 2013 in Central Europe and what differences in use appeared among different kinds of media. We found that Twitter played its most important part in exchange of current and factual information on the state of the event while Facebook prevalently was used for emotional support and organization of volunteers help. In a corpus-based comparative study we show how the different communicative modes prevalent in the registers German Facebook, Twitter and News are clearly reflected by the characteristic content, conceptualization and language of the respective register. The methods used include differential analysis, sentiment analysis, topic modeling, latent semantic analysis and distance matrix comparison.

1 Introduction

From the point of view of emergency management, a crisis has three basic dimensions: 1. the real event; 2. the actions of the involved organizations; 3. the perception of the crisis (BMI, 2008). In each of these dimensions, communication plays a central role. Social media have high potential to improve quality of communication in all three dimensions, and systematic usage of the new possibilities has just begun. Following the rise of social media, "Emergency management, once the province of official channels, is going where the people are." (Yasin, 2010) Using social media as output channel, emergency managers can reach a wider audience and pass information directly and more quickly (dimension 1); social media enables interaction with affected people and cooperation with volunteers (dimension 2); input from social media can improve situational awareness of the emergency managers (dimension 1), and it allows for tracking of activities of volunteers and monitoring of opinions and moods (dimension 3). Social media, though, also include the potential to harm emergency management, for instance, by spreading erroneous information or propagating negative mood or panic.

But there are two basic challenges emergency management has to face when using social media. First, with the medium changing, the structure and culture of communication is changing as well and new modes of communication arise together with new contents. Second, in the wake of digitalization of communication, all so-far disparate branches of media are converging into one multi-modal, multi-medial, multi-lingual, multi-cultural digital room which consequently is going to contain an overwhelming amount of information of great diversity and makes the different types of media competing. Consequently, if emergency management wants to exploit the positive and control the negative potential of social media, it first has to know what kind of information is available in which medium, and there is high need for appropriate computer-based tools for searching, sorting and analyzing of relevant data.

The paper is going to contribute to the field of social media research by investigating the following questions:

1. What kind of content is distributed through social media in context of a disaster?

2. Are there differences in content among several types of social media?

3. What are the differences in content to other public digital media?

The results gained are relevant for several research disciplines, mainly corpus-based variational linguistics, communication studies, natural language processing, and crisis informatics.

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 33–37, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

2 Data and Methods

As our case study, we chose the historical flood event in spring 2013, which heavily affected large areas of Central Europe, mainly south and east German states, Austria, and western regions of Czech Republic. Covering the time span from May to June/July 2013, we have collected German data in three different types of media (registers): Facebook, Twitter and News. For each medium there was created a corpus of flood-related documents together with a general reference corpus. Standard pre-processing included deletion of stop words, numbers, punctuation and lemmatization.

The Facebook flood corpus, comprising 35.6k messages (1.2M word tokens), was collected from 264 public Facebook pages or groups containing the words *Hochwasser* “flood” or *Fluthilfe* “flood aid” in their names. A Facebook reference corpus of 1.7M messages (42.8M tokens) was semantically balanced using the category system of the “Socialbakers” online platform and collecting public messages from the top 10 ranked public pages or groups (regarding fans) in each category.

For the Twitter flood corpus we retrieved a current version of the research corpus of the QuOIMA project (QuOIMA, 2013), that was collected from the public Twitter stream and filtered by 65 hashtags selection of 29 accounts. The current version comprises 354k tweets (4M tokens). A reference corpus of 1.8M tweets (14M tokens) consists of 1 percent of the public twitter stream from March to May 2015.

Using the RSS feeds news corpus “Wortschatz” from the University of Leipzig NLP group (Quasthoff et al., 2013) we created the flood news corpus by choosing 10.3k documents (3,6M word tokens) published during the time of the flood event and containing the keyword *Hochwasser* “flood” somewhere in the text. An alternative news flood corpus (used for the topic model analysis) was built on basis of a topic model trained on the complete set of documents from the same collection and time. We chose the top 9.5k documents (2,8M words) that have the highest probabilities in the derived flood topic (threshold 0.1). As the news reference corpus we used the 1.1M documents (212M tokens) from 2012 from the same collection.

As a supplementary corpus for comparison we built a small corpus from 30 professional reports (475k tokens) that have been collected manually from public websites of emergency management organizations in Germany and Austria.

In the course of our investigation we applied the following methods:

Differential Analysis: We analyzed differences in relative frequency of terms from one corpus to the other using Log-Likelihood Ratio Test (Dunning, 1993). In a first test series, we compared flood corpus and reference corpus within each register to find flood-related content, respectively. In a second series, we did mutual comparison of equal sized (samples of) flood corpora from different media aiming at differences in semantic focus between the registers when talking about the same event.

Sentiment Analysis: In a comparative study of emotional involvement in flood-related messages in different registers, we compared relative frequencies of sentiment words for each case using the SentiWS resource from the University of Leipzig NLP group (Remus et al., 2010), a list of German positive and negative sentiment bearing words.

Topic Models: In order to reveal and cluster content in the flood corpora in each register we tested Topic Modeling techniques and finally applied a Hierarchical Dirichlet Process in form of a Chinese Restaurant Franchise Sampler (HDP CRF) (Teh and Jordan, 2010) . Topic modeling infers not directly observable variables considered as latent topics. Another hidden variable describes each of those topics in form of a probability distribution over the vocabulary of the text collection. These weighted topic words allow for an intuitive interpretation of the inferred topics.

Latent Semantic Analysis and Distance Matrix Comparison: Finally, we wanted statistically measure similarity between the flood corpora in the different registers. For each flood corpus (sample), a term-document matrix was constructed and weighted (tf-idf) and projected into an lower-dimensional space using Latent Semantic Analysis techniques (LSA) (Deerwester et al., 1990). From that, a distance matrix was computed on the basis of Euclidean Distance Measure.

3 Results

3.1 Facebook

As a first semantic focal point in the flood-related Facebook corpus, differential analysis with respect to the Facebook reference corpus and topic modeling alike extracted lexical clusters including information on the ***state of the event*** (*Hochwasser* “flood”, *Pegel* “water gauge”, *Deich* “dike”) and ***ac-***

tivities of organizations (*Feuerwehr* "fire brigade", *Einsatz* 'operation"). The information are rather general in nature, but as a variation of the topic there appears a cluster for ***direct effects on social life*** (*Schule* "school", *geschlossen* "closed", *Strasse* "road", *gesperrt* "closed"). However, as the most dominant topics of the Facebook flood corpus and as the striking difference to the flood corpora of the other registers (Twitter, News, Professional), there appear the following clusters: ***empathy and social interaction*** (*bitte* "please", *dringend* "urgent", *helfen* "help", *danke* "thank you", *Liebe* "love", *Leid* "suffering", *Opfer* "victims"); ***volunteers help*** (*Sandsäcke* "sandbags", *Helfer* "helpers", *gesucht* "wanted", *wer* "who"); ***donations*** (*spenden* "donate", *Konto* "bank account", *BLZ* "bank code number", *unterstützen* "support"), and ***donations in kind*** (*Kleidung* "clothes", *benötigen* "need", *Sammelstelle* "collection point", *Sachspenden* "donations in kind"). The aspect of emotional support also is revealed by sentiment analysis (Table 1), which shows much higher relative frequencies for positive sentiment markers in Facebook than in any other flood corpus and a greater positive-negative ratio. The most significant characteristic terms of the corpus compared to the other flood corpora are shown in Table 2. (For differential analysis, all location markers where mapped on the string "locationCity".)

Flood	positiv	negativ
Facebook	0.115	0.044
News	0.081	0.046
Twitter	0.075	0.023

Table 1: Sentiment

3.2 Twitter

In comparison with the topical focus of the Facebook flood corpus, the focus of the Twitter flood corpus is just switched. The main focus is on current information on the event; social engagement is also present in this register but appears as subordinated. Remarkably, current information on the event occur in different versions and as separate topics, respectively. Beside a topic with ***general information on the current state of the event*** (*Wasser* "water", *Pegel* "water gauge", *gestiegen* "risen"), there is one topic which includes very ***precise, objective, technical information on weather conditions*** (*Druck* "pressure", *Feuchte* "humidity", *kmh* "kilometer per hour"). More ***dynamic aspects of the situation*** form a separate cluster (*steigend* "increasing", *Alarmstufe* "alert level", *Tendenz* "tendency") that are connected with particularly severe states of the event. These two topics with objective stative or dynamic reports contrast with a more ***subjective weather topic*** where emotional involvement with respect to the current situation becomes visible and is reported in very informal style (*Dauerregen* "continuous rain", *scheiß* "shit", *endlich* "finally"). A more general style of reporting on the event occurs in two topics that include a ***larger geographical context*** (*Sachsen* "Saxony", *Passau* "*Bavarian city*") or ***social context*** (*Hochwasserhilfe* "flood aid", *Merkel* "*name of German chancellor*"). Furthermore, there appears a topic which reports on the ***activities of the public and volunteers organizations*** and asks for support (*Feuerwehr* "fire brigade", *Einsatz* 'operation", *Helfer* "helper", *gesucht* "wanted"). Related to that is a topic which includes ***warnings*** (*Wetterwarnung* "weather alert"). Finally, a separate topic includes links to ***further information sources and breaking news*** (*live-ticker*, *Katastrophenalarm* "red alert"). Differential analysis re-

Facebook		
vs. News	vs. Twitter	vs. Prof.
bitte	helfen	helfen
bitten	bitten	bitten
helfen	melden	locationCity
benötigen	bitte	malen
uhr	gerne	helfer
helfer	benötigen	bitte
hilfe	sachspende	uhr
gerne	abgeben	hilfe
melden	sache	melden
spenden	gruppe	leute
malen	gebrauchen	spenden
spende	hilfe	heuen
gebrauchen	ort	gerne
quell	helfer	spende
sachspende	leute	benötigen
dringen	spende	gebrauchen
leute	verfügung	dringen
abgeben	kind	fahren
aken	betroffen	schonen
heuen	wissen	sachspende

Table 2: Differential Analysis: Facebook flood vs. the other flood corpora, top 20 significant words.

vealed the most characteristic terms of the Twitter flood corpus in Table 3. Accordingly, the distinctive feature of the twitter flood corpus when compared to the registers Facebook and News is the precision and objectivity of most of its state reports. From the professional reports it is separated by concreteness of reports (i.e., many location names) and less formal style. But, generally, the lexical closeness between Twitter and professional reports is quite appealing, as can be seen from Table 4.

Twitter		
vs. Facebook	vs. Twitter	vs. Prof.
hochwasser	hochwasser	hochwasser
stand	stand	locationCity
pegel	locationCity	stand
locationCity	pegel	elbe
kmh	kmh	pegelmv
hpa	elbe	hpa
pegelmv	pegelmv	kmh
elbe	hpa	wind
wind	wind	unwetter
unwetter	rege	rege
rege	unwetter	temp
temp	temp	pegel
konstant	konstant	minute
doemitz	doemitz	konstant
luftdruck	tendenz	doemitz
dauerregen	luftdruck	min
lm^2	minute	tendenz
minute	min	luftdruck
untere-havel-wasserstrasse	lm^2	uhr
feuchte	untere-havel-wasserstrasse	sonne

Table 3: Differential Analysis: Twitter flood vs. the other flood corpora, top 20 significant words.

Flood	Facebook	News	Prof.	Twitter
Facebook	0			
News	641.2	0		
Prof.	388.9	424.4	0	
Twitter	343.2	408.8	56.6	0

Table 4: Lexical Similarity

News		
vs. Facebook	vs. Twitter	vs. Prof.
euro	jahr	euro
jahr	prozent	mensch
prozent	euro	locationCity
million	meter	schonen
überfluten	million	meter
milliarde	groß	stehen
fluss	liegen	mehren
schaden	stehen	merkel
merkel	angabe	flut
bund	mehren	dienstag
meter	sprecher	wasser
cdu	wasser	häuser
land	erklären	woche
betroffene	mensch	geld
niedersachse	insgesamt	malen
donau	dienstag	helfen
deutschland	gemeinde	prozent
dpa	häuser	million
erklären	mitteilen	heißen
spd	konnt	montag

Table 5: Differential Analysis: News flood vs. the other flood corpora, top 20 significant words.

3.3 News

In case of the News flood corpus, there are general descriptions of the state of the event in coarser or finer local and temporal granularity: ***supra-regional overview*** (*Österreich* “Austria”, *Bayern* “Bavaria”, *Sachsen* “Saxony”) or ***region and local*** (*Stadt* “city”, *Görlitz* “*name of town*”, *Straße* “street”, *Sonntag* “Sunday”). When describing the ***activities of the public organizations***, News reports differ from Twitter and professional reports by taking the perspective of the common people and add evaluations from a general public point of view (*Einsatz* “operation”, *gut* “good”, *Keller* “basements”, *Häuser* “houses”, *verlassen* “leave”, *Familie* “family”). The most characteristic feature of the News flood corpus, though, is the wide range of content coming from inclusion of broader context in several topics: ***finances*** (*Euro*, *Million en* “millions”, *Schaden* “damage”), ***society*** (*Fond* “fund”, *Maßnahmen* “measures”, *Projekt* “project”), ***politics*** (*Merkel* “*name of German chancellor*”, *CDU* “*German party*”, *Bürgermeister* “mayor”), ***traffic*** (*Straße* “street”, *gesperrt* “closed”, *Zug* “train”, *Verspätung* “delay”), ***nature*** (*Mückenplage* “mosquito plague”, *Biber* “beaver”). Furthermore, as the only one among the registers studied, News includes ***retrospective, discussion, and evaluation*** (*Maßnahmen* “measures”, *wohl* “arguably”, *bisher* "up to now", *Jahre* "years", *gut* "good"). The most significant differentiating terms of the News flood corpus are shown in Table 5.

4 Conclusion

Question 1: In our case study of the flood 2013 in Central Europe, German social media proved relevant for all three dimensions of a crisis mentioned at the start. Interestingly, there appeared something like a division of labour between two different social media platforms under investigation. With respect to the real event, social media were actively used for sharing of up-to-date information with broad public and, thereby, contributed to improvement of situational awareness. To this dimension, Twitter contributed more and more precise information than Facebook. In the dimension of the activities of involved organizations, they played an important part in the organization of volunteers activities and donations. In this dimension, Facebook was much more intensively used than Twitter. As for the perception of the crisis, social media were used to directly show empathy or for emotional (self-) management. This dimension was dominated by Facebook.

Question 2 and 3: Accordingly, when comparing social and other public digital media, the different prevalent communicative modes are clearly reflected by the characteristic content, conceptualization and language of the respective register. According to our observation, the focus of Facebook content is on empathy and emotions and on social engagement. The conceptualization generally takes the perspective of the affected people and the language is emotionally-involved and informal. Twitter, in contrast, is mainly used for exchange of current and concrete information on the event, and takes a more factual point of view on the event. The characteristic language is situative reporting and factive, stylistically ranging from quite technical to colloquial. Finally, News is the medium that allows for inclusion of a broader social context and for public debate. It takes the perspective of the general public. The characteristic language features of this register reflect the communicative modes of informing and discussing, and a more general linguistic style. Lexical distance between News and both kinds of social media is larger than the distance between the two social media.

The extracted keywords and topics will enable automatic filtering of relevant content in social media. Gained information on the type of language used can serve as a basis for correct setting of parameters and adaptation of methods for processing and analyzing the data. Improved tools will allow emergency managers to better use and control the potential of social media.

Acknowledgments

This work was carried out as part of the Slándáil project. The project is receiving funding from the European Union's Seventh Framework Programme for Security under the Grant number 6076921.

For support on the topic model analysis we thank our colleagues Patrick Jänichen and Dieu Thu Le.

References

BMI - Bundesministerium des Inneren. 2008. *Krisenkommunikation - Leitfaden für Behörden und Unternehmen*, Online: www.bmi.bund.de/SharedDocs/Downloads/DE/Broschueren/2008/Krisenkommunikation, 25.05.2015.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer und Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American society for information science*, 41(6):391–407.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.

Uwe Quasthoff, Dirk Goldhahn and Gerhard Heyer 2013. Deutscher Wortschatz 2012. Technical Report Series on Corpus Building 1. Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig.

QuOIMA. 2011. *QuOIMA - Open Source Integrated Multimedia Analysis*. Online: www.kiras.at/gefoerderte-projekte/detail/projekt/quoima-quelloffene-integrierte-multimedia-analyse, 25.05.2015.

Robert Remus, Uwe Quasthoff and Gerhard Heyer 2010. SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. *Proceedings of the 7th International Language Resources and Evaluation (LREC)*

Yee Whye Teh and Michael I. Jordan. 2010. Hierarchical Bayesian nonparametric models with applications. In: Nils Lid Hjort et al. (eds.) *Bayesian Nonparametrics.*,114–133. Cambridge University Press, Cambridge, UK.

Rutrell Yasin. 2010. 5 ways social media is changing emergency management Online: www.gcn.com/articles/2010/09/06/ social-media-emergency-management.aspx, 25.05.2015.

Unsupervised Induction of Part-of-Speech Information for OOV Words in German Internet Forum Posts

Jakob Prange and **Stefan Thater** and **Andrea Horbach**
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{jprange,stth,andrea}@coli.uni-saarland.de

Abstract

We show that the accuracy of part-of-speech (POS) tagging of German Internet forum posts can be improved substantially by exploiting distributional similarity information about out-of-vocabulary (OOV) words. Our best method increases the accuracy by +15.5% for OOV words compared to a standard tagger trained on newspaper texts, and by +12.7% if we use an already adapted tagger.

1 Introduction

A major challenge in the automatic linguistic processing of data from computer-mediated communication (CMC) is often the lack of appropriate training material. Tools like part-of-speech (POS) taggers are usually trained on and optimized for edited texts like newspaper articles, and their performance decreases substantially when applied to out-of-domain CMC data. The tagger used in our study, for instance, achieves an accuracy of 97.2% when trained on and applied to German newspaper text; when applied to posts from an Internet forum, performance goes down to 85.0%.

One important reason for this decrease in performance is that CMC texts often contain out-of-vocabulary (OOV) words which the tagger has not seen during training. Consider the following example from the Internet forum *www.chefkoch.de*:

(1) Bei mir gab ***kabeljau ihh*** also manche ***fische*** mag ich ***irklich*** nicht ***aba rollmops*** mit **gebackene** ***kartoffeln*** und das ist **leckerer**!

The words in boldface are unknown to the tagger. They range from misspellings (*[w]irklich*), action words or interjections (*ihh*) to creative new word formations or deliberate orthographical variation (*aba* instead of *aber*) up to words that are perfectly acceptable but were not covered in the training material (*leckerer*) due to domain differences between test and training data. Words that are mis-tagged by an out-of-the box tagger model are printed in italics. We can see that, in this case, the mis-tagged words are a subset of the unknown words. Apart from this example, the frequency of mis-tagging is generally high and the percentage of mis-taggings is dramatically higher within the unknown words.

In this paper, we explore different methods to automatically induce possible POS tags for OOV words and compare different ways to exploit this information in a POS tagger. More precisely, we explore the idea that distributionally similar words tend to belong to the same lexical class and thus their POS tags can be used to induce possible POS tags of OOV words. We evaluate several ways of integrating this information into a POS-tagger: As a post-processing step, as an additional lexicon of a HMM-based tagger and as features in a CRF-based tagger. Our best approach increases the accuracy for OOV words by +15.5% for a tagger trained on standard newspaper text, and by +12.7% for an already adapted tagger.

2 Related Work

The problem that CMC texts usually contain many OOV words can be addressed in several ways. One can normalize the input text by mapping OOV words to known words in a preprocessing step, correct the POS tags of OOV words after tagging in a post-processing step, or adapt the tagger itself so that additional knowledge about possible POS tags of OOV words can be used directly during tagging.

The first two options have been explored *e.g.* by Gadde et al. (2011), who use word clusters based on string similarity to relate OOV words to known words and obtain an improvement of 4.5% over the baseline tagger on a small SMS corpus.

The third option has been investigated, amongst others, by Rehbein (2013), who trains a CRF-based

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 38–43, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

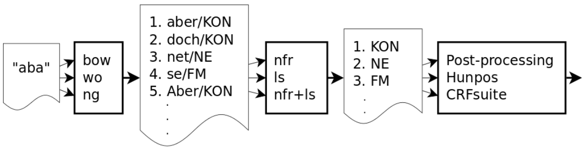


Figure 1: Example run of our pipeline with the OOV word “aba” (“aber”).

tagger for German Twitter tweets on features derived from word clusters, an automatically created dictionary for OOV words and additional out-of-domain training data. The tagger achieves an accuracy of 89% on a corpus of 506 German tweets. (See Owoputi et al. (2013) for using cluster features for English data.)

While we also use a CRF-based tagger in our experiments, our approach is more closely related to the work of Han et al. (2012), who use a combination of distributional and string similarity to induce a normalization dictionary for microtexts from Twitter. The main difference is that we use the normalization dictionary only indirectly to learn possible POS tags for OOV words.

3 Our Approach

The key idea underlying our approach is that distributionally similar words tend to belong to the same lexical class and thus their POS tags can be used to induce possible POS tags of OOV words. Figure 1 describes the workflow of our approach in more detail: Given an OOV word such as *aba*, we compute the list of 20 distributionally most similar known words together with their POS tags. Based on this list of similar words we then create a lexicon that lists possible POS tags of OOV words, which we use to increase tagging accuracy of OOV words in different ways.

Distributional models. We consider three different distributional models to compute similarity scores, which we train using the *chefkoch* dataset described in Section 4 below. We tag the dataset using the *hunpos* POS tagger (Halácsy et al., 2007) trained on the *Tiger* corpus (Brants et al., 2004) and use a sliding window approach to count frequencies of context words, using a fixed window size of ± 2 words around the target word. We restrict ourselves to contexts where all context words are known to the tagger; the target word itself can be OOV, in which case we replace the POS tag assigned by the tagger by the pseudo tag *X*.

We consider (i) a standard bag-of-words model (*bow*), (ii) a variant of the bow model where context words are indexed by their relative position to the target word (*wo*), and (iii) a model where we use 5-grams of the form $\langle t_1, t_2, *, t_3, t_4 \rangle$, where the t_i are the POS-tags of the context words (*ng*). In all cases, we use PMI scores derived from the frequency counts as weights in the word vectors.

POS-Lexicon. In order to induce a ranked list of possible POS tags of OOV words, we first compute a *candidate list* containing the 20 known words with the highest similarity scores to the OOV word, taking scalar product between the word-vectors of the respective model (*bow*, *wo*, *ng*) as similarity measure. Then, we extract all POS tags that occur in the candidate list and rank the tags using different methods. We report results for the following approaches:

n-first-ratio (nfr): POS tags are ranked based on the ratio of their frequency in the candidate list and the index at which they first occur.

Levenshtein distance (ls): POS tags are ranked based on the Levenshtein distance of the corresponding word in the candidate list to the OOV word; if a POS tag occurs several times in the candidate list, we take the value for the word with minimal distance.

nfr+ls: The two weights assigned to POS labels by the algorithms above are normalized and combined linearly.

We use this ranking to induce a lexicon that lists possible POS tags of OOV words. In the experiments, we consider two variants, one which lists only the highest ranked POS tag and one which lists the three best POS tags.

Taggers. We consider two taggers in our experiments: The *hunpos* tagger already mentioned above, which is based on Hidden Markov Models, and a re-implementation of Rehbein (2013)’s CRF tagger using the CRFsuite package (Okazaki, 2007). The list of possible POS tags for OOV words can be used directly as a “morphological lexicon” in the *hunpos* tagger; the tagger uses the POS tags in this lexicon to limit the search space when emission probabilities for OOV words are estimated. In order to give the distributional information to the CRF tagger, we expand a baseline feature set (Rehbein, 2013) by the top 1 and top 3 suggested POS labels, respectively, for OOV words

feature	description	example
wrd	word form	mann
len	word length	4
cap	word capitalized?	false
upper	number upper case	0
digit	number digits	0
sym	number other non-chars	0
pre 1	first char	m
⋮	⋮	⋮
pre n	first n chars	
suf 1	last char	n
⋮	⋮	⋮
suf n	last n chars	
simpos	top n POS suggestions	⟨NN, PIS, PPER⟩

Table 1: Feature set used for experiments with CRF.

(see Table 1); for known words we take the most frequent POS label(s) of the word in the training set.

4 Experiments and Results

We train the distributional models using forum articles downloaded from the German online cooking platform *www.chefkoch.de*. This dataset has been used in previous work by Horbach et al. (2014) and consists of about half a billion tokens from forum posts about a variety of daily-life topics. A small subset of 12,337 tokens comes with manually annotated POS information. Following previous work, we use two thirds (8,675 tokens) of the annotated subset as gold standard for the evaluation and one third as additional training material to re-train the tagger (see Experiment 4). The gold standard contains 1,500 OOV tokens.

The manual annotations use a CMC-specific extension of the STTS tagset (Schiller et al., 1999) proposed by Bartz et al. (2014), covering CMC specific phenomena such as contractions, emoticons or action words. About 4% of the OOV tokens in the gold standard use tags from the extended tagset, which cannot be predicted correctly in our first three experiments.

Experiment 1. Our first experiment compares the three distributional model variants against each other. We tag the test set using the *hunpos* tagger trained on standard newspaper text (*Tiger* corpus) and then replace the POS tags of all OOV words by the POS tag of the word in the candidate list with the highest distributional similarity (*hs*) according to the respective model in a postprocessing step (*pp*).

	all	IV	OOV
baseline	85.0	93.1	46.6
bow	85.3	93.1	48.9
wo	86.6	93.1	56.7
n-gram	**87.2**	93.1	**59.9**

Table 2: Accuracy of the baseline tagger and combinations with different distributional models.

As Table 2 shows, all three distributional models achieve an improvement over the *hunpos* tagger (baseline). The difference to the baseline is small for the *bow* model, but both the *wo* and the *n-gram* model achieve substantial improvements of +10.1% and +13.3%, respectively, for OOV words. The good performance of the *n-gram* model might be surprising as n-gram information is also used directly by the tagger. The added value from the distributional model is, however, that it is trained on a much larger corpus, and abstracts away from the individual context of an OOV word and considers all contexts of this word in the complete training corpus.

Experiment 2. Next, we evaluate the effect of the methods used to rank the POS tags in the induced POS lexicon. Again, we replace the POS tags of OOV words predicted by the tagger in a postprocessing step, but this time using the tag that is ranked highest by each of the three methods considered here, instead of just the distributionally most similar one.

Table 3 shows the results. Levenshtein distance does not improve tagging performance over the *hs* result in our first experiment. However, the *n-first ratio* produces a substantial improvement, and the combination of both methods gives an additional small improvement, showing that these two methods complement each other. The approaches which use the *n-gram* model give the best results, with an improvement of +15.5% on OOV words compared to the baseline. *Upper bound* shows how often the correct POS tag occurs at least once in the candidate list in the first place. We can see that the *nfr+ls* ranking method performs quite well wrt. this upper bound; at the same time, we see that in around one third of the cases the candidate list does not contain the correct POS tag, which obviously

model	hs	nfr	ls	nfr+ls	upper bound
bow	48.9	53.5	48.1	55.3	67.0
wo	56.7	59.7	55.9	60.5	68.7
n-gram	59.9	61.4	57.7	**62.1**	67.7

Table 3: Accuracy of different ranking methods for OOV words.

leaves room for future improvements.

Experiment 3. The results obtained in our first two experiments are quite encouraging, but the method of replacing the POS tag of an OOV word with the highest ranking alternative in a post-processing step is somewhat unsatisfactory, as it does not use potentially helpful information of the context in which the target OOV word occurs. An OOV word will always get the same new POS label, even if the word is ambiguous, and known words in the context cannot benefit from context effects of a correct tag for an OOV word.

To overcome this problem, we use the induced POS lexicon as a "morphological lexicon" for the *hunpos* tagger considering the 3 highest ranked POS tags as ranked by *nfr+ls*. When the tagger sees an OOV word, it uses one of the tags listed in this lexicon. We also consider a re-implementation of the CRF-tagger used by Rehbein (2013) in this experiment, where we add the suggested POS labels to a standard CRF feature set.

Surprisingly, neither *hunpos* nor our CRF-tagger profit from this additional information (see Table 4). To the contrary, the performance decreases substantially. However, if we consider only the highest ranked POS tag (*top 1*), we do get a small improvement for *hunpos* over the *pp* baseline(s), ranging between +0.2% and +0.3%. These results show that the context does not help in picking the correct POS tag among the three candidates listed in the *top 3* lexicon, but forcing the tagger to use the highest-ranked POS tag for OOV words (*top 1*) has a positive effect on the tagging accuracy of the words in the OOV word's context.

Experiment 4. Our final experiment investigates whether a similar performance gain can be achieved when we use a tagger model that has already been adapted to CMC data. We follow Horbach et al. (2015) and use one third of the manually annotated subset of the *chefkoch* corpus in addition to the *Tiger* corpus to train the *hunpos* tagger, reaching a new baseline accuracy of 91.5% . We tag the complete *chefkoch* corpus using this adapted tagger model and train our distributional models on this dataset. Thereby we gain the ability to retrieve also POS tags that only occur in the extended STTS tagset.

	pp	hunpos top1	crfsuite top3
baseline	91.5 (69.4)	91.5 (69.4)	90.8 (72.1)
bow	92.1 (75.2)	92.2 (75.2)	92.7 (78.4)
wo	92.8 (81.3)	93.0 (81.3)	93.1 (81.4)
n-gram	92.9 (**82.1**)	93.1 (**82.1**)	**93.2** (81.9)

Table 5: Results of experiments with already adapted training data. In parenthesis accuracy on unknwon words.

Table 5 shows the results. We observe similar tendencies in performance compared to the previous experiment. The overall best performance is achieved by the *n-gram* model, followed by *wo* and *bow*. Interestingly, the CRF tagger achieves with 93.2% the best overall result (+1.7% over the *hunpos* baseline and +2.4% over the CRF baseline) although it does not reach the performance of *hunpos* on OOV words.

The relative improvements over the baseline(s) are a bit smaller. One reason for this is that the adapted tagger model covers some of the most frequent OOV words in the whole *chefkoch* corpus so that these frequent and presumably easier cases for the distributional model do not need to be handled any more. Another reason is that some tags from the extended STTS tagset, specifically emoticons, often appear in syntactically not integrated positions and show high distributional similarity with punctuation, which makes the prediction of POS tags of OOV punctuation much harder.

Error analysis. Having shown that using our system does have a positive effect on the POS tagging of OOV words, it is still interesting to known, what kind of errors are made by the baseline tagger in the first place and which of these can be handled by our system.

The confusion matrix in Table 6 shows the classifier's performance and different classification errors made by the baseline tagger as well as the effects of our best system compared to the baseline in parentheses. We collapse POS tags into five groups for nouns, adjectives, verbs, other standard

	pp	hunpos top1	hunpos top3	crfsuite top1	crfsuite top3
baseline	85.0 (46.6)	85.0 (46.6)	85.0 (46.6)	85.0 (50.1)	85.0 (50.1)
bow	86.4 (55.3)	86.7 (55.3)	86.3 (53.7)	87.0 (57.3)	86.9 (56.5)
wo	87.4 (60.5)	87.6 (60.5)	86.8 (56.5)	87.5 (60.0)	87.6 (60.4)
n-gram	87.7 (62.1)	**87.9** (**62.1**)	86.7 (55.9)	87.6 (61.1)	87.6 (60.4)

Table 4: Accuracy for different ways of integrating the information into the taggers. *top 3* gives the results when the three highest ranked POS tags are considered; *top 1* gives the results when only the highest ranked POS tag is used. In parenthesis is the accuracy on only the unknown words.

		baseline tagger (effect of best configuration)				
		N	A	V	other	new
gold	N	87.2 (+8.4)	7.4 (-5.8)	2.1 (-1.1)	3.3 (-1.5)	0.0 (+0.1)
	A	2.1 (+0.2)	92.6 (±0)	3.1 (-1.6)	2.1 (+1.2)	0.0 (+0.2)
	V	2.1 (-1.7)	1.1 (-0.5)	96.6 (+1.9)	0.2 (+0.2)	0.0 (+0.1)
	other	3.0 (-2.6)	2.1 (-1.5)	0.9 (-0.9)	94.0 (+4.7)	0.0 (+0.4)
	new	13.2 (-3.4)	8.3 (-6.0)	9.4 (-5.7)	69.1 (-57.4)	0.0 (+72.5)

Table 6: Confusion matrix between our baseline tagging model and the gold standard. In parentheses is the absolute difference to this baseline for our best-performing model. POS categories are collapsed into nouns, adjectives, verbs, other standard STTS tags and the new STTS 2.0 tags.

STTS tags and the new STTS 2.0 tags.

We can observe three interesting phenomena: Firstly, due to a lot of lower-cased – and thus unknown – noun forms, there is a high rate of nouns getting erroneously tagged as adjectives (7.4%). In fact, out of the 111 nouns tagged as adjectives by the baseline tagger, 94 are lower-case. This problem is mostly solved by our system (-5.8%).

Another frequent mistake is the tagging of interjections (included in *other*) as proper nouns. This is also handled quite well ($3.0\% \rightarrow 0.4\%$).

Finally, the baseline tagging model is of course not able to cope with new tags from the extended STTS tagset. The adapted model leads to an accuracy of 72.5% for these tags, which – while not quite reaching the per-class accuracy of the other classes – is a reasonable result, given the limited amount of training data.

5 Conclusions

We have shown that distributional similarity information can be used to learn possible POS tags of out-of-vocabulary words and thereby improve the performance of POS taggers on CMC data. Our best performing approach increases the overall tagging accuracy on German internet forum posts by $+2.9\%$ compared to a tagger that has been trained on standard newspaper text; for a tagger that has already been adapted to CMC data, our approach increases accuracy by $+1.7\%$ / $+2.4\%$ to 93.2%.

We use two different taggers in our experiments, a HMM-based tagger and one based on CRF. One interesting observation is that the HMM-tagger generally performs better on OOV words, while the CRF tagger gives the overall best results when we use an already adapted tagger. This observation suggests that information about OOV words is not encoded optimally in the CRF-based tagger, and that we can improve our approach in future work.

Our approach is completely unsupervised in the sense that it does not rely on any additional manually annotated data, so it can be applied to other kinds of CMC data as well.

Acknowledgements. This work is part of the BMBF-funded project “Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen.” We would like to thank Ines Rehbein for helpful discussion and support with the implementation of the CRF-tagger.

References

Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2014. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics*, 28(1):157–198.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther Koenig, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Journal of Language and Computation, Special Issue*, 2(4):597–620.

Phani Gadde, L. Venkata Subramaniam, and Tanveer A. Faruquie. 2011. Adapting a WSJ trained part-of-speech tagger to noisy text: preliminary results. In *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data*, page 5. ACM.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 421–432. Association for Computational Linguistics.

Andrea Horbach, Diana Steffen, Stefan Thater, and Manfred Pinkal. 2014. Improving the performance of standard part-of-speech taggers for computer-mediated communication. In Josef Ruppenhofer and Gertrud Faaß, editors, *Proceedings of the 12th Edition of the Konvens Conference, Hildesheim, Germany, October 8-10, 2014*, pages 171–177. Universitätsbibliothek Hildesheim.

Andrea Horbach, Stefan Thater, Diana Steffen, Peter M. Fischer, Andreas Witt, and Manfred Pinkal. 2015. Internet corpora: A challenge for linguistic processing. *Datenbank-Spektrum*, 15(1):41–47.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Ines Rehbein. 2013. Fine-grained pos tagging of german tweets. In *Language Processing and Knowledge in the Web*, pages 162–175. Springer.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS-CL, University Stuttgart.

Bootstrapped Extraction of Index Terms from Normalized User-Generated Content

Piroska Lendvai
Saarland University
Dept. of Computational Linguistics
Saarbrücken, Germany
piroska.r@gmail.com

Thierry Declerck
Saarland University
Dept. of Computational Linguistics
Saarbrücken, Germany
declerck@dfki.de

Abstract

We report on the extraction of key phrases for news events, based on string alignment between social media posts and user-linked web documents. Hashtag normalization is tested for enhancing string similarity, while both token-based tweet similarity and manual event annotations are tested for transferring web links to posts that do not refer to external documents. We are able to identify more terms via web link transfer compared to no link transfer, and obtain syntactically and semantically more complex terms compared to general document-based term extraction.

1 Introduction

Creating the logical representation of a document collection in terms of index terms is a crucial step in information retrieval. The extraction of a meaningful set of index terms for a document collection, instead of making every word (noun) an index term, can profit from human insights. Our goal is to find, collect, and utilize such insights from social media content. Our core assumption is that users who include a reference to an external web document in their social media post are implicitly encoding a relevance signal; this assumption is analogous with utilizing landing page information from click data to classify user intent (see e.g. Joachims (2002)).

We investigate the extraction of key terms for news events, based on string alignment between social media posts and user-linked web documents, as described in Lendvai and Declerck (2015). Manually assigned event annotations are used to transfer the web links to posts that do not refer to external documents, thereby boosting the amount and quality of extracted index terms. Then, token-based tweet similarity is going to be used to the same end.

Hashtag harmonization is supposed to enhance string similarity; in Declerck and Lendvai (2015a) we reported on a hashtag processing approach that we test in our present study as well. Hashtags are normalized, lemmatized and segmented in a data-driven way in a simple offline procedure that generates a gazetteer of hashtag elements. In Declerck and Lendvai (2015b) we developed the basic Linguistic Linked Open Data (LLOD)[1] infrastructure for representing hashtags from social media posts. We explained how the OntoLex model[2] is used both to encode and to enrich the hashtags and their elements by linking them to existing semantic and lexical LOD resources: DBpedia and Wiktionary.

Our goal in the current study is to give a pilot evaluation on application- and data-driven, language-independent approaches for term extraction, comparing the obtained terms with document-based term extraction, and comparing the terms after hashtag harmonization and web link transfer against non-harmonized data and no link transfer. We also report on how term extraction is affected by link transfer based on automatically assigned tweet similarity instead of manual annotations.

2 Hashtag harmonization

Hashtags allow users to classify their social media text, especially Twitter messages, into semantic categories. Those tags are typically named entities such as ”#Ottawa”, terms such as ”#Shooting”, or concatenated phrases such as ”#WearewithCanada”. The relevance of hashtags to identify text topics has been utilized by several approaches. Laniado and Mika (2010) find hashtags to qualify as strong identifiers for Semantic Web applications. However, the

[1] See (Chiarcos et al., 2013) and http://linguistic-lod.org/llod-cloud

[2] OntoLex is a model for the representation of lexicons (and machine readable dictionaries) relative to ontologies. It has been developed in the context of the W3C Ontology-Lexica Community Group, see https://www.w3.org/community/ontolex/.

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 44–48, University of Duisburg-Essen, September 28, 2015. ©2015 GSCL

analysis of lexical variation to identify semantically coincident hashtags has not yet been considered, but is important to identify messages relating to the same topic. Semantic clustering approaches (Pöschko, 2011; D. Antenucci et al., 2011) focus on semantic topics and their relations, but neglect variation within hashtags. This kind of processing is necessary to obtain more precise information on the exact semantics represented by hashtags and identify all related tweets within a dataset.

Hashtags appear in different cases and need to be normalized first. Secondly, hashtags need to be lemmatized to automatically match singular and plural uses of words. Finally, the segmentation of complex hashtags into its individual components is needed if one wants to recognize hashtag paraphrases in related documents. By reducing the (ortho)graphical variation of hashtags, basic string and substring matching across document types is hypothesized to be made more effective.

The corpus we were working on in (Declerck and Lendvai, 2015a) was a UK-Riots corpus established by the Guardian[3]. We are now testing and extending our approach to datasets that have been collected in the context of the Pheme project[4], relating to the events of the Ottawa Shooting and the Gurlitt art collection, as described in (Lendvai and Declerck, 2015). The corpus contains 40,201 tweets (including many retweets) in which we identified 22,825 hashtags.

Normalization We normalize hashtags by lowercasing all letters. On Twitter, typographical errors and misspellings are common. We used the string similarity measure implemented in the Python module *difflib*[5] to detect basic spelling mistakes such as "#shotting". In order to avoid valid words to be corrected as misspellings, e.g. 'from' and 'form', the strings are matched to the unix words list[6]. If one of the strings is not in the list, the change is made.

Lemmatization Variation in hashtags also originates from suffixation, a frequent suffix is the plural sign. While there might be some semantic difference due to the use of plural or singular , it is worth reducing the plural in hashtags to the singular in order to be able to compare and to link hashtags to documents external to the Twitter sphere. We use a straightforward approach: comparing words ending with an '-s' to the unix word list. If the word ending in 's' is present in this list, like for example the word 'news', no action is taken. Otherwise we perform lemmatization. We are currently evaluating if this approach is accurate for hashtags compared to a proper lemmatizer adapted to user-generated content. We assume that this step will be needed in any case for languages with a richer morphology as English.[7].

Segmentation Deriving components from segmented hashtags as search terms presumably facilitates the automatic linking of tweets to documents from other genres, which do not contain hashtags, such as news articles. We use a simple approach to segmentation, starting from hashtags that use camel notation (see Declerck and Lendvai, 2015a), e.g. '#OttawaShooting', yielding the segments 'ottawa' and 'shooting', which will in turn be utilized to segment its casing-variant '#ottawashooting'. In our corpus we have 1,363 occurrences of 'OttawaShooting' and 232 occurrences of 'ottawashooting', whereas '#shooting' is used only 18 times as a standalone string. Hashtag segmentation is able to impact hashtag distribution, resulting in e.g. 1,611 occurrences of '#shooting', enabling better term relevance metrics.

Our simple approach to segmentation includes the risk to generate arbitrary segments (e.g. 'Wearewith')[8]. The unix words list can again be put to use for checking the validity of the components resulting from segmentation. Additionally, we apply queries to named entities resources in the LOD for validating such components.[9] These validation procedures make the harmonization of hashtags to be considered as an offline procedure, generating specialized gazetteers. We are investigating whether rules or patterns are possible to be extracted from the results of the current experiments, to be reused on incoming tweet streams for online processing.

[3] http://www.theguardian.com/news/datablog/2011/dec/08/twitter-riots-interactive

[4] http://www.pheme.eu/

[5] https://pymotw.com/3/difflib/

[6] https://en.wikipedia.org/wiki/Words_\%28Unix\%29

[7] See for example the work by (Horbach et al., 2014) on improving the performance of PoS taggers applied to German Computer mediated Communication

[8] We are grateful to an anonymous reviewer pointing out this issue.

[9] The querying procedure, implemented on Python, is described in details in (Declerck and Lendvai, 2015b).

3 Tweet-to-Document Linking

Very recently, creating systems for Semantic Textual Similarity judgements on Twitter data has been a Shared Task in the Natural Language Processing community (Xu et al, 2015). Given two sentences, the participating systems needed to determine a numerical score between 0 (no relation) and 1 (semantic equivalence) to indicate semantic similarity on the hand-annotated Twitter Paraphrase Corpus. The sentences were linguistically preprocessed by tokenization, part-of-speech and named entity tagging. The system outputs are compared by Pearson correlation with human scores: the best systems reach above 0.80 Pearson correlation scores on well-formed texts. The organizers stress that "while the best performed systems are supervised, the best unsupervised system still outperforms some supervised systems and the state-of-the-art unsupervised baseline."

In Lendvai and Declerck (2015) we proposed a cross-media (CM) linking algorithm in the PHEME project to connect User-Generated Content (UGC) to topically relevant information in complementary media, which we use in the current study as well. Each tweet in our datasets is manually annotated for an event. E.g. the tweet 'RT @SWRinfo: Das Kunstmuseum Bern nimmt das Erbe des Kunstsammlers Cornelius #gurlitt an.' is assigned the event *'The Bern Museum will accept the Gurlitt collection'*, while 'NORAD increases number of planes on higher alert status ready to respond if necessary, official says. http://t.co/qsAnGNqBEw #OttawaShooting' is assigned the event *'NORAD on high-alert posture'*, etc.

For each URL-containing tweet within each event, a tweet-to-document similarity calculation cycle is run between tweets that link an external web document, and the linked web document. Similarity is evaluated in terms of the Longest Common Subsequence (LCS) metric. LCS returns a similarity value between 0 (lowest) and 1 (highest) based on the longest shared n-gram for each text pair, without the need for predefined n-gram length and contiguity of tokens (cf. Lin (2004)).[10]

3.1 LCS terms extraction

We use LCS to collect the top-5 scored longest common token subsequences identified for a linked document, based on a series of LCS computations producing LCSs between one, but sometimes more, tweets linking this document and each sentence of the document. No linguistic knowledge is used, except for stopword filtering by the NLTK toolkit[11]. Then the LCS cycle is applied to the same document set but paired with tweets that did *not* link external documents, based on the hand-labeled events. We are able to extract more, and lexically different phrases due to the link transfer.[12] For example, for the web document with the headlines *"Swiss museum accepts part of Nazi art trove with 'sorrow' — World news — The Guardian"* the extracted top terms based on tweets linking to this document are: 'swiss museum accepts part nazi art trove', 'nazi art', 'swiss museum', 'part nazi', 'nazi', whereas the extracted top terms based on tweets *not linking any document* but being annotated with the same event as the tweets referring to this document, are 'kunstmuseum bern cornelius gurlitt', 'fine accept collection', 'museum art', 'kunstmuseum bern cornelius gurlitt', 'kunstmuseum bern gurlitt', exemplifying that the Gurlitt dataset holds multilingual data, since we obtain terms not only in English, but in German as well.

3.2 Term extraction evaluation

3.2.1 No transfer to URL-less tweets

We are able to grow the set of extracted unique terms significantly if we perform the web link transfer step, when compared to not performing this step: from 110 to 186 in Gurlitt, and from 171 to 320 in Ottawa. The obtained term sets are highly complementary: about 70-90% of the phrases extracted from URL-less tweets are unseen in the phrase set extracted from URL-ed tweets.

3.2.2 Transfer based on automatically grouped tweets

We have compared the results of our LCS approach to experimental results where instead of using tweet clusters based on manual event annotations, we create tweet clusters by computing tweet similarity between each tweet and a centroid tweet for each event (designated by the phrase used in the manual event annotation), via a LCS similarity threshold. Inspired by Bosma and Callison-Burch (2007) who use an entailment threshold value of 0.75 for detecting paraphrases, we obtained our LCS similarity

[10] For details please see (Lendvai and Declerck, 2015).

[11] http://www.nltk.org/index.html

[12] For more details we again refer to (Lendvai and Declerck, 2015).

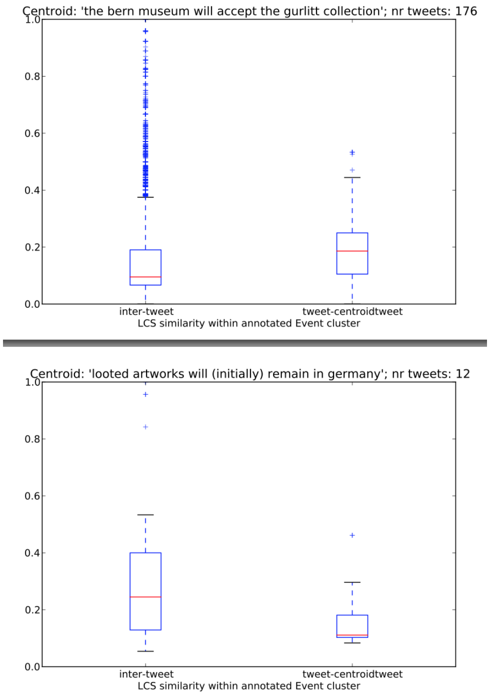


Figure 1: Tweet similarity distribution in terms of LCS values for two events from the Gurlitt dataset: tweet-tweet similarities within an event cluster, as well as centroid tweet - tweet similarities are plotted.

threshold t empirically by averaging the third quartile of LCS value distributions relating to an event over all events in a dataset ($t > 0.22$). Figure 1 illustrates tweet similarity distribution in terms of LCS values for two events from the Gurlitt dataset. We computed LCS values both in an intra-tweet way (i.e., LCS for all pairs of tweets within a tweet event cluster, the size of which is indicated in the upper right corner), and in the centroid-tweet way (i.e., LCS for all centroid-tweet pairs within the event cluster). Since Gurlitt is a multilingual set, the LCS scores often have a very wide distribution, also indicated by the large number of outliers in the plot.

The approach is rather crude and on the current toy datasets achieves a event-based-mean precision of 1.0 for Gurlitt and 0.32 for Ottawa, while a event-based-mean recall of 0.67 for Gurlitt and 0.78 for Ottawa. With this approach, we get much less URL-less tweets (Gurlitt: 16 vs 43, Ottawa:117 vs 182), but this seems to have an impact only on the Gurlitt multilingual dataset on the amount of extracted unique phrases from URL-less tweets (Gurlitt: 64 vs 93, Ottawa: 178 vs 197). Importantly, the quality and semantics of the extracted phrases for both datasets remain in line with those based on link transfer via hand-labeled events.

3.2.3 Frequency-based term extraction

We extracted a document-based term set from all tokens in the fetched documents that were automatically classified as nouns; part-of-speech information was obtained from the NLTK platform. These sets seem semantically more general than the terms obtained by the LCS approach (e.g. 'ausstellung', 'sammlung', 'suisse', i.e., 'exhibition', 'collection', 'switzerland') and are also smaller in size: 75 unique terms from all documents linked from the Gurlitt set, obtained in a top-5-per-document cycle to simulate the LCS procedure, and 116 for Ottawa. The obtained term set consists of single tokens only, while the average phrase length using the LCS approach is 3.65 for Gurlitt and 3.13for Ottawa.

4 Results and Conclusion

Our approach, based on longest common subsequence computation, uses human input for extracting semantically meaningful terms of flexible length. We link tweets to authoritative web documents, and create lexical descriptors extracted from tweets aligned with documents. The method is language-independent and unsupervised.

The extracted phrases are expected to have indexing potential and could be used in their multi-word form or could be tokenized further. Hashtag normalization has currently no significant impact on our toy-sized datasets; we have tested it on German data for the first time and plan to improve it in a data-driven way.

Scaling up from our current pilot setup, we are going to report on qualitative and quantitative results on cross-media, cross-lingual text linking in forthcoming studies.

Acknowledgment

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233).

References

D. Antenucci, G. Handy, A. Modi, and M. Tinerhess (2011). Classification of tweets via clustering of hashtags. EECS 545 Final Project, 545:1-11.

W. Bosma and C. Callison-Burch (2007). Paraphrase substitution for recognizing textual entailment. In:

Evaluation of Multilingual and Multi-modal Information Retrieval (pp. 502-509). Springer Berlin Heidelberg.

S. Bird, E. Klein, and E. Loper (2009). Natural Language Processing with Python, O'Reilly Media.

C. Chiarcos, P. Cimiano, T. Declerck, J.P. McCrae (2014). Linguistic Linked Open Data (LLOD) - Introduction and Overview in: Christian Chiarcos, Philipp Cimiano, Thierry Declerck, John P. McCrae (eds.): 2nd Workshop on Linked Data in Linguistics, Pages i-xi, Pisa, Italy, CEURS, 2013

T. Declerck and P. Lendvai (2015a). Processing and Normalizing Hashtags. in: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov (eds.): Proceedings of RANLP 2015, Pages 104-110, Hissar, Bulgaria, INCOMA Ltd, Shoumen, BULGARIA, 9/2015

T. Declerck and P. Lendvai (2015b). Towards the Representation of Hashtags in Linguistic Linked Open Data Format. in: Piek Vossen, German Rigau, Petya Osenova, Kiril Simov (eds.): Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, Hissar, Bulgaria, INCOMA Ltd, Shoumen, BULGARIA, 9/2015

A. Horbach, D. Steffen, S. Thater and M. Pinkal (2014). Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of the 12th edition of the Konvens conference (Konvens 2014).

T. Joachims (2002). Optimizing Search Engines Using Clickthrough Data. Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).

D. Laniado and P. Mika (2010). Making sense of Twitter. In Proceedings of the Semantic Web Conference – ISWC 2010, pages 470-485.

P. Lendvai and T. Declerck (2015). Similarity-Based Cross-Media Retrieval for Events. In: Ralph Bergmann, Sebastian Görg, Gilbert Müller (eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB, Trier, Germany, CEURS, 10/2015

Chin-Yew Lin (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8.

J. Pöschko (2011). Exploring twitter hashtags. arXiv preprint arXiv:1111.6553, 2011.

Xu, Wei, Chris Callison-Burch, and William B. Dolan. (2015). SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).

Author Index