Bochumer Linguistische Arbeitsberichte 17

NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication
22 September 2016

Bochumer Linguistische Arbeitsberichte

Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 17 (September 2016)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr 2016
ISSN **2190-0949**

Michael Beißwenger, Michael Wojatzki
and Torsten Zesch (Eds.)

NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication 22 September 2016

2016

Bochumer Linguistische Arbeitsberichte

(BLA 17)

Contents

Natural language processing for computer-mediated communication and social media discourse: (still) 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) and social media interactions such as chats, blogs, forums, tweets, newsgroups, messaging applications (SMS, WhatsApp), interactions on social network sites and on wiki talk pages: As part of large corpora which crawled from the web, CMC data are often regarded as an unloved bycatch that proves for linguistic annotation by means of standard natural language processing (NLP) tools that are optimized for 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 discourse is an important resource that closes the "CMC gap" in corpora of contemporary written language and language-in-interaction. With a considerable part of contemporary everyday communication being mediated through CMC technologies, up-to-date investigations of language change and linguistic variation need to be able to include CMC discourse in their empirical analyses.

The goal of the NLP4CMC workshops which are organized by the special interest group *social media / internet-based communication* of the *German Society for Language Technology and Computational Linguistics (GSCL)* 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 workshop was held in September 2014 at *KONVENS* at the University of Hildesheim.[1] The 2nd workshop was held in September 2015 at the *GSCL* Conference at the University of Duisburg–Essen.[2] This volume presents proceedings of the 3rd workshop which was held in September 2016 at *KONVENS* at the Ruhr-University

[1] Workshop Proceedings of the 12th Edition of the Konvens Conference. Hildesheim, Germany, October 8–10, 2014. Hildesheim: Universitätsverlag. http://www.uni-hildesheim.de/konvens2014/data/konvens2014-workshop-proceedings.pdf

[2] Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC 2015). Essen. http://sites.google.com/site/nlp4cmc2015/NLP4CMC-2015.pdf?attredirects=0&d=1

Bochum. Besides three individual papers the workshop included a round table on the results of *EmpiriST*, a community shared task on the automatic linguistic annotation of CMC and web corpora.[3] The goal of the round table was to identify perspectives for future work in adapting tools for tokenization and part-of-speech tagging for processing and annotating written CMC discourse.

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

Duisburg and Essen, September 2016
Michael Beißwenger
Michael Wojatzki
Torsten Zesch

[3] Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26). http://aclweb.org/anthology/W16-26

A Discourse-structured Blog Corpus for German: Challenges of Compilation and Annotation

Holger Grumt Suárez
Justus-Liebig-Universität Gießen
ASCL
Otto-Behaghel-Str. 10 D (D 410)
35394 Gießen
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de

Natali Karlova-Bourbonus
Justus-Liebig-Universität Gießen
ASCL
Otto-Behaghel-Str. 10 D (D 406)
35394 Gießen
Natali.Karlova-Bourbonus@germanistik.uni-giessen.de

Henning Lobin
Justus-Liebig-Universität Gießen
ASCL
Otto-Behaghel-Str. 10 D (D 407)
35394 Gießen
Henning.Lobin@uni-giessen.de

Abstract

The present paper reports the first results of the compilation and annotation of a blog corpus for German. The main aim of the project is the representation of the blog discourse structure and relations between its elements (blog posts, comments) and participants (bloggers, commentators). The data included in the corpus were manually collected from the scientific blog portal SciLogs. The feature catalogue for the corpus annotation includes three types of information which is directly or indirectly provided in the blog or can be construed by means of statistical analysis or computational tools. At this point, only directly available information (e.g., title of the blog post, name of the blogger etc.) has been annotated. We believe, our blog corpus can be of interest for the general study of blog structure or related research questions as well as for the development of NLP methods and techniques (e.g. for authorship detection).

1 Introduction

In our opinion, two views on computer-mediated communication (CMC) – linguistic and structural – have so far been established. According to the linguistic view, the language of CMC represents a distinct type of language form besides written and spoken language. Moreover, it combines characteristics of these two traditional language forms thus constituting a bridge between them. The structural view in its turn concentrates on building up of CMC. Two different kinds of CMC structure can be distinguished – external and internal. External structure relates to the representation, or layout, of CMC by means of HTML mark-up language. External structure of the most blogs includes for example a header (title), content, a footer (contact information) and a sidebar (site navigation). Internal structure in its turn relates to the generic structure of the CMC content. It describes a set of structural elements (e.g., post, comment, thread, word cloud etc.), properties and principles a CMC is constructed of and built on to function as a holistic construct and to match its purpose.

The identification of the full spectrum of CMC characteristics – linguistic or structural – still faces some major challenges primarily as a result of lacking valid annotated data. Storrer (2014: 189) claims that for this purpose a special – third - kind of corpus besides the written and spoken corpora is needed. She also adds that appropriate standards, methods and quality criteria for the study of CMC are crucially important as well.

In the present study, the structural nature of the weblog (henceforth blog) as a representative genre of CMC is of interest. We describe the genre blog as a dynamic, "living" construct of interrelated and interacting elements. The dynamics of a blog arise from its constant expansion as a result of ever more comments and blog posts as well as on the account of new blog par-

ticipants. Additionally, the author of the blog (henceforth blogger) can edit his post any time and add new information on request. The interrelatedness and interaction between elements (blog post, comments) and agents (blogger, commentators) of the blog contribute to the dynamics of the blog as well.

To demonstrate this idea, we compiled the first version of an annotated blog corpus in German using the scientific blog portal SciLogs (SciLogs, 2016) as a data source. The corpus includes both blog posts and related comments. The catalogue of features for the annotation of the corpus is based on three types of information directly or indirectly available from the data source. The typology of information is proposed in Section 4.1.

The general structure of the paper is as follows. Section 2 provides an overview of the studies related to the topic of the present project. In Section 3 the process of data collection is presented. Section 4 describes the main steps of the blog corpus annotation including annotation scheme. Some observed challenges for the automation of the task and possible solutions are also included in this section. Finally, Section 5 summarizes the results of the project and outlines the next steps.

2 Related Work

Currently, there is a limited number of publicly-available, large-scale blog corpora, which is surprising given the great influence of blogs on the web in general.

An example of a large-scale blog corpus is the German language wordpress blog corpus by Barbaresi and Würzner (Babaresi/Würzner, 2014). The corpus consists of 158,719 German wordpress blogs released under the Creative Commons license. The content of the blogs is divided into two different parts, the blog posts and the blog comments. The corpus can be used for example "to find relevant examples for lexicography and dictionary building projects, and/or to test linguistic annotation chains for robustness" (Babaresi/Würzner, 2014). Additionally, it gives a good insight into a German blog language.

Another example of a blog corpus is the bilingual (German, French) corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne), which is part of the LETEC (Learning and Teaching Corpus) (Abendroth-Timmer, 2014). This corpus is included in the CoMeRe (Communication médiée par les réseaux) project, which "aims to build a Kernel corpus assembling existing corpora of different CMC […] genres and new corpora build on data extracted from the Internet" (Abendroth-Timmer, 2014). The INFRAL blog corpus consists of posts and comments from two groups: a group of ten francophone learners of German as a foreign language from l'Université de Franche-Comté and a group of nine German-speaking learners of French as a foreign language from the University of Bremen who e.g. had to discuss various intercultural topics. One task for compilation of this corpus was to model the structure of interactions. Every comment is given a reference to the ID of the post, but the links between the comments are not included. The TEI schema developed for the CoMeRe project – this project is also part of the TEI special interest group (SIG) "computer-mediated communication" (CMC) – is an important basis for our own annotation schema.

3 Data Collection

So far, we have compiled a German language test corpus, which contains 21 blog posts and 195 comments. The source of the data is the prerelaunched version of the scientific blog portal SciLogs (SciLogs, 2016). SciLogs is subdivided into different sections (BrainLogs, ChonoLogs, KosmoLogs, WissenLogs) where scientists and those interested in science can interact on different topics. The data collection for the test corpus was done manually and focused on blog posts and comments appeared in the period of one week (randomly chosen week 49 of 2015). Moreover, we did not focus on a particular section and extracted the blogs from different sections.

Our next step will be to complete our corpus with the data appeared in 2015 considering all SciLogs sections. According to our current knowledge, the SciLogs data of 2015 includes about 1.200 blog posts and 12.000 comments. Retrieval of the blog data from the web will be conducted semi-automatically. For this purpose, an open source program HTTrack Website Copier (Roche, 2016) will be used. HTTrack enables the download of all kinds of the website data stored on the server including HTML pages, images and other files to a local directory on a computer. After the retrieval step, the data will be cleaned from the noise in the data and represented in form of HTMl pages (external structure). Finally, the relevant content will be extracted from the HTML pages and annotated

with TEI annotation standard (internal structure) by using the programming language Python and its libraries for parsing HTML/XML files.

4 Data Annotation

4.1 Typology of Blog Information

In our opinion, three types of information provided in the blog based on how the former is made available can be distinguished. The first type (type A) incorporates information which is directly available in the blog or from the source code of the blog site. In the blog post structure, it includes the blog post itself along with the meta information such as the title of the blog post, date of creation, the name of the blogger, the categories the entry belongs to and main keywords. In the structure of the comments, type A information is represented by the total number of comments as well as the name of the commentator, date and comment ID. The second type (type B) includes information which is not directly available but can be inferred from type A information, e.g. usual activity time of a commentator (at what time a particular commentator usually writes his comments). Finally, the third type (type C) is an interpretative information type. This kind of information is neither directly nor indirectly provided in the blog but is rather the result of statistical (basic statistics), linguistic (e.g., part-of-speeches) and discourse (e.g., topic identification with topic modelling) interpretation and analysis of the blog entries. The interpretative information type can either be collected manually or by use of computational tools.

4.2 Annotation Standard

Up to now, no standard exists for representing CMC data. One option could be to design an XML schema for CMC from scratch, which would perfectly fit the needs of our project. The main reason as to why we are not going along with XML is that the schema would be idiosyncratic and the corpus would not interoperate properly without causing difficulty with other resources. When searching for a standard for the representation of texts in digital form, one will take a look at the Text Encoding Initiative (TEI). However, none of the modules in the current version of the TEI Guidelines (P5) can be adopted for our project. Fortunately, the SIG CMC group under the direction of Beißwenger (TU Dortmund) has been working on the adaption of TEI guidelines to the presentation of genres of CMC since 2012 (Beißwenger 2015). Given that no module for CMC is so far ready to use, we have started to look for schema drafts by the SIG CMC group and up to now, a couple of corpora have been released by the SIG CMC group. Among them are CMC genres like tweets, email, text chat, wiki discussions and weblogs (Chanier 2014, Beißwenger 2013, Storrer 2015). The schema that is most useful for our needs, is the one released in 2014 by the French network CoMeRe (Communication médiée par les réseaux) (Hriba 2013). The CoMeRe schema is based on the previous schema draft by DeRiK (Beißwenger2013) and includes e.g. the metadata schema for CMC. But still, there is no possibility for representing the full structure of a blog and especially the related comments. Our goal is to take the latest schema draft provided by the SIG CMC (Beißwenger 2016) without changing the main characteristics of the schema. The status of that schema is that of a "core model for the representation of CMC" (Beißwenger et al. 2012: 6). And so we will probably need to redefine some elements while also introducing some new ones. Another possibility is to use the existing text structure module and investigate how many modifications have to be done in order the schema fits the purpose of our project.

4.3 Some Challenges of Discourse Structure Annotation

There is a number of challenging aspects which have to be dealt with for the task of blog corpus annotation. These challenges are in most cases the result of the particularities of the content management system (CMS) functionality used by our blog data source. Most of the challenges deal with the structure of the comments. As we are at an early stage of our project, only a limited number of challenges and solutions will be described here.

The first challenge is due to the absence of an editing function for the comments. The commentator who edits the text of the comment creates a new entry which appears in the timeline as an autonomous comment. Thus, the comments structure of our blog corpus includes both original comments and their edited versions appeared to the time of the data collection. Though, this aspect does not have an impact on the difficulty of the automation of the annotation task. However, it first impacts the accuracy of the total number of distinct comments (type A information). Second, it creates confusing linkages in the comments structure.

The latter problem also arises as the result of the second challenge – the possibility that one comment refers to more than one previous comment. Unfortunately, the CMS of our blog source does not offer any special options to mark or highlight multiple comment references. In some cases, the commentators use constructions such as [@name]* to overcome this problem. In other cases, an additional analysis of the comment content is required. For the purpose of the study, only explicit references are taken into consideration. No deeper content analysis has been conducted. The identification of multiple references and their annotation with TEI was processed automatically and then manually checked for mistakes in order to achieve accurate and reliable results. We believe that it is be less time- and cost-consuming than fully manual processing of the data. The automatic part is conducted based on explicit marks of multiple reference such as [@name]*. In the TEI blog annotation the multiple references are specified by enumeration of the ids of their comments (<replyTo>).

Finally, the third challenge is the task of the correct assignment of the comments to the level in the hierarchical structure of the comments. At present, the number of possible level assignments is limited to five. All comments appearing after the first comment on the fifth level are (wrongly) assigned to the fifth level. In order to solve this problem, we developed a simple algorithm to compute the correct level of the comments. The algorithm first takes the person reference ("@name", "[name] schrieb (engl.: wrote)" etc.) included in the text of the analyzed comment as the input. In the case of multiple references, only the first reference is taken into consideration. The algorithm then searches backwards for the matches between the person reference and the name of the commentator in the previous comments. Through matches, level of the analyzed comment is computed as the sum of the level assignment of the comment which the person reference belongs to and 1. By absence of the references, the level of the comment is computed subsequently.

5 Summary

The main steps conducted for the purposes of a scientific blog corpus compilation as well as challenges faced during this process were described in the present study. The current version of the corpus contains 21 blog posts and 195 related comments written in the period of one week. We are convinced that comments are an essential part of a blog corpus. On their own or in connection with the correspondent blog post, they provide valuable information for processing diverse research questions on the language of the blog and its structure. For example, based on the name of the commentator and the time of his comments, we can compute at what time a particular commentator is active in the blog.

The data for our blog corpus was manually collected and annotated according to the TEI schema drafts developed by the TEI special interest group. For the annotation, three types of information (direct, indirect and interpretative) based on the availability of the latter have been identified. The present version of the corpus includes annotation of the first type - directly retrieved information (e.g., the name of the blogger, title of the blog entry, the name of the commentator etc.). The next objective of the project is an expansion and full annotation of the corpus as well as the automation of the data collection and annotation task. At the final stage of the present project, our annotated corpus will be made available to the interested community to perform diverse kinds of research and experiments. Our aim is to enable the access to the corpus through a searchable online database. Additionally, we plan to make a part of the corpus to be available upon request. For the legal aspects of the SciLogs data usage and publication an external competent institution will be consulted.

Acknowledgments

We would like to thank our anonymous reviewers for their insightful comments and suggestions. Following the feedback, we included several improvements in our paper.

Reference

Barbaresi, A., Würzner, K.-M. (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In KONVENS 2014, NLP4CMC workshop proceedings, p. 2–10.

Beißwenger, M. et al. (2012). A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI), Issue 3.

Beißwenger, M. et al. (2013). DeRiK: A German reference corpus of computer-mediated communication. pp. 531-537. In: M. A. Finlayson (Eds.), LLC. The Journal of Digital Scholar-

ship in the Humanities, Volume 28, Number 4. Oxford, OUP, pp. 531-537.

Beißwenger, M. (2015). Computer-Mediated Communication SIG. In TEI Website. http://www.tei-c.org/Activities/SIG/CMC/ (last retrieved 20 April 2016).

Beißwenger, M. (2016). SIG:Computer-Mediated Communication. In TEI Website. http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication (last retrieved 20 April 2016).

Abendroth-Timmer, D. et al. (2014). Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne). Banque de corpus CoMeRe. Ortolang.fr: Nancy. https://hdl.handle.net/11403/comere/cmr-infralhttps://hdl.handle.net/11403/comere/cmr-infral (last retrieved 23 August 2016).

Chanier,T. et al. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Journal of Language Technology and Computational Linguistics. Berlin, GSCL, pp1-31.

Hriba, L., Chanier, T. (2013). Projet européen TEI-CMC. Comere: Corpuscomere. Communication médiée par les réseaux. In Comere Website. https://corpuscomere.wordpress.com/tei/ (last retrieved 20 April 2016).

Roche, X. (2016). HTTrack. Website Copier. http://www.httrack.com/ (last retrieved 20 April 2016).

SciLogs (2016). SciLogs. Tagebücher der Wissenschaft. Spektrum der Wissenschaft Verlagsgesellschaft mbH. http://www.scilogs.de/impressum/ (last retrieved 20 April 2016).

Storrer, A. (2014). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In: A. Plewina & W. Andreas (Eds.), Sprachverfall? Berlin, De Gruyter, pp. 171-196.

Storrer, A. (2015). ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D. In CLARIN-D Website. http://de.clarin.eu/en/curation-project-1-3-german-philology (last retrieved 20 April 2016).

Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis

Björn Ross Michael Rist Guillermo Carbonell
Benjamin Cabrera Nils Kurowsky Michael Wojatzki
Research Training Group "User-Centred Social Media"
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen
firstname.lastname@uni-due.de

Abstract

Some users of social media are spreading racist, sexist, and otherwise hateful content. For the purpose of training a hate speech detection system, the reliability of the annotations is crucial, but there is no universally agreed-upon definition. We collected potentially hateful messages and asked two groups of internet users to determine whether they were hate speech or not, whether they should be banned or not and to rate their degree of offensiveness. One of the groups was shown a definition prior to completing the survey. We aimed to assess whether hate speech can be annotated reliably, and the extent to which existing definitions are in accordance with subjective ratings. Our results indicate that showing users a definition caused them to partially align their own opinion with the definition but did not improve reliability, which was very low overall. We conclude that the presence of hate speech should perhaps not be considered a binary yes-or-no decision, and raters need more detailed instructions for the annotation.

1 Introduction

Social media are sometimes used to disseminate hateful messages. In Europe, the current surge in hate speech has been linked to the ongoing refugee crisis. Lawmakers and social media sites are increasingly aware of the problem and are developing approaches to deal with it, for example promising to remove illegal messages within 24 hours after they are reported (Titcomb, 2016).

This raises the question of how hate speech can be detected automatically. Such an automatic detection method could be used to scan the large amount of text generated on the internet for hateful content and report it to the relevant authorities. It would also make it easier for researchers to examine the diffusion of hateful content through social media on a large scale.

From a natural language processing perspective, hate speech detection can be considered a classification task: given an utterance, determine whether or not it contains hate speech. Training a classifier requires a large amount of data that is unambiguously hate speech. This data is typically obtained by manually annotating a set of texts based on whether a certain element contains hate speech.

The reliability of the human annotations is essential, both to ensure that the algorithm can accurately learn the characteristics of hate speech, and as an upper bound on the expected performance (Warner and Hirschberg, 2012; Waseem and Hovy, 2016). As a preliminary step, six annotators rated 469 tweets. We found that agreement was very low (see Section 3). We then carried out group discussions to find possible reasons. They revealed that there is considerable ambiguity in existing definitions. A given statement may be considered hate speech or not depending on someone's cultural background and personal sensibilities. The wording of the question may also play a role.

We decided to investigate the issue of reliability further by conducting a more comprehensive study across a large number of annotators, which we present in this paper.

Our contribution in this paper is threefold:

- To the best of our knowledge, this paper presents the first attempt at compiling a German hate speech corpus for the refugee crisis.[1]
- We provide an estimate of the reliability of hate speech annotations.
- We investigate how the reliability of the annotations is affected by the exact question asked.

[1] Available at https://github.com/UCSM-DUE/IWG_hatespeech_public

2 Hate Speech

For the purpose of building a classifier, Warner and Hirschberg (2012) define hate speech as "abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation". More recent approaches rely on lists of guidelines such as a tweet being hate speech if it "uses a sexist or racial slur" (Waseem and Hovy, 2016). These approaches are similar in that they leave plenty of room for personal interpretation, since there may be differences in what is considered offensive. For instance, while the utterance *"the refugees will live off our money"* is clearly generalising and maybe unfair, it is unclear if this is already hate speech. More precise definitions from law are specific to certain jurisdictions and therefore do not capture all forms of offensive, hateful speech, see e.g. Matsuda (1993). In practice, social media services are using their own definitions which have been subject to adjustments over the years (Jeong, 2016). As of June 2016, Twitter bans *hateful conduct*[2].

With the rise in popularity of social media, the presence of hate speech has grown on the internet. Posting a tweet takes little more than a working internet connection but may be seen by users all over the world.

Along with the presence of hate speech, its real-life consequences are also growing. It can be a precursor and incentive for hate crimes, and it can be so severe that it can even be a health issue (Burnap and Williams, 2014). It is also known that hate speech does not only mirror existing opinions in the reader but can also induce new negative feelings towards its targets (Martin et al., 2013). Hate speech has recently gained some interest as a research topic on the one hand – e.g. (Djuric et al., 2014; Burnap and Williams, 2014; Silva et al., 2016) – but also as a problem to deal with in politics such as the *No Hate Speech Movement* by the Council of Europe.

The current refugee crisis has made it evident that governments, organisations and the public share an interest in controlling hate speech in social media. However, there seems to be little consensus on what hate speech actually is.

[2]"You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.", The Twitter Rules

3 Compiling A Hate Speech Corpus

As previously mentioned, there is no German hate speech corpus available for our needs, especially not for the very recent topic of the refugee crisis in Europe. We therefore had to compile our own corpus. We used Twitter as a source as it offers recent comments on current events. In our study we only considered the textual content of tweets that contain certain keywords, ignoring those that contain pictures or links. This section provides a detailed description of the approach we used to select the tweets and subsequently annotate them.

To find a large amount of hate speech on the refugee crisis, we used 10 hashtags[3] that can be used in an insulting or offensive way. Using these hashtags we gathered 13 766 tweets in total, roughly dating from February to March 2016. However, these tweets contained a lot of non-textual content which we filtered out automatically by removing tweets consisting solely of links or images. We also only considered original tweets, as retweets or replies to other tweets might only be clearly understandable when reading both tweets together. In addition, we removed duplicates and near-duplicates by discarding tweets that had a normalised *Levenshtein* edit distance smaller than .85 to an aforementioned tweet. A first inspection of the remaining tweets indicated that not all search terms were equally suited for our needs. The search term *#Pack* (vermin or lowlife) found a potentially large amount of hate speech not directly linked to the refugee crisis. It was therefore discarded. As a last step, the remaining tweets were manually read to eliminate those which were difficult to understand or incomprehensible. After these filtering steps, our corpus consists of 541 tweets, none of which are duplicates, contain links or pictures, or are retweets or replies.

As a first measurement of the frequency of hate speech in our corpus, we personally annotated them based on our previous expertise. The 541 tweets were split into six parts and each part was annotated by two out of six annotators in order to determine if hate speech was present or not. The annotators were rotated so that each pair of annotators only evaluated one part. Additionally the offensiveness of a tweet was rated on a 6-point Likert scale, the same scale used later in the study.

[3]*#Pack*, *#Aslyanten*, *#WehrDich*, *#Krimmigranten*, *#Rapefugees*, *#Islamfaschisten*, *#RefugeesNotWelcome*, *#Islamisierung*, *#AsylantenInvasion*, *#Scharia*

Even among researchers familiar with the definitions outlined above, there was still a low level of agreement (Krippendorff's $\alpha = .38$). This supports our claim that a clearer definition is necessary in order to be able to train a reliable classifier. The low reliability could of course be explained by varying personal attitudes or backgrounds, but clearly needs more consideration.

4 Methods

In order to assess the reliability of the hate speech definitions on social media more comprehensively, we developed two online surveys in a between-subjects design. They were completed by 56 participants in total (see Table 1). The main goal was to examine the extent to which non-experts agree upon their understanding of hate speech given a diversity of social media content. We used the Twitter definition of *hateful conduct* in the first survey. This definition was presented at the beginning, and again above every tweet. The second survey did not contain any definition. Participants were randomly assigned one of the two surveys.

The surveys consisted of 20 tweets presented in a random order. For each tweet, each participant was asked three questions. Depending on the survey, participants were asked **(1)** to answer (yes/no) if they considered the tweet hate speech, either based on the definition or based on their personal opinion. Afterwards they were asked **(2)** to answer (yes/no) if the tweet should be banned from Twitter. Participants were finally asked **(3)** to answer how offensive they thought the tweet was on a 6-point Likert scale from 1 (Not offensive at all) to 6 (Very offensive). If they answered 4 or higher, the participants had the option to state which particular words they found offensive.

After the annotation of the 20 tweets, participants were asked to voluntarily answer an open question regarding the definition of hate speech. In the survey with the definition, they were asked if the definition of Twitter was sufficient. In the survey without the definition, the participants were asked to suggest a definition themselves. Finally, sociodemographic data were collected, including age, gender and more specific information regarding the participant's political orientation, migration background, and personal position regarding the refugee situation in Europe.

The surveys were approved by the ethical committee of the Department of Computer Science and Applied Cognitive Science of the Faculty of Engineering at the University of Duisburg-Essen.

5 Preliminary Results and Discussion

Since the surveys were completed by 56 participants, they resulted in 1120 annotations. Table 1 shows some summary statistics.

	Def.	No def.	p	r
Participants	25	31		
Age (mean)	33.3	30.5		
Gender (% female)	43.5	58.6		
Hate Speech (% yes)	32.6	40.3	.26	.15
Ban (% yes)	32.6	17.6	.01	-.32
Offensive (mean)	3.49	3.42	.55	-.08

Table 1: Summary statistics with p values and effect size estimates from WMW tests. Not all participants chose to report their age or gender.

To assess whether the definition had any effect, we calculated, for each participant, the percentage of tweets they considered hate speech or suggested to ban and their mean offensiveness rating. This allowed us to compare the two samples for each of the three questions. Preliminary Shapiro-Wilk tests indicated that some of the data were not normally distributed. We therefore used the Wilcoxon-Mann-Whitney (WMW) test to compare the three pairs of series. The results are reported in Table 1.

Participants who were shown the definition were more likely to suggest to ban the tweet. In fact, participants in group one very rarely gave different answers to questions one and two (18 of 500 instances or 3.6%). This suggests that participants in that group aligned their own opinion with the definition.

We chose Krippendorff's α to assess reliability, a measure from content analysis, where human coders are required to be interchangeable. Therefore, it measures agreement instead of association, which leaves no room for the individual predilections of coders. It can be applied to any number of coders and to interval as well as nominal data. (Krippendorff, 2004)

This allowed us to compare agreement between both groups for all three questions. Figure 1 visualises the results. Overall, agreement was very low, ranging from $\alpha = .18$ to .29. In contrast, for the purpose of content analysis, Krippendorff recommends a minimum of $\alpha = .80$, or a minimum of .66 for applications where some uncertainty is un-

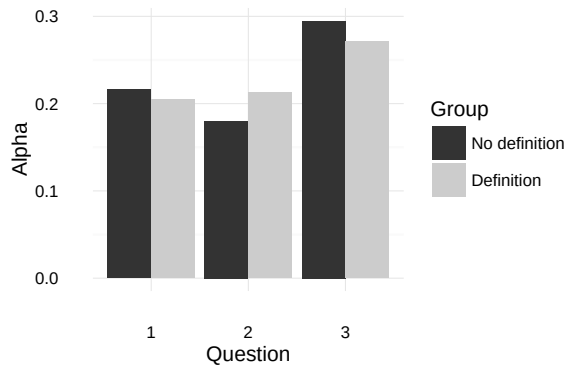


Figure 1: Reliability (Krippendorff's α) for the different groups and questions

problematic (Krippendorff, 2004). Reliability did not consistently increase when participants were shown a definition.

To measure the extent to which the annotations using the Twitter definition (question one in group one) were in accordance with participants' opinions (question one in group two), we calculated, for each tweet, the percentage of participants in each group who considered it hate speech, and then calculated Pearson's correlation coefficient. The two series correlate strongly ($r = .895, p < .0001$), indicating that they measure the same underlying construct.

6 Conclusion and Future Work

This paper describes the creation of our hate speech corpus and offers first insights into the low agreement among users when it comes to identifying hateful messages. Our results imply that hate speech is a vague concept that requires significantly better definitions and guidelines in order to be annotated reliably. Based on the present findings, we are planning to develop a new coding scheme which includes clear-cut criteria that let people distinguish hate speech from other content.

Researchers who are building a hate speech detection system might want to collect multiple labels for each tweet and average the results. Of course this approach does not make the original data any more reliable (Krippendorff, 2004). Yet, collecting the opinions of more users gives a more detailed picture of objective (or intersubjective) hatefulness. For the same reason, researchers might want to consider hate speech detection a regression problem, predicting, for example, the degree of hatefulness of a message, instead of a binary yes-or-no classification task.

In the future, finding the characteristics that make users consider content hateful will be useful for building a model that automatically detects hate speech and users who spread hateful content, and for determining what makes users disseminate hateful content.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

Peter Burnap and Matthew Leighton Williams. 2014. Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making. In *Proceedings of IPP 2014*, pages 1–18.

Nemanja Djuric, Robin Morris Jing Zhou, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2014. Hate Speech Detection with Comment Embeddings. In *ICML 2014*, volume 32, pages 1188–1196.

Sarah Jeong. 2016. The History of Twitter's Rules. *VICE Motherboard.*

Klaus Krippendorff. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *HCR*, 30(3):411–433.

Ryan C Martin, Kelsey Ryan Coyier, Leah M VanSistine, and Kelly L Schroeder. 2013. Anger on the Internet: the Perceived Value of Rant-Sites. *Cyberpsychology, behavior and social networking*, 16(2):119–22.

Mari J Matsuda. 1993. *Words that Wound - Critical Race Theory, Assaultive Speech, and the First Amendment*. Westview Press, New York.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of ICWSM 2016*, pages 687–90.

James Titcomb. 2016. Facebook and Twitter promise to crack down on internet hate speech. *The Telegraph*.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of LSM 2012*, pages 19–26. ACL.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of NAACL-HLT*, pages 88–93.

Towards the Harmonization and Segmentation of German Hashtags

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

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

Abstract

We present on-going work on the harmonization and segmentation of German hashtags. Our aim is to reduce the number of variants of hashtags expressing the same content to one harmonized hashtag that can thus serve as a unique "annotation tag" for a large set of tweets.

1 Introduction

When looking at hashtags used in Twitter posts (and probably in all social media) one can observe that one content is often expressed by various hashtags, whereas the degree of variance between the hashtags can heavily differ. Sometimes only the use of lowercase vs. uppercase letters marks the variance, like #EM2016 vs. #em2016. But there are more complex variants, as shown by the use of abbreviations or acronyms, like for example #EURO2016 vs. #europeanchampionship2016 or #Europameisterschaft2016. For the human reader, if she understands both English and German, the three hashtags are clearly related to the soccer event that took place in France in 2016. But for the machine processing of tweets and for supporting queries to them, it might be useful to formally establish this relationship. Both #europeanchampionship2016 and #Europameisterschaft2016 could be marked as a variant of #EURO2016 (or of #euro2016) or vice versa.

While Declerck & Lendvai (2015b) describe a proposal for the formal representation of such hashtag variants, there is, at the best of our knowledge, not yet any implemented method for detecting and marking such hashtag variants in German tweets.

We expect this harmonization step to also improve results of queries addressed to social media, as this has already been suggested in (Berardi et al., 2011).

2 Related Work

Our investigation dealing with the harmonization and segmentation of German hashtag in Twitter posts is influenced by the work applied to hashtags used in English tweets (Declerck & Lendvai, 2015a). Kotsakos et al. (2015) are proposing a very interesting approach to the filtering of meme-hashtags, including German hashtags, but the hashtags harmonization step they implement is limited to lowercasing.

3 Use of Hashtags in German Twitter Texts

An interesting aspect of hashtags in English posts is that they are showing a move to compounding, generating more and more "glued" word constructions, which are not only in use in social media, but are also getting more popular in "classical" text. This is making word decomposition a more and more relevant task for the automated analysis of English.

Now, compounding is an important feature of German and there exist already some segmentation algorithms for the analysis of German text.[1] But we see that the "compounding" mechanisms applied to hashtags are showing relevant differences to the compounding rules applied to the generation of "normal" text. There is therefore a need to develop specific algorithms for the decomposition of German hashtags. And this is even more necessary if one considers the fact that German tweets are making a large use of hashtags, substantially more as in English tweets, as this has been reported in (Weerkamp et al., 2011) on a comparative study of Twitter texts in several languages. This study reveals that 14% of English tweets are tagged by a hashtag, whereas 25% of German tweets include a hashtag. And German tweets used significantly more hashtags

[1] See for example (Henrich & Hinrichs, 2011).

than English ones: 1.9 hashtags per tweet, compared to 1.4 for English tweets.

3.1 Examples of German Hashtags

In this section we present few examples of hashtags we found by just reading some German tweets.

In a first case we are dealing with normal German words, which can simply be lowercased for achieving the intended harmonization:

```
#europameisterschaft, #Euro-
pameisterschaft, #EUROPA-
MEISTERSCHAFT, #EUROPAmeis-
terschaft
#Europameisterschaft2016,
#europameisterschaft2016
```

But compared to English hashtags, we can see here that we have "classical" compounds within the hashtag, and thus no use of camelCase can further help for segmenting[2]. In this case we will use just standard segmentation algorithms for German.

In a second case, we are dealing with abbreviated hashtags, which can also be lowercased:

```
#EM2016, #em2016
```

While those cases are straightforward candidates for harmonization, it is a bit more challenging to reduce `#Europameisterschaft2016` to `#em2016`, also due to the fact that no camelCase is used (in this case one could relate "E" and "M" to "em"). For this we take advantage of the use of non-classical compound effects, like the addition of digits at the end of the hashtag. An indicator is also given by the fact that those distinct hashtags are sometimes used in the same tweet, although this is more often the case with the use of the `#EURO2016` hashtag. In fact the latter hashtag can be used as a pivot over tweets in different languages, but we are not dealing with multilingual issues in this study.

A third case is given by examples like:

```
#StandortDeutschland
#FußballEM2016
```

We assume that the use of camelCase notation in German tweets is really an indicator of non-classical compounding, so that we can segment the hashtag here, and possibly harmonize it with the following sequence: `#Fußball-#Europameisterschaft 2016`. The last example is an interesting one: grouping two hashtags via a hyphen sign, which seems to be specific to German hashtags.

A fourth case is given by:

```
#Brexit-Befürworter
#Brexit-Votum
```

This case is very similar to `#Fußball-#Europameisterschaft` with the difference that the word after the hyphen sign is not a hashtag. We can here harmonize to `#Brexit` – and ultimately to #brexit -- just storing the second word as a modifier.

A fifth and more complicated case is:

```
#warumeuropa
#bestemannschaft
#mussverscrapptwerden
```

Those cases show examples of real chunks or phrases included in one hashtag. It is still not clear if there is any advantage in trying to segment to cases.

4 Our Approach for Segmenting German Hashtags

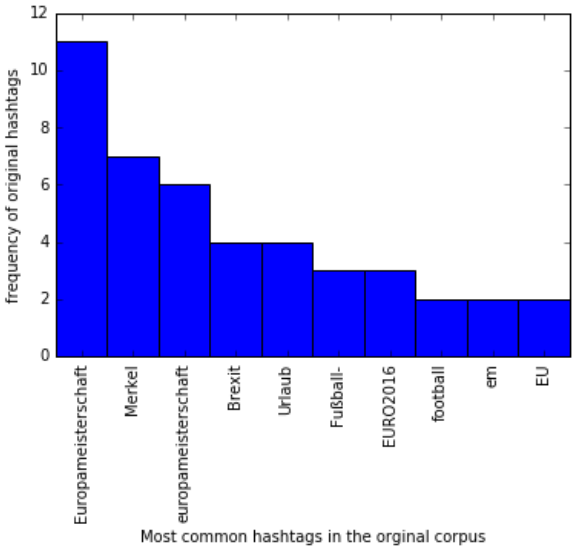
At the current stage of our investigation we have been implementing solutions for the four first cases mentioned in the preceding section. The data basis for our first experiment consists of 164 hashtag tokens just copied from some Twitter threads. The main topics were the 2016 European Championship in soccer[3] and the Brexit[4]. For now, we want to test some few algorithms on this small data set. In a next step we will apply and evaluate the algorithms to larger corpora.

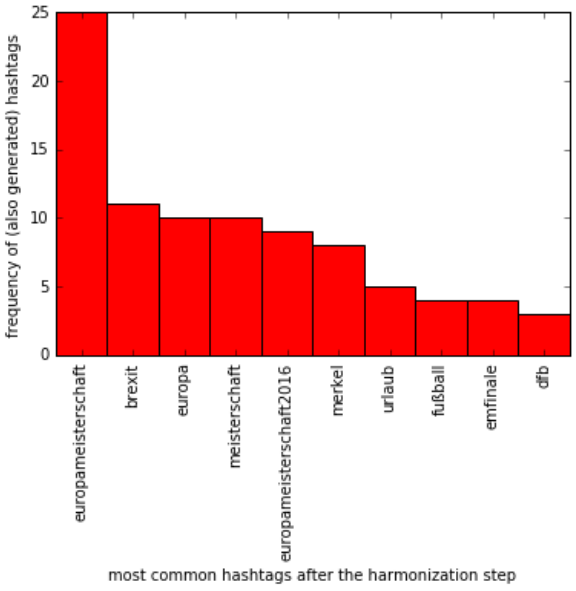
Below we display two charts showing first the 10 most frequent hashtags in our small data set before any segmentation and harmonization steps. The second chart shows the frequency of hashtags or words that result from the application of the current version of our segmentation and harmonization process to our data set.

[2] The intensive use of camelCase notation in English hashtags was a feature helping to segment those in the study reported in (Declerck & Lendvai, 2015a).

[3] https://en.wikipedia.org/wiki/UEFA_Euro_2016

[4] https://en.wikipedia.org/wiki/Brexit





We observe some significant changes in the ranking, so that in the second chart the term "brexit" is now the second in frequency (we had in the original data set both #Brexit" and "#brexit" as standalone hashtags but also compounds like "#Brexit-Befürworter" or "BrexitVote").

Interesting is also the emergence of the term "meisterschaft", which was not appearing as a standalone hashtag in the original collection. But as in this case they were some examples of camelCase notation, the term "meisterschaft" has been extracted by the algorithms.

We also observe the rise of frequency for the harmonized hashtag "europameisterschaft". This is not only resulting from the addition of the frequency of the uppercase term "Europameisterschaft", but also to an acronym resolution step linking "em" to "europameisterschaft" (and "EM" to "Europameisteschaft"). As a consequence, the hashtags "#em" and "#EM" are deleted from the ranking list, and other topics are now visible in this list. It is for now not clear which hashtag should be selected as the harmonized one: we expect the application scenarios to specify this point.

We are currently analyzing those first results, while we immediately see that we have to mark the context or the domain in which "europa" or "meisterschaft" are occurring. The same is valid for "brexit". This aspect is relevant for queries: we aim at suggesting this context to the users submitting the queries.

We are currently working on processing also the hashtags containing numerical and other special symbols and extending the investigation to a larger selection of hashtags, also in the full context of the tweets they are occurring in.

Acknowledgments

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

Reference

Giacomo Berardi, Andrea Esuli, Diego Marcheggiani, and Fabrizio Sebastiani. 2011. ISTI @ TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking. Proceedings of the 20th Text Retrieval Conference (TREC 2011), Gaithersburg, US, 2011.

Thierry Declerck, Piroska Lendvai. 2015a. Processing and Normalizing Hashtags. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitko (eds.): Proceedings of RANLP 2015, Pages 104-110, Hissar, Bulgaria.

Thierry Declerck, Piroska 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.

Henrich, V. & E. Hinrichs (2011). Determining Immediate Constituents of Compounds in GermaNet. In: *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*. Hissar, Bulgaria. pp. 420-426.

Dimitrios Kotsakos, Panos Sakkos, Ioannis Katakis, Dimitrios Gunopulos. 2015. Language agnostic meme-filtering for hashtag-based social network analysis. Social Network Analysis and Mining 5 (1), 1-14

Wouter Weerkamp, Simon Carter and Manos Tsagkias. 2011. How People use Twitter in Different Languages. In: Proceedings of Web Science.