**IJCLA VOL. 3, NO. 1, JAN-JUN 2012, PP. 71–88**
**RECEIVED 30/10/11 ACCEPTED 09/12/11 FINAL 13/06/12**

# Detecting Irregularities in Blog Comment Language Affecting POS Tagging Accuracy

MELANIE NEUNERDT, BIANKA TREVISAN,
RUDOLF MATHAR, AND EVA-MARIA JAKOBS

*RWTH Aachen University, Germany*

ABSTRACT

*Studying technology acceptance requires the survey and analysis of user opinions to identify acceptance-relevant factors. In addition to surveys, Web 2.0 poses a huge collection of user comments regarding different technologies. Extracting acceptance-relevant factors and user opinions from such comments requires the application of Natural Language Processing (NLP) methods, particularly Part-of-Speech (POS) tagging. Applied to typical blog language POS tagging often suffers from high error rates. In this paper, we present multiple user-specific studies of blog comments to analyze the relation between blog language and performance of NLP methods. We propose an approach, which leads to enhancement of POS tagging and lemmatizing quality. Furthermore, we present an ontology-based corpus generation tool to improve the identification of topic- and user-specific blog comments. Utilizing the generation tool, a corpus dealing with mobile communication systems (MCS) is exemplarily created. Furthermore, we analyze and transform the identified comments into structured datasets.*

KEYWORDS**:** *Natural Language Processing, Part-of-Speech Tagging, weblog, user writing style, ontology search, corpus generation.*

## 1 INTRODUCTION

Typical instruments used in acceptance research are questionnaires, interviews, or focus groups. The according outcomes reveal user opinions about a particular topic such as MCS. Nevertheless, traditional methods in acceptance research have the major shortcoming of being subject to numerous methodological effects. In surveys, respondents tend to answer dishonestly (*Social-Desirability-Response-Set*). As a result, *artificial* or *falsified data* is collected. Moreover, the utilization of traditional surveys calls for high user willingness, hence collecting a sufficient amount of data is a very arduous and time-consuming task. However, this type of data provides the advantage of being structured in a predefined manner, which ensures the availability of information as well as efficient data access. As a complement to traditional methods in acceptance research, we propose an innovative approach in which natural language discourses from web data, such as blog comments, are analyzed with the aim of identifying frequently discussed topics in a particular field and current user evaluations on this topic. Particularly, this approach benefits from the fact that the data is *natural* or *authentic*, that it is accessible, and that it might be downloaded quite efficiently.

However, the analysis of natural language discourses is problematic since Internet users tend to a more colloquial formulation or expression style. More precisely, the language used in blogs suffers from numerous lexical, syntactic, semantic, stylistic, and typographical phenomena, e.g., unconventional use of punctuation marks such as *?!?!*, novel typographical means of evaluation such as :-), or frequent use of interjections such as *haha*. Common NLP methods such as POS tagging cannot process these text type-specific phenomena correctly and, in consequence, high error rates appear. As a result, no exact statements can be made. Usually, POS tagging is the first step of text processing for further text analysis. The output can be used for further NLP processing steps, e.g., opinion detection. Therefore, a high POS tagging accuracy is very important for further investigations. Modern taggers achieve a per-word-accuracy of 97.53% when tagging German newspaper corpora [1]. Unfortunately, the accuracy drops significantly when applying taggers to web corpora [2]. First, low tagging accuracy is caused by the use of topic-specific terms and abbreviations, e.g., *Datenbandbreite (data bandwidth), iPhone, UMTS*. Second, blog comments are non-standardized texts, and characterized by different users' writing styles. Therefore, POS tagging has to be adapted to blog-

specific linguistic phenomena. Our work focuses on the identification of causes for POS tagging decision errors in blog comments. For this purpose, an ontology-based corpus generation tool is developed which is used to create a topic-specific corpus. The corpus is analysed for blog-specific linguistic phenomena.

The paper is structured as follows. In Section 2, the ontology-based corpus generation tool is introduced. Afterwards, the created corpus is presented in Section 3. In Section 4, empirical results of the corpus analysis on blog-specific linguistic phenomena are shown. In section 5, we present a first sketch of a POS tagger adapted to the language used in blog comments. Section 6 presents an example for blog comment transformation into suitable data representation. Finally, we present our conclusion.

## 2 Ontology-Based Corpus Generation

The increasing number of user-generated web content provides a large amount of opinionated data. However, people express their opinion differently and use different terminology in written web discussions. Hence, it becomes hard to access and extract topic-specific user opinions by simple keyword search. Particularly, usage of ontology-based search that considers keyword relations, such as synonymy, leads to an increase of quantity and quality of retrieved search results. Our tool named CROW can be used to selectively search for blog comments. It offers the ability to create a corpus according to a predefined ontology. The resulting corpus serves for further linguistic analysis and mathematical calculations.

### 2.1 *Ontology principles*

Ontologies play an important role in the field of knowledge engineering and semantic web research. Typical applications are ontology learning [3] and ontology-based focused crawling [4,5]. Ontologies contain a collection of concepts represented by terms that exist in a certain domain. The relations are determined according to linguistic usage or to human semantic association, respectively. Thus, ontologies can be described by a directed graph where the nodes represent concepts and the edges represent semantic relations. In this work, the edges describe the dependency between MCS components, properties,

and instances. The ontology consists of four relation types: (1) Hierarchical and (2) non-hierarchical relations as described in the literature [6,7], (3) attributive relations, and (4) type-token relations.

(1) Hierarchical relations are hyperonymy and meronymy. With these relations, over- and subordinated concepts (e.g. phone as a hyperonym of mobile phone) as well as part-whole relations between concepts (e.g. display as a part of mobile phone) of MCS can be expressed [6].
(2) A synonymy describes a non-hierarchical relation. The relation links concepts (terms) that have an identical or similar meaning, e.g., mobile phone and cell phone. In the ontology, synonyms are summarized into a term set (synset) [7].
(3) Attributive relations indicate which properties or utilization-types are ascribed to an object. The relation property indicates object properties, e.g., robustness and longevity. With the relation association concepts are connected to each other that have no lexical-semantic connection. Rather, they are related to each other on the basis of user experiences or usage types.
(4) The type-token relation instance assigns real world examples or class representatives to classes, e.g., *iPhone* a respresentative for *cell phones*.

### 2.2 *Tool functionalities*

CROW is a web application providing a graphical user interface. The application enables the user to specify an ontology manually or to reload it from the storage; see Figure 1. Furthermore, a comment corpus can be selected or filtered by user name or time stamp.

Different edge types illustrate different relation types. Various statistic scores for the underlying comments and the concept related terms are calculated and plotted. In addition to the overall number of comments, e.g., the *document frequency* (*DF*) including single concepts and combination of related concepts is computed. It shows the co-occurrence of concepts in user discussions. For all related concepts, the correlation coefficient is calculated between the *term frequencies* (*TF*). The correlation strength is indicated by the color of the relation line according to a given color table; see Figure 1. These values support the user evaluating relevant blog comments and give an impression of the used terminology regarding MCS. Furthermore, users can use a particular ontology path for comment extraction according to the

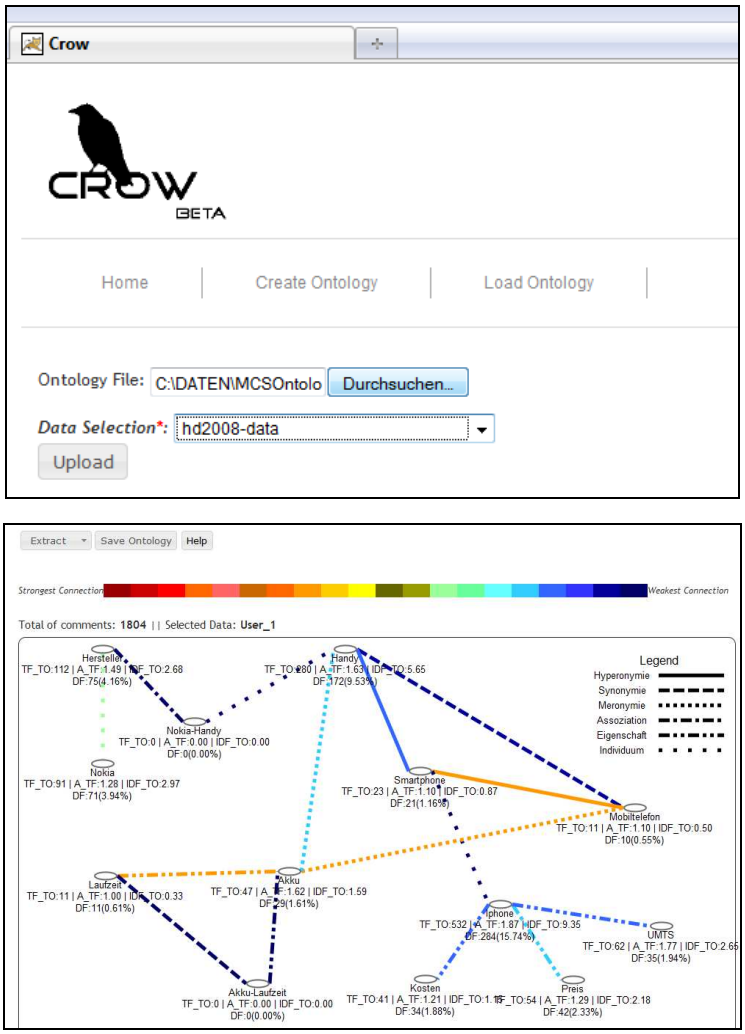


**Fig. 1.** User interface and ontology visualization (CROW).

statistical results and their interpretations. For instance, the correlation coefficient allows identification of how often different users use a synonym. Hence, the topic-selection can be refined and a blog comment corpus with high relevance to MCS terminology can be generated.

Two types of comment extraction options are provided: First, the intersection set of comments covers all comments including all concepts of the selected ontology path (sub ontology). Second, the union set of comments contains at least the occurrence of one concept in the selected sub ontology.

## 3 BLOG COMMENT CORPUS

For blog comment analysis, a set of approximately 166 thousand German blog comments posted from January 2008 to December 2009 is considered. In a number of preprocessing steps, posted comments are bowdlerized from enclosing webpage elements, html-tags, and corresponding meta information, e.g., user name is extracted and added as metadata to the comment. Table 1 illustrates some statistics about the corpus collection, particularly, in terms of covered users and their blogging frequency. Comparing the statistical values for both years shows that the data corpus for 2008 and 2009 follows a very similar distribution.

**Table 1.** Comment corpus statistics.

| | Data 2008 | Data 2009 |
|---|---|---|
| Articles | 1,252 | 1,289 |
| Comments | 84,203 | 81,831 |
| Users | 10,474 | 9,509 |
| #Comments per article | 67 | 63 |
| #Comments per user | 8 | 9 |

For further analysis, we use the ontology-based corpus generation tool described in Section 2. Therefore, a specific topic of MCS is focused. As an example, we select the topic *Handy (cell phone)*; the sub ontology is created manually and is part of the whole MCS ontology. This ontology is used to extract the MCS corpus.

Based on the generated MCS corpus, all users are ranked in decreasing order according to their posting frequency. We assume that users who post regularly tend to develop their own writing style. Among these users non-standardized expressions, colloquial expressions, and emoticons are used more frequently. Moreover, grammar rules are violated more often than in comments posted by users that post only once or twice. Considering the posting frequency distribution, the 12 most active users are selected for further statistical

and linguistic analysis. The 3,249 comments of the 12 users form the analysis corpus with a total of 675,762 tokens. The sample corpus is used to identify and evaluate acceptance-relevant factors or user opinions considering the design of user devices, e.g., cell phones.

## 4 Analysis of Writing Style Irregularities

Blog-specific writing styles lead to processing errors in NLP methods, e.g., POS tagging. Analysing and evaluating users' writing styles aimes at enhancing NLP methods with respect to blog comment processing and, in particular, adapting existing approaches to the language in blog comments for automatic opinion extraction. To detect user opinions, all text characters, e.g., emoticons, are important. Hence, bowdlerizing comments is not a sufficient solution for our task. The goal is to investigate causes for decision errors in NLP methods and handle those text irregularities.

Opinion detection in the field of text retrieval is still a challenging task [8,9]. Two different approaches are used for opinion detection and classification: First, machine learning-based approaches based on training data. And second, lexicon-based approaches, which use a lexicon of sentimental words, e.g., list of positive and negative words, provided by linguistic quantization and weightings [10,11].

### 4.1. *User-specific statistical analysis*

In this section, some statistical analysis is performed to identify non-standardized text-patterns in blog comments. We address the task of detecting different writing styles, which need to be considered for enhancing the accuracy of NLP methods. Blog comment users evaluate objects by using different non-standardized evaluative expressions. Therefore, we choose features according to three different types of writing style: 1) emoticon usage, 2) usage of colloquialisms, and 3) punctuation marks usage. Representatives or examples of the different expression types are counted by frequency, which is the basic and most popular feature set used in text classification and clustering tasks [11,12]. Consequently, we count the frequencies of all features. To make the results comparable, we normalize each feature value by the feature-specific maximum, determined over all users.

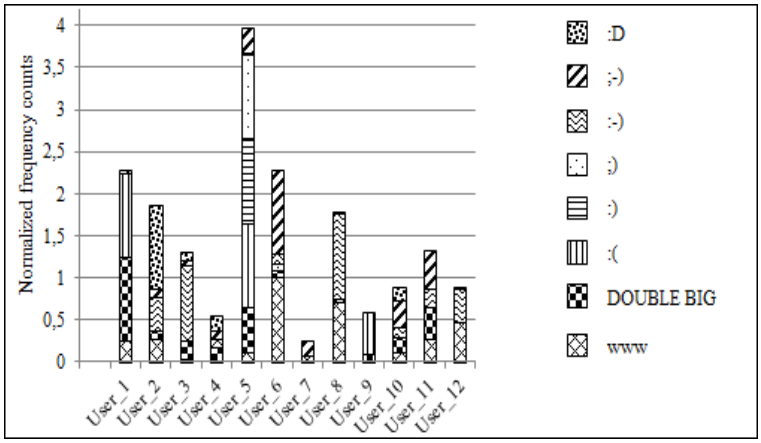


**Fig. 2.** User-specific emoticon usage.

**Emoticon usage.** To measure the occurrence of emoticons, we count the six most common emoticons composed of two and three characters. Considering the goal of opinion extraction, three emoticons with positive expression (*:D, :), :-)* ) two ironic emoticons (*;), ;-)* ), and one emoticon with negative expression (*:(* ) are considered. Empirical results show that only 5 out of 70 (< 10%) users do not use any emoticons to express their opinion. Figure 2 shows the distribution of used emoticons for the first 12 users based on the comment corpus described in Section 3.

**Colloquialisms.** A large amount of comment data shows that German grammar is generally not maintained. Users are very modest in using capitalization and produce numerous orthographical errors and letter transpositions. Furthermore, they tend to shorten words, e.g., *ne, draus (hence), drum (therefore),* and introduce new terms to express their opinions. Therefore, we suggest measuring the level of colloquialism and introduce a number of different features.

Firstly, we use count of words features to create manually a small list of terms typically used in comments, e.g., *lol, haha, hehe, nö, ne, naja*. Secondly, we count the number of complete capitalized words and words where the first two letters are capitalized, e.g., *TElefon.* The second type of words has a high occurrence in user comments due to fast writing style. Since there are numerous capitalized acronyms where at least the first two letters are capitalized, e.g., *WLAN, UMTS, GBit,*

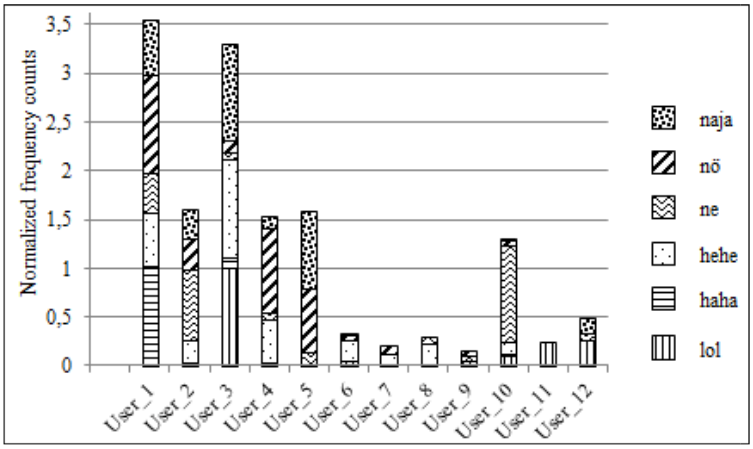


**Fig. 3.** User-specific colloquial expressions.

this has to be considered in counting. Therefore, we construct a list with acronyms that are related to the context of MCS. Our data corpus, including articles and comments, serves as basic component to determine context-relevant acronyms. The list is generated automatically using some regular expressions. Finally, we count forgotten space characters after dots and commas, due to users' carelessness. In order to avoid counting of digits, including commas or punctuations. Figure 3 shows the results.

**Punctuation usage.** To describe the punctuation usage, we use frequency counts of various punctuation marks. Users disregard punctuations on one hand and introduce new ways of punctuation to express the intensity of their opinion on the other hand. Therefore, we distinguish between two types of punctuations: single punctuations used in the conventional manner according to the German punctuation rules and multiple punctuations which indicate evaluative utterances, e.g., *???, !!!, *, and '*. Multiple punctuations are measured by counting bigrams and trigrams in the comment. The results are depicted in Figure 4.

To measure the degree of irregularity, we consider all feature types together. Therefore, we sum up the feature values for each user. The result is depicted in Figure 5.

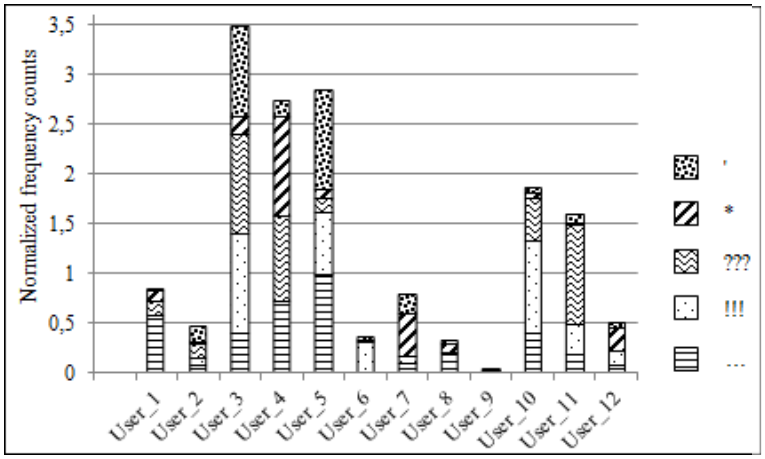


**Fig. 4.** User-specific punctuation usage.

### 4.2. *User-specific evaluation style analysis*

Written user comments in weblogs are characterized by text type-specific expressions and formulation styles, e.g., colloquial or ironic expressions. In Section 4.1 statistical analysis concerning the usage of specific non-standardized tokens/expressions is performed, whereas in this section evaluative expressions containing more than one word are analysed. The focus is to detect sequences of evaluative expressions in blog comments, which need to be considered for the extraction of user opinions. Some evaluative expressions are collected systematically and defined for the purpose of automatic processing.

The blog comments are analysed manually using the content analysis software MaxQDA. The manual analysis is a necessary preliminary data investigation; the results form the linguistic knowledge basis for the subsequent automatical extraction of user evaluations. Each user comment of the corpus is sifted in order to identify evaluative expressions. If an evaluative expression is discovered, all relevant parts of the expression are marked and categorized according to the literature [13,14,15]. The text segments are assigned to the following categories: Noun phrase, I-sentence, dialection, weighting, irony, negation, onomatopoeia, idiom, relationalization, rhetorical question, and comparison.

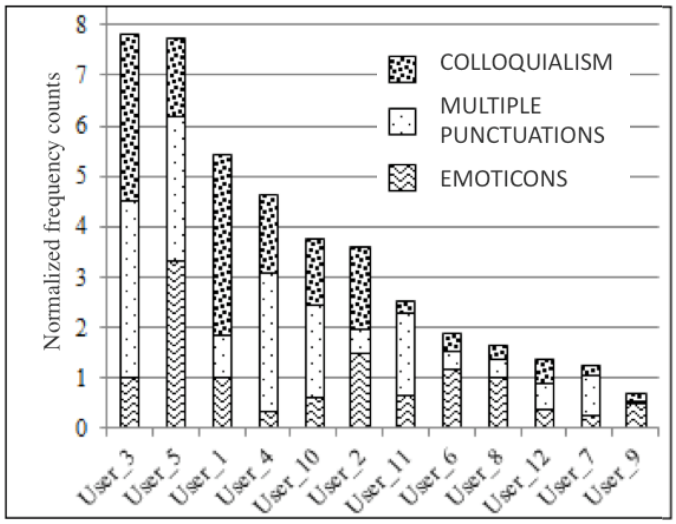


**Fig. 5.** User ranking: grammatical accuracy combining all feature types.

(1) Noun phrase: Expressions that consist of a valuing adjective and a noun;
(2) I-sentence: Expressions that express the stance and attitude of the user;
(3) Dialection: Expressions that are weakened or reinforced by question-answer phrases;
(4) Weighting: Expressions, that indicate the weighting of the evaluation aspect or topic;
(5) Irony: Expressions in which the writer says the opposite of what he means. Statements contrary to regular, shared knowledge and opinions of society;
(6) Negation: Expressions that get a negative rating by using a negation particle;
(7) Onomatopoeia: Expressions in the form of a loud imitation of a natural or other non-linguistic acoustic phenomenon;
(8) Idiom: Evaluative expressions that consist of a language typical phrase;
(9) Relationalization: Expressions that weaken an opinion or statement or put it into perspective;
(10) Rhetorical question: Expression, in which a question is used to intensify an opinion;

(11) Comparison: Expressions, in which two or more objects or evaluation aspects are compared with each other.

To ensure the reliability of the results, two independent coders analyze the comments. Lastly, the coders check their categorization results in order to reach coding agreement (*intercoder reliability*). The evaluation of the categorization results is performed numerically: the coding frequency for each type of evaluative expression and user are counted. The results of the analysis are summarized in Table 2.

**Table 2.** Users with the highest number of evaluative expression and corresponding categories.

| | User 7 | User 12 | User 1 | User 5 | User 2 |
|---|---|---|---|---|---|
| #blog comments | 370 | 39 | 475 | 316 | 418 |
| #tokens | 12,3213 | 9,632 | 10,5764 | 53,364 | 53,148 |
| Noun phrase | 15 | 2 | 2 | 14 | 6 |
| I-sentence | 35 | 61 | 21 | 13 | 22 |
| Dialection | 7 | 0 | 0 | 0 | 0 |
| Weighting | 20 | 3 | 10 | 12 | 10 |
| Irony | 15 | 37 | 8 | 2 | 5 |
| Negation | 5 | 0 | 3 | 2 | 3 |
| Onomatopoeia | 0 | 1 | 7 | 1 | 1 |
| Idiom | 6 | 5 | 2 | 7 | 3 |
| Relationalization | 4 | 0 | 1 | 2 | 7 |
| Rhetorical question | 6 | 7 | 11 | 1 | 4 |
| Comparison | 18 | 3 | 16 | 20 | 10 |
| #Total codings | 131 | 119 | 81 | 74 | 71 |

Regarding Table 2, it is evident that users evaluate differently. Furthermore, the frequencies of evaluative expressions differ a lot. For instance, 39 comments of User 12 build only a small fraction of the corpus. Nevertheless, 119 evaluative expressions are identified in the user's posted comments. Compared to the other users, User 12 evaluates on MCS more frequently. The ratio between the amount of evaluative expressions and the number of analysed comments is in balance for the other users. Furthermore, we observe that there is a kind of user preference for the expression of certain evaluative expression types. While User 7, 12, and 5 most often use noun phrases, I-sentences, weightings, and irony, users 1 and 2 use more frequently evaluative types like rhetorical questions. Examples 1 to 3 show a selection of user utterances from the analyzed corpus.

*Ein Fass ohne Boden.* / *A bottomless pit.* (1)
(Idiom, User 12)

*Wo gibt es perfekte Geräte, die für alle genau richtig sind?* / *Where there are perfect devices that are just right for all?* (2)
(Rhetorical question, User 1)

*App Store liefert dann noch diverse "nice to have" Erweiterungen.* / *App Store still provides several "nice to have" extensions.* (3)
(Noun phrase, User 7)

Evaluative expressions like these cause processing errors. For instance, the idiom in (1) would not be recognized as a phrase. Instead, each word would be extracted and processed separately. In the same way, analysis and evaluation errors would occur for (2). Furthermore, the use of anglicism is problematic (3), since POS tagging tools do not reliably recognize foreign words.

## 5 Enhancement of POS Tagging Results

A number of approaches aim at the enhancement of NLP methods by means of preprocessing with the goal to bowdlerize comments. Since we want to detect user opinions, all text characters, e.g., emoticons and multiple punctuations, are important for interpretation. Therefore, our approach does not remove characters but further enables NLP methods to handle such irregularities [16]. To analyze the accuracy of NLP methods applied to blog comments, we choose a POS tagger as an example. A POS tagger annotates every word with a POS tag and a lemma. A tool for automatic German text corpora annotation is the TreeTagger, developed at the University of Stuttgart [17]. The TreeTagger adds a POS tag according to the Stuttgart-Tübingen Tagset (STTS) and a lemma according to a special lexicon to each word [17].

In our approach, the TreeTagger is improved by using a blog-specific lexicon (BS lexicon) as well as a topic-specific lexicon (MCS lexicon). The BS lexicon contains blog-specific expressions and is created according to the results of sections 4.1 and 4.2. Analysing the POS tagging results without the BS lexicon shows that comments indeed suffer from a high number of incorrect POS tags and unknown lemmas, but the results for neighboring words in the same sentence are not negatively affected. Hence, integrating blog-specific expressions by means of a lexicon enhances POS tagging results. Table 3 shows a tagged sentence part starting with a word not contained in the lexicon.

**Table 3.** Extract of the tagged corpus.
Incorrect POS tags are marked with a frame.

| | **Without auxiliary lexicon** | | **With auxiliary lexicon** | |
|---|---|---|---|---|
| **Token** | **POS tag** | **Lemma** | **POS tag** | **Lemma** |
| *Hmm* | NN | <unknown> | ITJ | Hm |
| , | $, | , | $, | , |
| *komisch* | ADJD | komisch | ADJD | komisch |
| *dass* | KOUS | dass | KOUS | dass |
| *ich* | PPER | ich | PPER | ich |
| [...] | | | | |
| *HSDPA* | NN | <unknown> | NE | HSDPA |
| *und* | KON | und | KON | und |
| *EDGE* | NN | <unknown> | NE | EDGE |
| *Sender* | NN | Sender | NN | Sender |

The example shows that the colloquial expression *Hmm* does not affect the POS tagging results in the remaining sentence. Thus, it is sufficient to define the term in the BS lexicon. Tagging results considering the MCS and BS lexicon are shown in the right part of the table. The MCS corpus is used to create the auxiliary lexica, whereas the comments posted by the selected 12 users are not considered. Moreover, the selected comments of the 12 users serve as a test sample for lexicon evaluation. Finally, the lexicon contains about 2,000 topic-specific terms related to MCS [18]. The MCS lexicon works as an auxiliary lexicon and complements the embedded standard lexicon of the TreeTagger.

The evaluation part (a test sample of about 3,500 comments and 390,000 token) is tagged with (a) and without (b) using the MCS lexicon and BS lexicon. As a result, the mean number of tagging errors dropped from (a) 9.44% to (b) 7.58% (mean value). Table 4 illustrates the improvements that are achieved.

In the first column of Table 4, the error rates applying the standard TreeTagger are shown. The ranking according to these error rates is strongly related to the ranking shown in Figure 4, which confirms our assumption that the POS tagging accuracy is strongly related to the degree of irregularity in users' writing style. Furthermore, in the rightmost column the improvements for different users are depicted.

**Table 4.** User-specific POS tagging results (in %).

| | Tagging Error Rates | | Improvement |
|---|---|---|---|
| | without auxiliary lexica | with auxiliary lexica | |
| User 1 | 7.64 | 5.95 | 1.69 |
| User 2 | 9.69 | 6.91 | 2.78 |
| User 3 | 12.15 | 9.09 | 3.06 |
| User 4 | 9.32 | 7.87 | 1.45 |
| User 5 | 12.20 | 10.16 | 2.04 |
| User 6 | 8.74 | 6.68 | 2.06 |
| User 7 | 7.91 | 5.91 | 2.00 |
| User 8 | 8.59 | 7.23 | 1.34 |
| User 9 | 10.64 | 8.71 | 1.93 |
| User 10 | 11.22 | 8.52 | 2.70 |
| User 11 | 10.36 | 8.35 | 2.01 |
| User 12 | 7.98 | 6.90 | 1.08 |
| Average | 9.7 | 7.69 | 1.86 |

For some users, a very high improvement in the tagged data is achieved when using the auxiliary lexica (e.g. User 3, User 10); in contrast, the tagged results of other users change little (e.g. User 9, User 12). In total, the tagging accuracy for each user is improved.

## 6 DATASET GENERATION

The overall goal of this work is to extract user opinions from comments and to generate suitable datasets for the integration into an acceptance model. By means of the ontology-based corpus generation tool, some user comments dealing with user devices (cell phones) as a sub-topic of MCS are extracted. With respect to the results in Sections 4.1 and 4.2, the extracted user comments are evaluated with the aim of creating user-specific datasets. Table 5 depicts dataset examples for two users evaluating five components of particular user devices. For representation of acceptance strength, we choose a scale from 1 (low acceptance) to 6 (high acceptance). The mapping is performed according to used features described in Section 4.1, e.g., emoticons or multiple exclamation points, and evaluative expressions listed in Section 4.2. Value -99 denotes that no information on the stated issue is available. Furthermore, interpreting user expressions allows for the extraction of some demographic information; see Table 5, column 2.

**Table 5.** Example datasets constructed based on user comments.

| User | Occupation | #Comments | Object | Display | Material | Usability | Camera | Battery | Reference |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Emp-loyed | 1,804 | iPhone | -99 | 2 | 3 | 1 | 4 | Nokia (E 51), iPod |
| | | | Nokia E51 | -99 | 2 | 1 | 1 | -99 | |
| 5 | Emp-loyed | 796 | iPhone | 6 | 6 | 5 | 4 | -99 | Siemens S55 |
| | | | Siemens S55 | 3 | -99 | 5 | -99 | -99 | |

## 7 CONCLUSION

In this work, we presented a study of text type-specific writing styles in blog comments to identify causes for NLP decision errors. By means of an ontology tool, a topic-specific blog comment corpus is created. Based on this corpus, we perform an analysis of text type-specific linguistic phenomena such as punctuation marks, emoticon usage, and colloquial expressions. Results show that the degree of grammatical or stylistic irregularity in blog comments differs significantly for different users. As an example for NLP methods, TreeTagger results based on comments for 12 different users are presented and discussed. Combination of the TreeTagger with topic- and blog-specific lexica enhances the tagging results for blog comments. The results show that the error rates as well as improvements are strongly related to the degree of irregularity in the users' writing style.

Nevertheless, further improvement, particulary with respect to opinion detection in blog comments, is desirable. A still unsolved task is the correct tagging of emoticons and multiple punctuation marks, which is crucial for opinion detection. Therefore, the adaption of the tokenizer is necessary and new training of the TreeTagger with annotated blog comments is required. Future work will further enhance POS tagging results for blog comments. Moreover, a gold standard for blog comments will be created.

ACKNOWLEDGMENTS. This work was partially supported by the Project House HumTec at RWTH Aachen University, Germany.

## REFERENCES


1. Schmid, H.: Improvements in Part-of-Speech Tagging with an Application to German. In: ACL SIGDAT-Workshop, pp. 47–50 (1995)
2. Giesbrecht, E., Evert, S.: Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In: 5th Web as Corpus Workshop (2009)
3. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: EKAW. Springer (2006)
4. Ehrig M., Maedche, A.: Ontology-focused crawling of web documents. In: Symposium on Applied Computing, pp. 1174–1178. ACM, New York (2009)
5. Zheng, H.-T., Kang, B.-Y., Kim, H.-G.: An ontology-based approach to learnable focused crawling. Information Sciences 178, 4512–4522 (2008)
6. Roman, S.: Eine Ontologie für die Grammatik. Modellierung und Einsatzgebiete domänenspezifischer Wissensstrukturen. In: KONVENS, pp. 125–129 (2006)
7. Fellbaum, C.: Wordnet. In: Poli, R., Healy, M., Karneas, A. (eds), pp. 231–243 (2010)
8. Missen, M., Boughanem, M., Cabanac. G.: Challenges for Sentence Level Opinion Detection in Blogs. In: 8th ICIS, pp. 347–351. IEEE Press (2009)
9. Missen, M.M., Boughanem, M.: Using WordNet's Semantic Relations for Opinion Detection in Blogs. In: ECIR, pp. 729–733. Springer-Verlag, Berlin, Heidelberg (2009)
10. Taboada, M., Brooke, J., Toloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics 37, pp. 267–307 (2011)
11. Sebastian, F.: Machine learning in automated text categorization. ACM Computing Surveys 34, 1–47 (2002)
12. Mishne, G., Glance, N.: Leave a Reply: An Analysis of Weblog Comments. In: 3rd Annual Workshop on the Weblogging Ecosystem at WWW (2006)
13. Bednarek, M.: Language patterns and ATTITUDE, Functions of Language 16, 165–192 (2009)
14. Bednarek, M.: Evaluation in Media Discourse: Analysis of a Newspaper Corpus. Continuum, London, New York (2006)
15. Alston, W.P.: Illocutionary acts and linguistic meaning. In: Tsohatzidis, S.L. (eds), pp. 29– 49 (1994)
16. Ruiz-Rube, C. M. Cornejo, J., Dodero, M., Garcia, V.: Development Issues on Linked Data Weblog Enrichment. In: 4th MTSR, pp. 235–246. Springer, Berlin, Heidelberg (2010)
17. Schiller, A., Teufel, S., Stöckert, C., Thielen, C.: Guidelines für das Tagging deutscher Textkorpora mit STTS. Technischer Bericht, Institut

für maschinelle Sprachverarbeitung, Universität Stuttgart und Seminar für Sprachwissenschaft, Universität Tübingen (1999)
18. Trevisan, B. Jakobs, E.-M.: Talking about mobile communication systems: Verbal comments in the web as a source for acceptance research in large-scale technologies. In: IPCC, pp. 93–100 (2010)

**MELANIE NEUNERDT**
INSTITUTE FOR THEORETICAL INFORMATION TECHNOLOGY,
RWTH AACHEN UNIVERSITY,
SOMMERFELDSTR. 24, 52074 AACHEN, GERMANY
E-MAIL: <NEUNERDT@TI.RWTH-AACHEN.DE>

**BIANKA TREVISAN**
INSTITUTE FOR THEORETICAL INFORMATION TECHNOLOGY,
RWTH AACHEN UNIVERSITY,
SOMMERFELDSTR. 24, 52074 AACHEN, GERMANY
E-MAIL: <B.TREVISAN @TK.RWTH-AACHEN.DE>

**RUDOLF MATHAR**
INSTITUTE FOR THEORETICAL INFORMATION TECHNOLOGY,
RWTH AACHEN UNIVERSITY,
SOMMERFELDSTR. 24, 52074 AACHEN, GERMANY
E-MAIL: <MATHAR@TI.RWTH-AACHEN.DE>

**EVA-MARIA JAKOBS**
INSTITUTE FOR THEORETICAL INFORMATION TECHNOLOGY,
RWTH AACHEN UNIVERSITY,
SOMMERFELDSTR. 24, 52074 AACHEN, GERMANY
E-MAIL: <E.M.JAKOBS@TK.RWTH-AACHEN.DE>