Mining corpora of computer-mediated communication: Analysis of linguistic features in Wikipedia talk pages using machine learning methods

Michael Beißwenger (TU Dortmund)
Harald Lüngen (IDS Mannheim)
Eliza Margaretha (IDS Mannheim)
Christian Pölitz (TU Dortmund)

Abstract

Machine learning methods offer a great potential to automatically investigate large amounts of data in the humanities. Our contribution to the workshop reports about ongoing work in the BMBF project KobRA (http://www.kobra.tu-dortmund.de) where we apply machine learning methods to the analysis of big corpora in language-focused research of computer-mediated communication (CMC). At the workshop, we will discuss first results from training a Support Vector Machine (SVM) for the classification of selected linguistic features in talk pages of the German Wikipedia corpus in DeReKo provided by the IDS Mannheim. We will investigate different representations of the data to integrate complex syntactic and semantic information for the SVM. The results shall foster both corpus-based research of CMC and the annotation of linguistic features in CMC corpora.[1]

1 Introduction

Up to now there have been very few annotated corpora of CMC freely available for the scientific community. Scholars doing data-based research of CMC discourse therefore often face the following limitations:

(a) They have to collect corpora for their research projects by themselves.

(b) "Off the shelf" tools for the linguistic annotation of written language data do not perform on CMC data in a satisfying way.

(c) Given (a) and (b), the researchers either have to annotate their corpora manually or confine themselves to analyzing their corpora as raw data (without the possibility to query linguistic annotations).

(d) The corpora they are able to analyze (taking into consideration that (a) and (c) are consuming a lot of their time and effort) are rather small than big.

The methods and experiments described in this paper are driven by the vision that the application of machine learning methods can improve the situation and possibilities of building corpora and doing corpus-based analysis of CMC discourse in several respects:

1. If we succeed to adapt machine learning methods for the automatization of typical routine tasks in corpus-based analysis (e.g. the cleaning and classification of query results), then these methods can support linguists in analyzing bigger data than they could analyze when every routine task would have to be done manually. "Big data", here, refers to amounts of data which are too large to be analyzed intellectually. For a linguist, the Wikipedia which is used as the test bed for the experiments reported here definitely *is* "big data": The German Wikipedia corpus in DeReKo comprises more than 1.5 million article pages (consisting of 678 million word tokens) and more than 555,000 talk pages (consisting of 264 million word tokens).

2. The methods applied can be used not only for mining the big data for those "gold nuggets" which are relevant for a particular linguistic research question; they may additionally be

used as a basis for automatically annotating the retrieval and classification results. In this respect, machine learning methods also enhance the conditions for building annotated CMC corpora.

In the following sections we give an overview of the project background of our work (sect. 2), a description of the Wikipedia corpus in DeReKo (sect. 3), and a description of the linguistic phenomena under observation (sect. 4). Sect. 5 describes the machine learning methods applied and sect. 6 gives an outlook on ongoing and future work.

2 Project background

The work presented in our paper is part of the KobRA project ("Corpus-based linguistic research and analysis using data mining") funded by the eHumanities program of the BMBF 2012-2015.[2] The project brings together researchers from linguistics, language technology and artificial intelligence to adapt machine learning methods for recurrent and time-consuming routine tasks that linguists have to perform when doing corpus-based linguistic analysis (e.g. classification and disambiguation of results from corpus queries) and thus to enable researchers to work with amounts of data that are too big to be be analyzed intellectually. The application scenario for the methods developed in the project is defined in case studies from several fields of linguistic research: diachronic linguistics, lexicography, variational linguistics/computer-mediated communication.

The data basis and test bed for the experiments reported in this paper is the German Wikipedia corpus in DeReKo provided by the IDS Mannheim (cf. sect. 3) on which the methods are trained and evaluated and which allows for a comparison of language use in monologic texts (= "article pages") and in dialogic written conversations (the sequences of user postings that can be found on "talk pages") which, cum grano salis, are both usually written by the same user group (= those users who contribute to writing Wikipedia articles as authors, moderators, reviewers etc.). Previous research has shown that Wikipedia is a fruitful resource for studies in linguistic variation on the internet (Storrer, 2013).

The scope of the experiments is on the retrieval and automatic classification of selected linguistic phenomena which can be considered as either specific for language use in written CMC or as elements which are typical of language use under the conditions of spontaneous, dialogic interaction and which occur both in spoken conversations as well as in written conversations on the internet (cf. sect. 4).

3 The corpus

The CMC corpus we used for the experiments is the 2013 conversion of the Wikipedia available within DeReKo, the German Reference Corpus (Kupietz and Lüngen, 2014), at the *Institut für Deutsche Sprache* in Mannheim.[3] It was built from the Wikipedia dump of July 27, 2013, and contains approx. 943 million tokens. Unlike other corpora derived from Wikipedia, it has been prepared as a linguistic corpus and comprises the whole German Wikipedia. It is represented in I5 (Lüngen and Sperberg-McQueen, 2012) the TEI P5 customization used to encode the texts in DeReKo.

Since the Wikipedia talk pages corpus was one of the first sub-corpora in DeReKo to contain CMC texts, the I5 format was on this occasion extended to incorporate macro-structural elements (most notably <posting>) and attributes to represent the thread and posting structure of CMC data as proposed in (Beißwenger et al., 2012).

In Wikipedia, each talk page (or: discussion) is paired with a Wikipedia article. On a talk page, the users, i.e. Wikipedia authors, can discuss an article, i.e. whether and how it should be revised or extended, what references or images to include etc. When an article is edited, the editor usually justifies his/her edit by a written contribution on the respective talk page. According to the Wikipedia talk page guidelines[4] and also in prac-

[2] See http://www.kobra.tu-dortmund.de. The project is headed by Angelika Storrer (U Mannheim/German Linguistics. The main partners of the project are Katharina Morik (TU Dortmund University/Artificial Intelligence), the IDS Mannheim (Marc Kupietz, Andreas Witt), the BBAW Berlin (Alexander Geyken) and the SfS at U Tübingen (Erhard Hinrichs/Computational Linguistics).

[3] see http://www.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html

[4] http://de.wikipedia.org/wiki/Wikipedia:Diskussionsseiten

tice, a talk page is structured much like a discussion forum, i.e. it comprises a sequence of discussion topics introduced by headings, and within such a topic, dialogue turn(Schegloff, 2007)-like units provided by a single user are delimited by means of paragraph indentation, thus forming a discussion thread. (Beißwenger et al., 2012) classify these turn-like units as *posting* units, and this view has also been adopted in the I5 representation of the Wikipedia corpus in DeReKo.[5]

The conversion of the wikitext data of the Wikipedia dump into the I5 format is described in detail in (Margaretha and Lüngen, In press), the source code of the conversion tools is available from GitHub.[6] The conversion pipeline also includes a heuristic method for identifying the posting segments in a talk page and an evaluation of this method. According to the evaluation on 49 talk pages, the performance of the automatic heuristic posting segmentation yielded approximately 60% micro average precision and 80% micro average recall when compared with posting segmentations provided by human annotators. The agreement between the two human annotators themselves was κ=0.76, which suggests that the exact identification of posting boundaries is not an unambiguous tasks for humans, either, when reading a talk page. Altogether 5.4 million posting segments were identified and annotated in the talk pages corpus by the automatic segmentation. For the corpus, PoS annotations from the Stuttgart TreeTagger are also available (though they have not been used in the experiments described here), and we have prepared Wikipedia corpora in the same fashion for other languages, too.

[5]A posting in CMC is originally defined as a piece of text sent to the server by the author at one specific point in time. Hence, the turn-like sections in Wikipedia talk pages are strictly speaking not postings, as a wiki user always posts a new version of the whole wiki page, i.e. (s)he might have edited the page in different places, even might have modified or deleted previous contributions by other users. But since on a talk page, the dialogue structure with its sequentially ordered threads of turns prevails, the turn-like units have been identified with postings as defined in (Beißwenger et al., 2012) in the present I5 representation.

[6]https://github.com/IDS-Mannheim/Wikipedia-Corpus-Converter

4 Machine learning tasks

For our first experiments with adapting machine learning methods for the analysis and annotation of Wikipedia, we selected two types of linguistic features which are of interest for studies in language-focused CMC research as well as for research on linguistic variation in written and spoken language.

4.1 Interaction words

Interaction words are units which are based on a word or a phrase of a given language describing expressions, gestures, bodily actions, or virtual events. In German CMC, simple forms of interaction words typically have the form of non-inflected verb stems (*grins, lach, freu*) whereas complex forms additionally may incorporate objects and/or adverbials (*lautlach*, *diabolischgrins, kopfschüttel, schulterzuck, nachlinksrutsch*). Some interaction words have the form of acronyms (*lol*, *rofl*, *g*). Interaction words are usually not part of the syntactic structure of the utterance they accompany; instead, they are used for the description of emotions or mental activity, as illocution or irony markers, or to playfully mimic bodily activity (Beißwenger et al., 2012). They are often (but not necessarily) enclosed in asterisks (**grins**, **freu**).

As a starting point for our experiments in automatically detecting interaction words, we assume that a researcher who wants to analyze interaction words in a corpus where these units are not explicitly annotated would usually define a query pattern for expressions which s/he considers as typical forms of interaction words – for example forms which are frequently used as interaction words in other corpora or random expressions between asterisks. We defined tasks for both of these two scenarios:

Task #1a:

- *Data basis:* Query results for the most frequent forms of interaction words according to the annotations in the Dortmund Chat Corpus (*lol, lach, freu, grins, wink, seufz*; cf. (Storrer, 2013). Each match is represented in a snippet with a context size of max. 999 characters (extracted from the corpus).

- *Training and evaluation data:* Random sample with 600 matches from the data basis that have been independently classified by two human annotators as "contains an interaction word" (type 1) or "does not contain an interaction word" (type 0).
- *Task:* Learn a classification model for separating the snippets into type 1 and type 0 snippets.

Task #1b:

- *Data basis:* Query results for expressions between asterisks. Each match is represented in a snippet with a context size of max. 999 characters (extracted from the corpus).
- *Training and evaluation data:* Random sample with 600 matches from the data basis that have been independently classified by two human annotators as "contains an interaction word" (type 1) or "does not contain an interaction word" (type 0).
- *Task:* Learn a classification model for separating the snippets into type 1 and type 0 snippets.

4.2 "Non-canonical" uses of *weil* and *obwohl*

In the written German standard, *weil* and *obwohl* are conjunctions which introduce subordinate clauses with the finite verb form in sentence-final position. Under conditions of conceptual orality (prototypically but not limited to spontaneously spoken language), *weil* and *obwohl* also occur in the pre-front position of sentences with the finite verb in a position other than sentence-final (typically V2; examples: *"ja toll aber so richtig steht es nicht drin* ***weil*** *damals sollten wir nämlich eine arbeit in informatik machen über das dualsystem", "Ja ich bin auch 96 Fan aber trotzdem, er hätte auch im Spiel sein fehler noch ändern können.* ***Weil*** *ich bin selber Schiedsrichter, und hatte auch schon so eine Situation"*). In popular discussions about language change, cases like these are often considered as degenerated grammar and as an example of language decline (cf. critically on this discussion: (Günthner, 2008) while analysis in the field of spoken language research/interactional linguistics could show that in their "non-canonical" uses *weil* and *obwohl* often have functions which are different from those of the "canonical" use as subordinate conjunctions (cf. e.g. (Gohl and Günthner, 1999), (Günthner and Auer, 2005), (Imo, 2012). It is an open question inhowfar "non-canonical" uses of *weil* and *obwohl* in written CMC have the same or similar functions as "non-canonical" uses in spoken language. Corpus-based analyses on this question will help to develop a better understanding of how much the encoding medium (writing vs. articulated sound) and the structure of the encoding process (private composition before transmission vs. 'on-line') affect the structure of utterances in written and spoken conversations.[7]

Our first experiments addressed the classification of matches for *weil* in the corpus:

Task #2:

- *Data basis:* All 305,708 matches for *weil* in the talk pages subcorpus. Each match is represented in a snippet with a context size of max. 999 characters (extracted from the corpus).
- *Training and evaluation data:* Random sample with 1,200 matches from the data basis that have been independently classified by two human annotators as "non-canonical use" (type 1) or as "canonical use" (type 0).
- *Task:* Learn a classification model for separating the snippets into type 1 and type 0 snippets.

5 Machine learning methods

Machine learning methods offer automatic classification and filter methods for large scale data. Based on examples, a decision function is extracted that can be applied to large amounts of data to classify and filter them with respect to the CMC phenomena like those described in section 4. The collection of all these extracted rules is summarized by a single classification model. The derivation of such rules depends on the features of the

[7] Cf. the discussion of the effect of written 'en bloc' encoding on the process of message composition and the system of turn-taking in (Beißwenger, 2007)

data as well as on the complexity and regularities in the texts.

We use kernel methods (Shawe-Taylor and Cristianini, 2004) and Support Vector Machines to integrate different feature representations of the corpus snippets into a classification model. A Kernel encodes similarity information for pairs of snippets based on a certain feature representation. Kernel methods enable us to directly integrate all possible feature representations of the data – even complex representations such as syntactic structures or semantic relations – into a single classification model. This model is a Support Vector Machine that uses the Kernels to decide which snippets belong to a certain class and which not.

We use three different kernels to represent the snippets from the Wikipedia corpus: A *tree kernel* is used to integrate syntactic information from parse trees as proposed by (Moschitti, 2006). To derive the parse trees for German sentences, we use the Stanford Parser (Rafferty and Manning, 2008). Further information is integrated via *Substring kernels* that count the presence of certain substrings in a given text (Lodhi et al., 2002). Last, a linear kernel is used on the *bag-of-words* representations of the corpus snippets. In the bag-of-words representation, each snippet is represented via a large vector. Each component of such a vector gives the (normalized) frequency of a certain word appearing in the text. This is the baseline approach which we compare to the kernel methods.

In order to use the kernels for the classification of the phenomena under observation, we generate a Gram matrix for each of them. The Gram matrix contains the kernel evaluations for each pair of snippets from the training data. These evaluations are everything needed to learn our classification model.

For each Gram matrix, we train a Support Vector Machine using the LibSVM library (Chang and Lin, 2011). The Support Vector Machine uses the Gram matrix to learn a decision function that is used to classify any snippet for the respective phenomena. For both the training of the classification model and its application on test data, we only use kernel evaluations from the Gram matrix.

The training is done on a part of the hand-classified training data described in section 4. Then we apply the Support Vector Machine to the rest of the data to classify them for the phenomenon. Based on this independent test set, the performance of the classifier can be evaluated and we can estimate which kernel is best suited for the task.

In order to estimate the performance, we perform a 10-fold cross validation evaluation. The measure of the performance is the F1 score, that is the mean of the precision and the recall of the trained classifier. Finally, the model is applied to the unlabeled test data. In order to get information on what snippets are difficult to classify, we additionally estimate confidence values of the classification. These values are used to propose additional hand classifications for some of the snippets. In an *Active Learning* (Settles, 2009) setting, this potentially results in better training data by actively choosing which snippets to classify by hand.

6 State of work and future agenda

At the KONVENS workshop, we will present and discuss first results from adapting the machine learning methods outlined in sect. 5 for the retrieval and disambiguation tasks described in sect. 4. As next steps, we are planning to further improve these results by using additional methods (Active Sampling), by doing experiments with different data sets for the same phenomena and by adapting the models which perform well also to data sets from other CMC genres/corpora.

The optimized classification models shall finally be used for automatically annotating the results in the corpus data. For this purpose, we will use labels from the extended STTS tagset for the POS tagging of CMC corpora ("STTS-IBK") that has been defined for the Empirikom shared task on linguistic processing of German CMC (*EmpiriST2015*[8]).

As a part of our future agenda, we are planning to transfer the machine learning methods described in this paper also to other genres and phenomena: On the one hand, the classifiers trained on Wikipedia talk pages shall be evaluated with/adapted to data also from Wikipedia articles pages and from other CMC genres such as chats, tweets, or blog comments. On the other

[8] http://empirikom.net/bin/view/Themen/SharedTask

hand, the methods developed for the identification/classification of interaction words and "non-canonical" *weil/obwohl* shall be adapted also to other linguistic phenomena which are of interest for language-focused corpus investigations of CMC discourse. In this context, we will also investigate which approaches for text representations in the field of machine learning are important to safely apply our trained models to new and unseen texts and phenomena, and examine and compare our methods to previous domain adaptation methods like FLORS (Schnabel and Schuetze, 2014).

References

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012. A TEI schema for the representation of computer-mediated communication. *Journal of the Text Encoding Initiative*, 3.

Michael Beißwenger. 2007. *Sprachhandlungskoordination in der Chat-Kommunikation*, volume 26 of *Linguistik – Impulse & Tendenzen*. de Gruyter, Berlin. New York.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. Technical report, ACM Transactions on Intelligent Systems and Technology. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm .

Christine Gohl and Susanne Günthner. 1999. Grammatikalisierung von *weil* als Diskursmarker in der gesprochenen Sprache. *Zeitschrift für Sprachwissenschaft*, 18(12):39–75.

Susanne Günthner and Peter Auer. 2005. Die entstehung von diskursmarkern im deutschen – ein fall von grammatikalisierung? In Torsten Leuschner, Tanja Mortelsmans, and Sarah de Groodt, editors, *Grammatikalisierung im Deutschen*, pages 335–362. de Gruyter, Berlin.

Susanne Günthner. 2008. Geht die nebensatzstellung im deutschen verloren? In Markus Denkler, Susanne Günthner, Wolfgang Imo, Jürgen Macha, Dorothee Meer, Benjamin Stoltenburg, and Elvira Topalovicet, editors, *Frischwärts und unkaputtbar. Sprachverfall oder Sprachwandel im Deutschen*, pages 103–128. Aschendorff, Münster.

Wolfgang Imo. 2012. Wortart diskursmarker? In Björn Rothstein, editor, *Nicht-flektierende Wortarten*, pages 48–88. de Gruyter, Berlin.

Marc Kupietz and Harald Lüngen. 2014. Recent developments in dereko. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, March.

Harald Lüngen and Michael Sperberg-McQueen. 2012. A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative*, 3:1–18.

Eliza Margaretha and Harald Lüngen. In press. Building linguistic corpora from wikipedia articles and discussions. *Journal for Language Technology and Computational Linguistics (JLCL)*.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In Diana McCarthy and Shuly Wintner, editors, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), April 3-7, 2006, Trento, Italy*, pages 113–120. The Association for Computer Linguistics.

Anna Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the ACL Workshop on Parsing German*.

Emanuel Schegloff. 2007. *Sequence Organization in Interaction*, volume 1: A Primer in Conversation Analysis. Cambridge University Press, UK.

Tobias Schnabel and Hinrich Schuetze. 2014. Flors: Fast and simple domain adaptation for part-of-speech tagging. In *Transactions of Association for Computer Linguistics*, pages 15–26.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison. Computer Sciences Technical Report.

John Shawe-Taylor and Neollo Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Angelika Storrer. 2013. Sprachstil und Sprachvariation in sozialen Netzwerken. In Barbara Frank-Job, Alexander Mehler, and Tilmann Sutter, editors, *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*, pages 331–366. VS Verlag für Sozialwissenschaften, Wiesbaden.