Universität Potsdam, Karl-Liebknechtstr. 24-25, 14476 Potsdam/OT Golm

Humanwissenschaftliche Fakultät
Frau Lüßen-Koch

Humanwissenschaftliche Fakultät
Department Linguistik

Angewandte Computerlinguistik

Prof. Dr. Manfred Stede

Telefon: +49-331-977-2691
Telefax: +49-331-977-2087
Sekretariat: +49-331-977-2950

Datum: 19.6.2019

Review of the dissertation of Uladzimir Sidarenka

In his thesis, Uladzimir Sidarenka (henceforth: US) is concerned with applying sentiment analysis (henceforth: SA) to the German language, and in particular to the genre of Twitter messages. The running theme of his work is to systematically test how approaches that have previously been developed for English perform in comparison to each other when applied to German tweets, and to suggest his own modifications and extensions in various places. SA is a fairly broad field, and the thesis addresses an impressively wide range of subtopics: corpus construction, message-level polarity detection, fine-grained SA, and measuring the influence of discourse-level features.

The dissertation consists of six chapters, an introductory foreword, and an afterword, which here plays the role of the "Conclusions" chapter. The reader notices the absence of a "Related Work" chapter, but this is due to US folding this into the separate chapters (giving it different levels of attention, though, as pointed out below), which seems a good approach given the overall breadth of topics.

After succinctly summarizing the main research questions in the foreword, **Chapter 1** provides an introduction to sentiment analysis that is largely organized chronologically, points to the central milestones achieved in the field, and pays specific attention to work on microblogs. In doing this, it goes some way to providing an account of the state of the art, but one thing the reader misses is a concise definition of SA; here it would have been sufficient to quote one from influential literature, such as the book by Liu (2012). Also, giving a few more examples to illustrate the range of subtasks and possible domains would be helpful.

The first central contribution of this thesis is the construction of an annotated corpus of German tweets. Until recently this was the only one of its kind altogether, and it still is the only one with fine-grained annotation of sentiment. **Chapter 2** explains the method that US devised to collect data for a corpus that covers several domains and at the same time achieves a certain balance in terms of sentiment-bearing ("polar") and neutral Tweets. This thoughtful method might well serve as a blueprint for similar work and to my mind in itself constitutes a useful contribution. US proceeds to explain his annotation scheme, which draws on the early work of Wiebe and colleagues, and distinguishes polar words, sources and targets of opinions. Polar words get a two-valued *strength* attribute, where I wonder why the two values are *strong* and *medium,* rather than *weak*. In line with the careful procedure of corpus construction, the annotation process is a three-stage process, with an increasing role of discussion and adjudication. US measures inter-annotator agreement at each stage

and shows that it successively improves to a degree that can be considered good for this kind of project. For measuring IAA, the author introduces his proposal of "binary" and "proportional" kappa as a lean and strict measure for dealing with the problems of overlapping spans and fuzzy span boundaries; here a slightly broader discussion, which looks at IAA treatment in related work on span labeling, and maybe specifically considers the potential utility of Krippendorf's unitized alpha, would have been nice.

Finally, a breakdown of annotation volume for the various sections of the corpus is given, in order to verify that the initial balance goals have been achieved. US also shows the different IAA values for the sections and checks for correlations with topic and formal categories.

The net result of corpus construction is a set of two annotations; US does not build a single gold standard – a decision that could have been briefly discussed at the end of the chapter. Nonetheless, altogether the reader is provided with a thoughtful design of the data collection and annotation procedures, which are carefully evaluated, so that a solid basis for automatic experiments is now in place.

Chapter 3 addresses the first subtask, viz. the construction of sentiment lexicons for German. The author chooses 400 tweets from annotator 1 as development set, and tests the generated word lists on all 6000 polar terms (with emoticons ignored) that were labeled by annotator 2. In explaining his evaluation metric, he criticizes previous approaches for "usually" measuring the size of the intersection of a candidate lexicon with the well-established General Inquirer lexicon (an idea that, as US shows, has some problems). Providing some sample references to back up the criticism would be useful here.

In his first experiment, US evaluates three existing German lexicons (produced mostly by translating English resources) individually, as well as the performance of their intersection and their union. It turns out that the intersection performs best, both on micro- and macro-F1 (which US routinely determines for all experiments throughout the thesis).

The question then is to what extent methods for automatically building lists of polar words can compete. The author first turns to *lexicon-based* methods, which work by propagating the values of seed terms through WordNet, by different technical methods. US translates the standardly-used 14 English seed adjectives of Turney/Litman (2002) into German and adds neutral terms, so the result is a 10+10+10 seed list. He then reimplements six approaches from the literature and runs them on GermaNet, optimizing the number of iteration steps for each method. It turns out that the approaches perform considerably worse than the existing lexicons do, and that the different methods have individual strengths and weaknesses. US does not compare the performance to the original results the methods achieved on the English WordNet; while for absolute values that would not be very informative (he extended the task to 3-way classification, and WordNet has better coverage than GermaNet), but the relative ordering could be compared for the English and German scenarios.

Corpus-based methods induce word lists from corpora by measuring distributional similarity to given seed terms in various ways. US reimplements four methods, using the German Twitter month snapshot (Scheffler 2014) for estimating the word similarities. Here, the experiments run into a problem due to ambiguity in one of the translated words; so the result of one clear winner and a group of runners-up is to be taken with a grain of salt.

Finally, US turns to *neural word embedding* methods, where he reimplemented two successful approaches from the literature, again extended to 3-way classification. In addition, he tests four new methods based on nearest-centroids, k-NN clustering, PCA, and an algorithm for computing a linear projection (LP) in vector space. All are being tested with word2vec embeddings trained on the Twitter month snapshot. It turns out that US's LP method performs best not only among the new approaches, but also outperforms the other corpus- and lexicon-based methods, and comes very close to the best "previous lexicon" method (i.e., the intersection of the three lists).

In a follow-up experiment, US tests the effects of working with different word embeddings, to explore the relation between the two goals of (generally) predicting context words and (task-specifically) classifying tweet polarity. The results were very different for different approaches, with no clear "winner" for the 3-way classification problem. An experiment on the effects of vector normalization yields similar results. Still, it is important that US does not simply use methods out of the box, but is interested in dissecting their parts and measuring their contributions. In the same vein he then examines the influence of choosing different seed sets, as they have been used in the literature (but so far, to my knowledge, they have not been compared to each other). For the lexicon-based methods, this reveals differences between methods (degree of performance dependence on seeds) and between seed sets (how well do they perform across methods). For the corpus-based methods, on the other hand, such differences were considerably smaller, which US attributes to ambiguity arising from translating the English seed words. For the NWE methods, WS finds a new overall state of the art result, using his own LP approach with the seed set of Kim/Hovy (2004).

Finally, US takes a look at the output polarity word lists generated by the various methods. The top-10 polar terms of dictionary methods are intuitively OK, whereas those of corpus and NWE methods (with two exceptions, one of them being his LP method) feature many rare and largely neutral terms.

The reason why **Chapter 4** is called “Aspect-based sentiment analysis” is a bit unclear, since this term is usually reserved for approaches where opinions are analyzed not for the target as a whole but for various relevant *attributes*. However, other places in the chapter use the more suitable term "fine-grained" SA (and I recommend to use that also in the title). In any event, the task is to identify the overall text span of an opinion as well as its source and target expressions, and the polar words responsible for the opinion.

To evaluate the work, US proposes to use a token-sensitive measure suggested in related work for other purposes. For appreciating this decision, it would be good to get information on how fine-grained SA approaches for English usually handle this. Likewise, for methods and results a brief overview of related work would here be helpful.

The data description section does not say which annotation is now used as "gold". (Probably the same setup is used as in Chapter 3?) For replicability, this would be important to know.

Two technical approaches are being experiment with: a feature-rich CRF model, and two NN models. US first gives a thorough introduction to the CRF approach as such (which becomes relevant later, when he proposes alternative variants of the CRF topology). But there is little information on whether/how previous SA research has employed them.

The performance analysis shows a huge drop when going from training test to test set. For analysing this, WS runs ablation tests for the feature groups and finds that specifically the recognition of source and target strongly depend on selected feature groups. He also inspects the top-10 state and transition features (some of which are domain-specific lexemes), and discusses some prominent errors observed on the development set.

The two RNN architectures, an LSTM and a GRU, are run on a balanced variant of the dataset (obtained by upsampling). While the NNs suffer less from overfitting, their performance on the test set is somewhat lower than that of the CRF model. Next, similar to the analysis in Chapter 3, the potential utility of two types of word embeddings is being tested: Twitter-pretrained word2vec and least-square embeddings that provide a fallback for unknown words. The latter, in conjunction with the LSTM architecture, are now able to beat the CRF, though not by a wide margin.

The final section of the chapter describes experiments on determining potential influential factors for the performance of the various approaches. First, the author turns to the annotation scheme and considers the variant of labeling only polar words as "sentiment" (as opposed to the whole constituent, which might very well cause confusion for the classifiers due to many neutral words in these spans).

This reduction leads to a huge performance increase on sentiment, but unfortunately lowers the results for source and target. (Still, the macro-f1 is considerably better for the narrow variant.) US surmises that the introduced token gaps between sentiment and source/target make the identification of the latter more difficult.

Since the interplay between contextual and token-based features seems difficult to capture for the models, US proposes (and implements) higher-order CRFs, first- and higher-order semi-Markov models, and tree-structured CRFs as alternatives to the linear-chain first order CRF. These are evaluated on the training and on the development set (rather than on the test set, as done in the first CRF and NN experiment), the reason being unclear. The original CRFs can indeed be beaten, but different variants are successful for the three tasks, so there is no clear "winner". All in all, the original CRF seems to be an effective overall solution; likewise, testing more complex variants of the NN architectures did not yield improvements.

A final experiment tests whether running text normalization of the tweets was a good idea for fine-grained SA, and finds that indeed it was.

A generally more popular SA task in the community is to compute message-level polarity, which is the topic of **Chapter 5**. As a prerequisite, the corpus annotations need to be mapped to tweet level, for which US describes a plausible procedure.

In addition, all methods will be evaluated on 7500 tweets from a newly introduced German Twitter corpus by other researchers (SB10k), which is equipped with message-level polarity.

Similar to previous chapters, US first discusses lexicon-based methods, then traditional machine learning, and finally neural networks. For all of these, the related work is aptly summarized, and own ideas are added. His reimplementation of five *lexicon-based* methods uses the Zurich polarity list, and a harmonization of the output for making them comparable. It turns out the oldest approach by Hu/Liu (2004) performs best. In interesting experiments, US then checks the influence of common contextual polarity features: negation, intensifiers, and irrealis markers. Modulo a few specific exceptions, the bottom line is that – interestingly – switching off this type of context analysis improves the results. A sample error analysis reveals quite different problem sources for the systems, one of them being rules relying on (standard) English syntax that do not work well on the German Tweets. Further scrutiny could try to figure out whether the detrimental influence of context features is due to preprocessing errors, or to genuinely different contextual constellations found in Twitter messages.

In the section on *machine-learning* methods, I would appreciate a sub/section break between the extensive related work part and the author's own proposal and implementation. US reimplemented three systems that were successful on English, and ran ablation of feature groups to study their impact on performance. By and large, all features are helpful, except POS tags, which WS attributes to the problem of using a standard newspaper-trained tagger for German, as opposed to Twitter-specific ones in the English work. Similarly, WS offers explanations for some other slightly peculiar results, and he also determines the top-10 most important features for each classifier.

In the posthoc experiments, US replaced the SVM classifier with Naive Bayes and Logistic Regression. At the time when the work on English was published, SVMs were considered as the most successful ML approach, but US finds that on his data, LR performs better for two of the three systems. However, he is careful to note that such results always depend on the data sets, and one should not optimistically expect to find a similar improvement on the English SemEval datasets.

For *deep learning* methods, WS reimplements six approaches and furthermore proposes a modification to the successful approach of Baziotis et al (2017) by adding two more attention mechanisms that favor the presence of polar words and that of valence shifters. His evaluation demonstrates that these ideas indeed compete with the state of the art, yielding the best performance for some classes in the two corpora studied. Overall, DL lags behind the traditional ML, though. The next experiment thus adds different variants of word embeddings, with one using a least-square

method to handle lexical gaps. This step again boosts performance and in particular, his LBA method now yields the best macro-F1 results of *all* approaches (from all families) on the PoTS corpus by a margin of 0.06.

For the error analysis section, US uses the LIME tool, which allows for identifying the tokens that were most responsible for the decisions of the DL system. While this does not yield quite as much transparency as a feature impact analysis of traditional ML, it goes some way toward solving the "explainability" problem of NNs.

Of the final "parameter variation" experiments, I mention here only one that adds large amounts of training data by distant supervision: 4 million tweets from the Twitter corpus of Scheffler (2014), which could be (heuristically) identified as pos/neg/neut on the basis of emoticon presence (an approach whose validity is not beyond doubt, but has been routinely used for similar purposes). Interestingly, it turns out that the ML and DL approaches perform worse when adding that data; US's hypothesis is that the big difference in class distribution between the data sets is responsible. It would be nice to test this (by downsampling the noisy data appropriately), even if just for one approach (the author mention the enormous training time required for the methods). The main hope of this step was to overcome the DL systems' tendency of falling back to just predicting the majority class, which US had observed for various settings. This hope was unfulfilled, though.

Chapter 6 examines whether message-level SA can benefit from discourse-level information. To this end, the PoTS and SB10k corpora are run through an automatic discourse segmenter (to which US had also significantly contributed, outside of this thesis); this lead to roughly 8000 tweets with >1 segment. For each segment, polarity was assigned using the author's LBA classifier.

After providing a thorough review of related research, the author argues that RST is the most suitable approach to capturing discourse information for his purposes and thus runs a publicly-available RST parser (re-trained on the German PCC data) on the corpus. As often done for SA work, the set of discourse relations is being reduced to just two: contrast and no-contrast. US notes that the performance of the parser is significantly lower than what is reported for English (where much more training data is available). In the spirit of the preceding chapters, the author then reimplements three earlier systems and develops some variants himself: two CRF models and a Recursive Dirichlet process. For the latent CRF models, the RST tree is converted to a dependency structure where each EDU (and the whole tweet as root node) is represented by the three polarity scores assigned by LBA. US then explains his mechanism for dealing with unobserved labels, and then turns to the RDP implementation, whose introduction is rather brief and would benefit from providing some more detail.

Comparing the performances, the results are rather mixed: On the one hand, the author's own implementations comfortably beat two simple baselines (LAST and ROOT), as well as the three re-implementations. On the other hand, the advantage over a classifier that completely *ignores* the discourse information is unfortunately rather small (0.006 for the PoTS corpus, 0.002 for SB10k).

As in earlier chapters, WS then traces output errors made by his implementations as well as the baselines, and runs additional experiments with two alternative base classifiers. On PoTS, his own LBA is by and large the winner, but on SB10k, the system of Mohammad et al. (2013) performs better.

Given the overall modest success of adding discourse information, the author runs a follow-up experiment for testing the influence of automatic RST parsing. To this end, manual RST annotations were created by an assistant for 88% of the PoTS corpus. While the classifier results indeed increase in this scenario, the no-discourse system is – still – on a par with the best discourse approach. US also experiments with different relation sets that are drawn from the literature. Interestingly, the binary decision originally used (contrast) is best for span and relation, but not for nuclearity, where the approach using the full RST set is a clear winner. A follow-up result is that some of the discourse-SA approaches in fact perform better when larger relation sets are used.

Generally, the idea of combining automatic SA and automatic RST parsing – two error-prone tasks – on Twitter data is somewhat courageous. Still, the different results obtained in this chapter invite follow-up work that tries to more reliably identify the role of different discourse segments (a notion that is per se a bit problematic on Twitter) and/or tests the approach on similar but somewhat more formal short texts such as brief product reviews; for Twitter, the added value even of gold RST annotations seems clearly limited.

In an **Afterword**, the author gives a brief summary of the work, and then provides a list of "lessons learnt" with respect to the original research questions. In addition, yet another experiment is presented, which fills a gap in the previous comparative analyses: determine the relative contribution of the three steps of preprocessing, here applied to message-level SA using the LBA system.

Overall Evaluation: With some 160 pages, this dissertation is not unusually long. However, behind the scenes of the many result tables and their brief explanations, there is a huge amount of work involving implementing many approaches from previous work, coming up with ideas for extensions, and conducting a large number of follow-up experiments that test the relative impact of specific design decisions or parameters. As it happens with machine learning, the results are often somewhat inconclusive (why does system X work better with lexicon Y, etc.); but on the whole, Uladzimir Sidarenka is able to derive a range of interesting conclusions from his impressive body of work. These results pertain both to a "roadmap" for building SA systems for a new language, and to technical innovation for solving certain tasks; I mention here specifically his linear projection approach to lexicon construction and his lexicon-based attention mechanism for message-level SA). Furthermore, it is interesting to see that for the tasks considered here, deep learning approaches do not automatically lead to better results than those of well-established SVMs or CRFs; though Uladzimir may be right in insinuating his prognosis that eventually, DL architectures will be winning also on tasks like those tackled in this thesis.

In short, this thesis is impressive both for the thoroughness of comparisons and tireless tracing of influencing parameters, and for the creativity in designing new solutions for certain subtasks. My occasional critical remarks largely pertained (i) to a lack of connection to previous research in some places (while in others, that job is done very well). (ii) Error analyses are generally done but their scope is not made clear: While illustrative examples of errors are provided throughout, it is not clearly stated how many instances have been examined, and thus a statement on an error type being "rare" is hard to interpret. (iii) Overall, the thesis is well-written and remarkably free of typos or other calamities. The presentation style, however, could be improved by providing clearer borderlines between discussion of earlier research and own contributions, including a more explicit textual marking of the latter.

Altogether, I suggest only a very small list of recommended changes below, and am happy to grade the work with **magna cum laude**.

Suggested changes:
- Consider renaming chapter 4 to "Fine-grained sentiment analysis"
- in chapters 4-6 (or once for the whole thesis), clarify which annotation of the corpus has been used as gold (in the interest of replicability)
- p.94 3rd paragraph, spurious "enough"
- introduce sub/section break between related work and own approaches, e.g. in Sct 6.3

Prof. Dr. Manfred Stede