Proceedings of the 54th Hawaii International Conference on System Sciences | 2021

Machine Learning for Readability Assessment and Text Simplification in Crisis Communication: A Systematic Review

Hieronymus Hansen
University of Münster (ERCIS)
hieronymus.hansen@uni-muenster.de

Adam Widera
University of Münster (ERCIS)
adam.widera@wi.uni-muenster.de

Johannes Ponge
University of Münster (ERCIS)
johannes.ponge@wi.uni-muenster.de

Bernd Hellingrath
University of Münster (ERCIS)
bernd.hellingrath@wi.uni-muenster.de

Abstract

In times of social media, crisis managers can interact with the citizens in a variety of ways. Since machine learning has already been used to classify messages from the population, the question is, whether such technologies can play a role in the creation of messages from crisis managers to the population. This paper focuses on an explorative research revolving around selected machine learning solutions for crisis communication. We present systematic literature reviews of readability assessment and text simplification. Our research suggests that readability assessment has the potential for an effective use in crisis communication, but there is a lack of sufficient training data. This also applies to text simplification, where an exact assessment is only partly possible due to unreliable or non-existent training data and validation measures.

1. Introduction

Successful Crisis Communication (CC), be it in the wake of natural hazards, terrorist attacks or other comparable critical emergency situations, requires a rapid exchange of critical information between all actors involved in the crisis to respond accurately and timely in the given situation [25]. The aim is always to ensure the highest possible protection of the affected population [18, 52]. A prerequisite is that there is no confusion in the CC dialogue [35]. Researchers found that the process of cognitive message processing has so far played a subordinate role in CC [4, 49]. In the context of warning messages explicit reference was made to the lack of knowledge regarding the optimal message length, design and content [4, 62]. Since machine learning (ML) techniques for processing messages are considered an established tool in research and practice [e.g. in 44, 78], the question is, whether such technologies can also play a key role in CC in order to effectively communicate with the public. In this paper machine learning refers to ability of artificial intelligence systems "to acquire their own knowledge, by extracting patterns from raw data" [19] Our central research question is: *Which functions of ML-driven readability assessment and text simplification can be applied to support crisis communication?*

The required information varies from very generic (such as key facts about the event), to very specific questions (such as local availability of water pumps to dry basements). Besides the content perspective, the requirements for successful CC can also vary depending on the phase of the crisis management lifecycle. Warnings inform about upcoming short- and long-term threats and can contain behavioral instructions to minimize harm. Thus, warnings are useful not only during the preparation but also during the actual response phase. Though, requirements for CC differ in terms of urgency and target audience. Initial responses in the Covid-19 crisis included information about the origin of the virus and measures to be taken by the population to reduce the spread of the virus. Even nine months after the occurrence of SARS-CoV-2, reminders from governmental agencies to comply with existing hygiene regulations are prominent in public discourse [11]. Thus, drifts from early-warnings to educational CC messages can be observed when entering the recovery and rehabilitation phase. Last but not least, CC during the mitigation phase can have a fundamental impact on increasing risk-awareness on community level (see e.g. [47]).

Several generic characteristics or requirements for successful CC have been discussed in past works [4, 27, 34, 49, 62, 72]. Strengthening confidence in the sender of the message, and the willingness to cooperate are considered as overall objectives [7, 28]. Further, the messages should be sent at the right moment depending on the circumstances of the current crisis situation [34, 72]. Both, the source and the

URI: https://hdl.handle.net/10125/70890
978-0-9981331-4-0

content should appear credible to the recipient, correspond to reality and be free from contradictions [6, 7, 28]. The messages should be comprehensive without omitting key information [34]. The applied language should be as clear and simple as possible, without jargon, and understandable by anyone, including readers with language skills between the sixth and eighth grade [27, 40, 70]. In the following chapter, we present the applied methodology. Chapter 3 and 4 portray the results of these exploratory literature reviews on readability assessment (RA) and text simplification (TS). The findings are discussed in chapter 5. Chapter 6 concludes and mentions limitations of the findings.

2. Research Methods and Related Work

Our work is built upon a preceding systematic literature review on the requirements of effective crisis messages from crisis managers to the population in text form. It is based on the guidelines of Templier for "conducting rigorous IS literature reviews." [64]. We assigned the final requirements for crisis messages to three different categories. The first requirement category dealt with the *linguistic understanding* of the message. There are two requirements of this category relevant for this article: On the one hand the *comprehensibility* of the text through simple language [34]; on the other hand, the *completeness* of the message without losses of information relevant to the receiver. *Message framing,* the second category, deals with the impact of the words chosen on the readers' attitude. Lastly, the *components and content order* in the context of warning messages defined the last main category. A summary of the research process is given in Figure 1.

Our three main requirement categories of this review served as the foundation to identify ML tasks that could possibly support crisis message generation. An ML task defines the "terms of how the machine learning system should process [a collection of measured features] [19]." Three task categories were selected to assign fitting ML tasks for the requirement categories: The *classification* of data based on a certain characteristic, the *modification* of data and the *automatic creation* of texts without a given scripture. The task classes shown in Table 1 were derived from two literature reviews and an article identified during

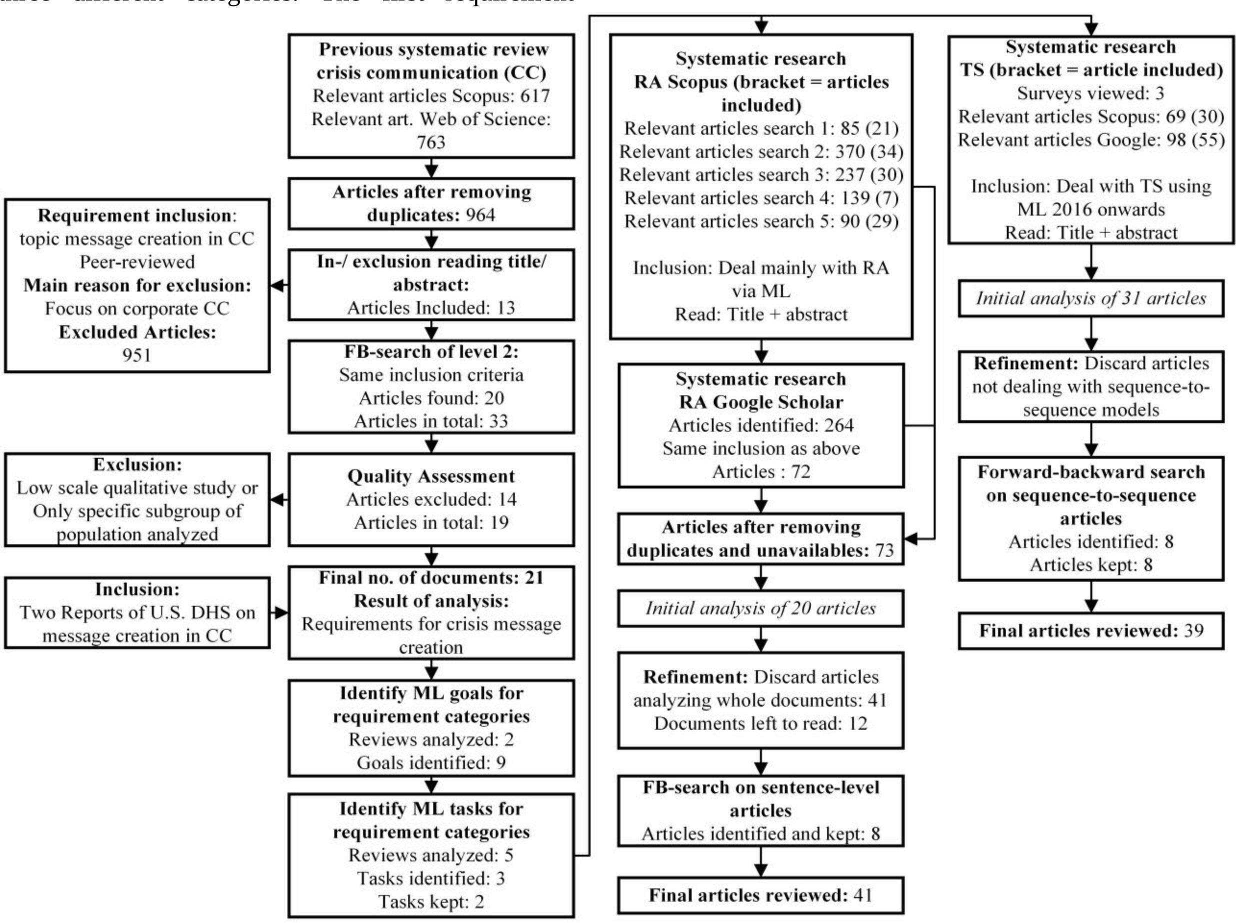

Figure 1. Reviewing process machine learning in crisis communication

the crisis message requirements review [23, 41, 63]. Three tasks, readability assessment, text simplification and content classification have also been identified via the same three publications. Readability assessment (RA) and text simplification (TS) were selected for a detailed analysis. Marked in Table 1, those tasks reflect the goals of assessing respectively adjusting text difficulty. Content analysis was not further analyzed, because it was researched extensively in the context of crisis management and social media, for example to classify tweets [23]. The third class (creation) was also not investigated further, because the initial search generated no relevant works.

The subsequent literature reviews on RA and TS are presented in Figure 1. RA describes the classification of a sentence based on its legibility. Legibility, refers to “the sum of elements of textual material that describe the understanding, reading speed, and degree of interest in the material [10].” In this paper, the term readability is used synonymously with comprehensibility. In the area of RA, there is the so-called *readability classification* in addition to relative comparisons of legibility between sentences and regression problems. In classification, the respective text is assigned to a pre-defined class depending on its readability level [10].

Table 2. Tasks text simplification

Process	Source*
Lexical Substitution	76
Sentence Splitting	38, 61, 76
Reordering	76
Paraphrasing	38, 61
Deletion	38, 61
**Note: Task at least mentioned*	

TS goes beyond the analytical nature of RA. The aim is to reduce the complexity of a text and make it easier to understand [31]. An overview on the set of tasks is given in Table 2. Modifications to the input took place either at word or at sentence level. The difficulty of TS lies in the fact that, despite the simplification of the sentence, it must not diminish the meaning and expressiveness in the respective context. Thus, the exchange of a certain word by a possibly more widely used synonym (lexical substitution) can lead to grammatical errors which tend to reduce the overall understanding [76]. Grammatical changes like word reordering or sentence splitting tend to cause some syntactical errors, while sentences are not always simplified [76]. Like RA, the ML solutions can be divided into two categories: Statistical solutions and artificial neural network solutions. Only the latter are considered in this work. The reason for this is that in the majority of articles found, this approach was labeled pre-dominant [29, 75, 76]. Only one article describes statistical solutions as the better choice [77].

The literature search on Scopus for RA using five different strings resulted in 121 included hits. The review on Google Scholar resulted in 72 articles. After reading title/abstract/keywords and removing duplicates 73 articles remained, of which an initial amount of 20 articles was analyzed, before adjustments were made. Only two articles of those 20 initially read articles dealt with RA to *analyze single sentences or short texts*. We decided to discard articles covering RA on longer documents (20 articles analyzed, 40 out of 53 remaining articles on longer documents were discarded, so 13 articles left on sentence-level RA). The search was therefore adjusted to balance the rate between document and sentence level analyses. After working through the remaining 13 articles a forward-backward-search was conducted on the papers that use RA for single sentences, in which eight more articles have been identified and subjected to a full-text analysis afterwards. In the end, 41 articles on RA via ML have been reviewed.

One goal for the review of TS was to avoid another review process including several changing search strings, as it was the case for RA. At first three surveys on text simplification were reviewed to identify important keywords for the upcoming searches [42, 54, 55]. The final strings for Scopus and Google Scholar are listed below:

TITLE-ABS-KEY("sentence simplification" OR "text simplification") AND (LIMIT-TO(SUBJAREA, "COMP")) AND (LIMIT-TO(PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO(PUBYEAR, 2017) OR LIMIT-TO(PUBYEAR, 2016)) AND (LIMIT-TO(LANGUAGE, "English"))

allintitle: "sentence simplification" OR "text simplification"

Table 1. Machine learning goals for requirement categories

Requirements / Task class/	Linguistic Understanding	Message Framing (Reaction)	Components and Content Order
Class 1: Classification	**Assess text difficulty**	(Emotional) Reaction prediction	Check completeness and correct order
Class 2: Modification	**Adjust the difficulty of the text**	Adjust the choice of words to cause the desired reaction	Adjust the content and order of information
Class 3: Creation	Automatic content creation for a given difficulty level	Automatic content creation according to the desired reaction	Create crisis warnings automatically

The scope of analysis was adjusted based on the review of the 29 identified articles. The so-called sequence-to-sequence (seq2seq, see 4.2) deep learning approach led to the best results in TS. Therefore, only those models were considered further on. As a next step, a forward-backward-search was conducted on the seq2seq articles in which eight further articles on that topic were localized. Additionally, potential updated research of the identified authors was searched and included. In total, 37 articles were reviewed. The topics covered in TS showed a higher degree of diversity than in RA, ranging from the construction of corpora, that are datasets of texts for training, to automated evaluation metrics. The research goal, as well as architecture or evaluation model in question, including features to classify texts (see 3.2 for specific examples), the corpora (if existent) and model performances were extracted for each article.

3. In-depth: Readability Assessment (RA)

3.1. Comparison to Traditional Formulas

Classical formulas in the field of RA, such as the Flesch-Kincaid and Coleman-Liau indices [16], are established tools in crisis management for the evaluation of news on social media [53] and websites [43]. However, these approaches reveal significant limitations in terms of reliability when applied to texts with fewer than 300 words [10, 26]. Also, they often ignore important factors for legibility, such as cohesion or ambiguities of individual words [10]. These limitations can lead to questionable results, especially with the evaluation of shorter messages. Hence, they do not seem suitable for the evaluation of CC. In contrast, ML solutions are used in various application-areas concerning the recognition and evaluation of complex semantic features in texts [10, 13, 14, 37, 74]. Several neural networks based solutions showed higher performances than statistical methods for shorter texts, scoring spearman rank correlations between around 0.5 and 0.7 from 25 respective 100 words, where statistical methods scored only between 0.1 and 0.4 [37].

3.2. Machine Learning Approaches in Readability Assessment

Within the reviewed RA articles, a general distinction was made between two different approaches to ML: Statistical machine learning methods based on a fixed selection of features on the one hand [10] and artificial neural network methods on the other hand [37]. The evaluation of the features in the statistical approach is trained by supervised ML architectures [10]. Prerequisite is the sufficient presentation of labeled training data for the respective features, like for example of lexical (e.g. word familiarity, ambiguous terms) or syntactic nature (e.g. sentence complexity) [10]. As shown by Vajjalla and Meurers (2014), features can also be of morphological, psycholinguistic nature [66]. In their work, morphological features include for example the derivations or compositions of words. Among others, Vajjala and Meurers name imageability or the age of acquisition as psycholinguistic features [66]. Often the number of features varies between 50 and 100 [9, 12, 14, 16, 44, 69, 74], sometimes more than 100 features are used [21, 66, 69]. Dell'Orletta et al. (2014) conclude that in a binary classification of Italian newspaper articles using 14 features on document level and 30 features on sentence level respectively, a further increase of features did not lead to significant performance improvement. It should be noted that this cannot be transferred one-to-one to other texts and languages, as levels of difficulty vary on the language analyzed [15].

The overwhelming majority of the articles found were based on the application of statistical solutions, such as support vector machines [9, 13, 21, 30, 67, 78]. On the other hand, only few articles considered neural networks. Two models use more complex deep learning architectures that are not based on comparatively simple neural networks [33, 37]. The networks of Nadeem and Ostendorf (2018) are equipped with a so-called attention head in four different setups, which enables weighting the semantic relevance of individual words and/or sentences [37].

3.3. Current Performance of Readability Assessment

Table 3 and Table 4 illustrate the accuracy scores of classification models run by the respective authors according to the number of assignment classes. In each study, text pieces are assigned depending on their readability. If several classifications were run, the setup scoring the best result is listed. In most cases the sentences were divided into two classes only, or compared in ranking procedures of two text pairs each. According to the results, the performance of the classification procedures tends to decline with an increased number of classes, at least for document level. In general, the accuracy of classification tends to decline for shorter texts. One could reason intuitively that a higher degree of difficulty stems from a smaller amount of text. Still, many scores reach more than 80% correct classifications. However, the high-performance values should not be overrated, because the performance highly depends on the complexity of the datasets and therefore limits comparability.

Table 3. Document level classifications

Publication	#Classes	Lang	Acc
Clercq and Hoste (2016)	2	Eng	96
Clercq and Hoste (2016)	2	Dut	98
Dalvean and Enkhbayar (2018)	2	Eng	89
Mesgar and Strube (2018)	2	Eng	97
Curto et al. (2015)	3	Por	81
Razon and Barnden (2015)	3	Eng	95
Pilán and Volodina (2016)	4	Swe	72
Clercq and Hoste (2016)	5	Eng	71
Clercq and Hoste (2016)	5	Dut	73
Curto et al. (2015)	5	Por	75
Hartmann et al. (2016)	5	Por	52
Vajjala and Meurers (2014)	5	Eng	90
Jiang et al. (2015)	6	Eng	92
Jiang et al. (2015)	6	Chi	51
Huang et al. (2018)	7	Eng	42
Lang = Language, Eng = English, Dut = Dutch, Por = Portuguese, Swe = Swedish, Chi = Chinese, Acc = Accuracy			

4. In-depth: Text Simplification (TS)

4.1. The Neuronal Sequence-to-Sequence Approach

TS using complex deep learning solutions is currently mostly based on the seq2seq approach. It consists of the two following basic steps: *encoding* and *decoding* [31]. In *encoding* a text sequence of any length is accepted as input and an output vector is calculated. This output vector serves as input for the second step, *decoding*. Depending on the properties of this vector, the output set is created word by word. Some researchers tune their model by using enhancements to improve performances. Guo et al. (2018) influence the output values of their model by results of two external auxiliary tasks [20]. Zhang and Lapata (2017) define a reward function, that includes several variables evaluating the potential reading flow, simplicity and relevance of the content [75]. Zhang et al. (2017) perform a purely lexical simplification of individual words which must be included in the output set [76]. Most TS solutions use subtypes of recurrent neural networks [5, 20, 31, 39, 57, 60, 61, 68, 75, 76]. The use of recurrent neural networks (RNN) is prevalent in the evaluation of languages, because the respective output depends on the previous or additionally subsequent inputs. This allows to select the decision of the next word, when creating a sentence depending on the surrounding terms [68]. A special case among the identified articles is the so-called multi-head-attention transformer model, which outperforms their RNN-based counterparts in two studies [29, 77].

Table 4. Sentence level classifications

Publication	#Classes	Lang	Acc
Ambati et al. (2016)	2*	Eng	78
Curtotti et al. (2015)	2	Eng	77
Liu and Matsumoto (2017)	2	Jap	84
Mesgar and Strube (2016)	2	Eng	76
Mukherjee et al. (2018)	2	Eng	90
Schumacher et al. (2016)	2*	Eng	84
Vajjala and Meurers (2014)	2	Eng	66
Vajjala and Meurers (2016)	2*	Eng	82
Azpiazu and Soledad Pera (2016)	3	Eng	81
Stajner et al. (2016)	3	Eng	57
Pilán et al. (2016)	5	Swe	63
**Ranking procedure of two text pieces Lang = Language, Eng = English, Jap = Japanese, Swe = Swedish, Acc = Accuracy*			

4.2. Current Performance of Text Simplification

Whether the given models for TS can already be used effectively in CC depends largely on their performances and ability to measure them efficiently. Table 5 shows human evaluations between the simplified model outputs and their original references in the dimensions of grammaticality, adequacy (i.e. meaning preservation) and simplicity of the text. We harmonized the values to fit into a 1 – 5 scale to improve comparability. Studies listed more than once show numbers from different corpora. If several models were tested, we selected the one with the best simplicity score for each corpus. Surprisingly, an increase in simplicity did not always result in losses in terms of grammaticality or content adequacy. This could be due to complexity differences of the given references. In addition, the models tend to differ in the number of simplification operations carried out, ranging from simple lexical substitutions only to the deletion and rephrasing of whole sentence-parts. A precise assessment on the suitability of individual models can hardly be made based on these values only, especially since there is no threshold defined for acceptance in CC.

Table 5. Performances of text simplification models

Publication/Metric	Grammar Reference	Grammar Model	Adequacy Reference	Adequacy Model	Simplicity Reference	Simplicity Model	Corpus
Guo et al., 2018	4,97	4,73	4,08	3,18	3,83	4,62	Newsela
Vu et al., 2018	4,58*	4,24	2,98*	3,03	3,99*	3,45	Newsela
Vu et al., 2018	4,63*	4,57	3,97*	3,28	3,59*	3,81	WikiSmall
Vu et al., 2018	4,59*	4,65	4,43*	3,95	2,38*	2,90	WikiLarge
Sulem et al., 2018	4,8*	3,98	5*	3,33	*3**	*3,68*	PWKP
Zhang & Lapata, 2017	3,9*	3,65	2,81*	2,94	3,42*	3,1	Newsela
Zhang & Lapata, 2017	3,74*	3,92	3,34*	3,36	3,13*	3,55	WikiSmall
Zhang & Lapata, 2017	3,79*	2,60	3,72*	2,42	2,86*	3,52	WikiLarge
Zhang et al., 2017	*5*	*3,60*	*5*	*3,65*	*1*	*2,62*	PWKP
Xu et al., 2016	*5*	*4,5*	*5*	*4,16*	0**	0,65**	Wiki by Coster

Legend: Italic entries harmonized onto 1 – 5 Likert scale
**Reference is an already human-simplified sentence*
***Average number of successful paraphrases of model (1,35 when sentence was simplified by humans)*

5. Discussion

In this section, functions and challenges for application in crisis communication will be discussed for both readability assessment and text simplification. For each subchapter we will discuss the *applications of CC and non-CC-specific corpora*, as well as the *reliability of existing solutions in static and turbulent environments*. Additionally, for RA the challenges include the *improvement of shorter texts and assessments towards reliability of binary- and multi-classifications.* Finally, TS specific challenges remain *improving automatic performance measures* and *balancing the simplification and meaning preservation* of simplified texts according to CC standards.

5.1. Challenges for Readability Assessment in Crisis Communication

The performance of ML approaches for single sentences or short texts (e.g. tweets) is lower compared to the document level scores, especially with more than two assignment classes used. The solutions found for shorter texts are often based on rather simple binary classifications, leaving room for improvement. Thus, common solutions for the evaluation of Twitter messages are rather unsuitable [3]. Nadeem and Ostendorf (2018) also note that in this context statistical methods often deliver very poor performances and point to the need for research regarding effective deep learning models to address this problem [37]. Still, there is potential to use RA methods on both document and sentence level to support crisis communications. Document RA could support the creation of texts in rather static environments, for example to check websites or vouchers. Sentence RA might be even more important, in case of short statements to the public, when timely action is required. In that sense it could support reaching the CC requirement of comprehensibility through signalizing if a text meets or exceeds the intended complexity. We recommend testing the reliability of binary and multi-classifications in CC contexts.

The potential added value of RA methods in CC highly depends to a large extent on the availability of sufficient high-quality training data [19]. There are already several larger corpora that could serve as a basis for initial tests. In line with the requirement to use sixth grade level language or lower, initial tests may be conducted using the WeeBit [65], or Common Core corpus [17]. These datasets contain texts classified by grade levels. It might also be discussed whether it makes sense to perform manual annotations for CC-specific corpora. Yaneva et al. (2017) conclude that, although small domain-specific corpora are not sufficient to produce a meaningful result, the data, in conjunction with a large general corpus, can provide improved performance in certain contexts [74]. Dell'Orletta et al. (2014) compared the performance of a small data set, which was created by manually selecting sentences, with some larger sets, in which the texts were extracted automatically without insight [15]. They recognized small advantages of the complex manual annotation set [15]. In this respect, the costly annotation of a corpus for CC might be a useful investment, especially when considering the danger of unknown jargon influencing the RA.

5.2. Challenges for Text Simplification in Crisis Communication

In theory, TS could enhance the analysis of RA by automatically simplifying sentences that do not meet the expected readability goals. While influencing the complexity of the task, this could be accomplished in any context where a text is to be received by the public (as in the examples given in 5.1). TS extends from analysis to modification, which explains the more complex challenges that must be tackled, before successfully adopting it in CC. Most of the following shortcomings of current solutions affect the challenge of balancing the goals of simplification and meaning preservation.

The main challenge with TS is the difficulty in recognizing words that are of central importance in a particular context. In standard seq2seq architectures, for example, there is no simple copying of the most important words, which can sometimes lead to severe losses in meaning preservation [5]. Often, simplification operations would be performed without considering the semantic relevance of individual phrases in the context of the text [31]. Ma and Sun (2017) extend their model by introducing a *self-gated encoder* to the standard encoding [31]. This provides input words with an additional factor that declares the importance of individual words according to information content and thus influences the inclusion of words in the output record. Others use a *pointer-generator-network* [20, 29] or similar modifications [5]. It enables the direct copy of a word into the output record. A probability value is calculated, which describes the inclusion on a new word from a vocabulary. The *pointer-copy-network* [29] is specifically dedicated to deal with *out-of-vocabulary*, words that the model was not trained with and whose meaning and relevance is therefore unknown. However, current models in the use of *out-of-vocabulary* are still very immature [29, 57]. From the CC perspective, the solution of this problem is particularly relevant as correct processing of domain-specific technical terms must be regarded as essential for communication with the population. Deleting or incorrectly replacing *out-of-vocabulary* could seriously affect the understanding of a message, especially if the recipient is under stress.

TS models show significant performance losses, especially with longer and syntactically more complex sentences [29], which could emerge due to the insufficient storage capacity of longer dependencies in LSTM models [68]. Attempts that facilitate the recognition of longer dependencies include the *pointer-generator-network* [29], and a *neural-semantic-encoder*, which stores additional dependencies in an additional matrix [68]. Sulem et al. (2018) try to address this problem by first performing a sentence splitting step that converts complex sentences into single shorter ones, which led to an increase in simplification operations [60].

Common TS models tend to underestimate the number of possible changes to a text or sentence [8]. Furthermore, neural networks specialize in the application of frequently occurring rules, so that difficulties can arise in syntactic exceptions [77]. This may result in the output not being optimally simplified or grammatically incorrect. One solution is the sentence splitting, which drastically increased the set of operations in a given dataset [60]. With regards to CC, the correctness of simplifications is indispensable. Hence, the correctness of simplification should be a more important goal than maximizing the operations performance. For crisis-warnings it has been shown that an increased amount of information resulted in higher message credibility [48, 62], resulting in a risk of gaining comprehensibility at the cost of completeness and ultimately credibility. The potential trade-off between comprehensibility and completeness is what we see as one of the main challenges to deploy TS successfully not only in rather static areas as websites, where some errors might be forgiven, but in rapidly evolving in-crisis-scenarios where credibility and trust in crisis managers is an important goal.

Deep learning solutions do not require manually defined rules for performing operations, but large amounts of training data instead. As with RA, those should preferably be available in annotated sentence pairs [77]. It was found that the current amount of annotated corpora is insufficient for TS [1, 45, 50, 71]. In addition, existing data sets were criticized for their lack of quality. The main complaint covered the existence of only one single simplified alternative [75]. Wikipedia datasets seem inefficient, as only half of the sentence pairs analyzed were actual simplifications [73]. Also, the low agreement of human annotators in the creation of manual corpora was criticized [8]. As with RA, the question arises as to whether CC specific corpora should be created.

The works in Table 5 also use automatic performance metrics to compare models more efficiently, compared to costly human evaluations. The problem is that the most popular metrics have been criticized heavily in former works and seem fairly unreliable to use in a real-world context [8, 56, 58, 59]. Therefore, we decided to not rely on these rather controversial metrics and leave this issue open for further research.

6. Limitations and Conclusion

In the area of linguistic comprehension, the RA performance shows decent potential through its successful application in other contexts, especially in the evaluation of longer documents. Since ML techniques were often described as superior to traditional methods, it should be of interest to examine the existing possibilities regarding CC. Meanwhile the assessment of TS solutions does not seem possible at this stage without initial testing. The main reasons for this are the unreliable evaluation methods of the output sequences and the scores of human evaluations which are difficult to interpret.

For both ML tasks, however, there is still no training data tailored to CC. Depending on the task, it should be examined whether existing corpora already achieve sufficient performance. Future pilot studies on the implementation of initial solutions could therefore examine the potential presumed in this work. First tests on existing architectures and the potential value of generating crisis communication specific training corpora could provide more in-depth assessment on the application of ML in CC. For RA, an initial binary classification of crisis management documents or websites could possibly be carried out first using a large publicly accessible corpus, which classifies the texts into *below* or *at acceptable level* or *above acceptable level,* respectively. In case of TS, a first pilot study could provide initial insights into what results are possible with existent corpora. In any case, the lack of datasets to train the respective ML models seems to be one of the main problems in both tasks.

The ML tasks included here are based on the previously identified requirements of CC to messages in text form shown in Table 1. Further research could dive deeper into message requirements other than linguistic understanding. An area of particular interest would be the category of message framing. As mentioned in the beginning, this category deals with the emotional reaction and influence on the receiver. It would be interesting to see if current techniques of ML handle this task, to assess the effect of messages before sending. The topic of automated message creation was also underrepresented in the research and could be subject of future research. Overall, the different utilizations of machine learning in textual crisis communication remain widely unexplored.

7. References

[1] Alva-Manchego, F., J. Bingel, G. Paetzold, C. Scarton, and L. Specia, "Learning How to Simplify From Explicit Labeling of Complex-Simplified Text Pairs", in Proceedings of the Eighth International Joint Conference on Natural Language Processing, Taipei, Taiwan. 2017.

[2] Ambati, B.R., S. Reddy, and M. Steedman, "Assessing Relative Sentence Complexity using an Incremental CCG Parser", in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California. 2016.

[3] Azpiazu, I.M. and M. Soledad Pera, "Is Readability a Valuable Signal for Hashtag Recommendations?", RecSys Posters, 2016.

[4] Bean, H., B.F. Liu, S. Madden, J. Sutton, M.M. Wood, and D.S. Mileti, "Disaster Warnings in Your Pocket: How Audiences Interpret Mobile Alerts for an Unfamiliar Hazard", Journal of Contingencies and Crisis Management, 24(3), 2016, pp. 136–147.

[5] Cao, Z., C. Luo, W. Li, and S. Li, "Joint Copying and Restricted Generation for Paraphrase", CoRR, 2017.

[6] CDC, Crisis and Emergency Risk Communication: 2014 Edition, 2014.

[7] CDC, CERC: Messages and Audiences, 2018.

[8] Choshen, L. and O. Abend, "Inherent Biases in Reference-based Evaluation for Grammatical Error Correction and Text Simplification", CoRR, 2018.

[9] Clercq, O. de and V. Hoste, "All Mixed Up? Finding the Optimal Feature Set for General Readability Prediction and Its Application to English and Dutch", Computational Linguistics, 42(3), 2016, pp. 457–490.

[10] Collins-Thompson, K., "Computational assessment of text readability A survey of current and future research", ITL - International Journal of Applied Linguistics, 165(2), 2014, pp. 97–135.

[11] https://www.cdc.gov/coronavirus/2019-ncov/index.html, accessed 10-2-2020.

[12] Curto, P., N. Mamede, and J. Baptista, "Automatic Text Difficulty Classifier - Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching", in Proceedings of the 7th International Conference on Computer Supported Education, Lisbon, Portugal. 2015.

[13] Curtotti, M., E. McCreath, T. Bruce, S. Frug, W. Weibel, and N. Ceynowa, "Machine learning for readability of legislative sentences", in Proceedings of the 15th International Conference on Artificial Intelligence and Law - ICAIL '15, San Diego, California. 2015.

[14] Dalvean, M.C. and G. Enkhbayar, "A New Text Readability Measure for Fiction Texts", SSRN Electronic Journal, 2018.

[15] Dell'Orletta, F., M. Wieling, G. Venturi, A. Cimino, and S. Montemagni, "Assessing the Readability of Sentences: Which Corpora and Features?", in Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, Maryland. 2014.

[16] Denning, J., M.S. Pera, and Y.-K. Ng, "A readability level prediction tool for K-12 books", Journal of the Association for Information Science and Technology, 67(3), 2016, pp. 550–565.

[17] English Language Arts & Literacy in History/Social Studies, Science and Technical Subjects: Appendix B:

Text Exemplars and Sample Performance Tasks, Common Core State Standards Initiative.
[18] FEMA, Lesson 3. Communicating in an Emergency, 2014.
[19] Goodfellow, I., Y. Bentio, and A. Courville, Deep Learning: An MIT Press book, 2016.
[20] Guo, H., R. Pasunuru, and M. Bansal, "Dynamic Multi-Level Multi-Task Learning for Sentence Simplification", CoRR, 2018.
[21] Hartmann, N., L. Cucatto, D. Brants, and S. Aluisio, Automatic Classification of the Complexity of Nonfiction Texts in Portuguese for Early School Years, Springer International Publishing, Cham, 2016.
[22] Huang, Y.-T., M. Chang Chen, and Y.S. Sun, "Characterizing the Influence of Features on Reading Difficulty Estimation for Non-native Readers", CoRR, 2018.
[23] Imran, M., C. Castillo, F. Diaz, and S. Vieweg, "Processing Social Media Messages in Mass Emergency: A Survey", CoRR, 2015.
[24] Jiang, Z., G. Sun, Q. Gu, T. Bai, and D. Chen, "A Graph-based Readability Assessment Method using Word Coupling", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal. 2015. Association for Computational Linguistics: Stroudsburg, PA, USA.
[25] Kapucu, N., "Interagency Communication Networks During Emergencies", The American Review of Public Administration, 36(2), 2006, pp. 207–225.
[26] Kidwell, P., G. Lebanon, and K. Collins-Thompson, "Statistical Estimation of Word Acquisition With Application to Readability Prediction", Journal of the American Statistical Association, 106(493), 2011, pp. 21–30.
[27] Kuligowski, E. and P. Dootson, "Emergency Notification: Warnings and Alerts", in Encyclopedia of Wildfires and Wildland-Urban Interface (WUI) Fires. 2018. Springer International Publishing: Cham.
[28] Lang, S., L. Fewtrell, and J. Bartram, Water quality: Guidelines, Standards and Health: Assessment of risk and risk management for water-related infectious disease, IWA Publ, London, 2002.
[29] Li, T., Y. Li, J. Qiang, and Y.-H. Yuan, "Text Simplification with Self-Attention-Based Pointer-Generator Networks", in Neural Information Processing. 2018. Springer International Publishing: Cham.
[30] Liu, J. and Y. Matsumoto, "Sentence Complexity Estimation for Chinese-speaking Learners of Japanese", in Philippines (Hg.) 2017 – Proceedings of the 31st Pacific.
[31] Ma, S. and X. Sun, "A Semantic Relevance Based Neural Network for Text Summarization and Text Simplification", CoRR, 2017.
[32] Mesgar, M. and M. Strube, "Lexical Coherence Graph Modeling Using Word Embeddings", in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California. 2016.
[33] Mesgar, M. and M. Strube, "A Neural Local Coherence Model for Text Quality Assessment", in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018.
[34] Mileti, D.S. and J.H. Sorensen, Communication of Emergency Public Warnings: A Social Science Perspective and State-of-the-Art Assessment, 1990.
[35] Moran, B., Crisis and Emergency Risk Communication Basic Principles: Host: Belen Moran, Presenters: Bret Atkins Date/Time: August 4, 2011 3:00 pm ET, 2011.
[36] Mukherjee, P., G. Leroy, and D. Kauchak, "Using Lexical Chains to Identify Text Difficulty: A Corpus Statistics and Classification Study", IEEE journal of biomedical and health informatics, 2018.
[37] Nadeem, F. and M. Ostendorf, "Estimating Linguistic Complexity for Science Texts", in Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans, Louisiana. 2018.
[38] Narayan, S., C. Gardent, S.B. Cohen, and A. Shimorina, "Split and Rephrase", CoRR, 2017.
[39] Nisioi, S., S. Stajner, S.P. Ponzetto, and L.P. Dinu, "Exploring Neural Text Simplification Models", in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL, Vancouver, Canada. 2017.
[40] Omori, H., E.D. Kuligowski, S.M.V. Gwynne, and K.M. Butler, "Human Response to Emergency Communication: A Review of Guidance on Alerts and Warning Messages for Emergencies in Buildings", Fire Technology, 53(4), 2017, pp. 1641–1668.
[41] Otter, D.W., J.R. Medina, and J.K. Kalita, "A Survey of the Usages of Deep Learning in Natural Language Processing", CoRR, 2018.
[42] Paetzold, G.H. and L. Specia, "A Survey on Lexical Simplification", Journal of Artificial Intelligence Research, 60, 2017, pp. 549–593.
[43] Phelan, T., Literacy Matters: EM Messages & Readability Levels: Urban Assembly School for Emergency Management, Advisory Board, 2015.
[44] Pilán, I., S. Vajjala, and E. Volodina, "A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity", CoRR, 2016.
[45] Qiang, J., "Improving Neural Text Simplification Model with Simplified Corpora", CoRR, 2018.
[46] Razon, A. and J. Barnden, "A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features", in Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, September. 2015.
[47] Reynolds, B. and M. W. Seeger, "Crisis and emergency risk communication as an integrative model", Journal of health communication, 10(1), 2005, pp. 43–55.
[48] Sattler, D.N., K. Larpenteur, and G. Shipley, "Active Shooter on Campus: Evaluating Text and E-mail Warning Message Effectiveness", Journal of Homeland Security and Emergency Management, 8(1), 2011.

[49] Savoia, E., L. Lin, and K. Viswanath, "Communications in public health emergency preparedness: a systematic review of the literature", Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, 11(3), 2013, pp. 170–184.
[50] Scarton, C., G.H. Paetzold, and L. Specia, "Text Simplification from Professionally Produced Corpora", in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan, May. 2018.
[51] Schumacher, E., M. Eskenazi, G. Frishkoff, and K. Collins-Thompson, "Predicting the Relative Difficulty of Single Sentences With and Without Surrounding Context", CoRR, 2016.
[52] Schwarz, S. Andreas, Matthew W., and Auer Claudia, eds., The Handbook of International Crisis Communication Research, Wiley Blackwell, Chichester, 2016.
[53] Scott, K.K. and N.A. Errett, "Content, Accessibility, and Dissemination of Disaster Information via Social Media During the 2016 Louisiana Floods", Journal of public health management and practice: JPHMP, 24(4), 2018, pp. 370–379.
[54] Shardlow, M., "A Survey of Automated Text Simplification", International Journal of Advanced Computer Science and Applications, 4(1), 2014, pp. 58–70.
[55] Siddharthan, A., "A survey of research on text simplification", ITL - International Journal of Applied Linguistics, 165(2), 2014, pp. 259–298.
[56] Stajner, S., "Machine Translation Evaluation Metrics for Quality Assessment of Automatically Simplified Sentences", in QATS: Workshop on Quality Assessment for Text Simplification. 2016.
[57] Stajner, S. and S. Nisioi, "A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification", in LREC 2018: Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan. 2018.
[58] Stajner, S., M. Popovic, H. Saggion, and M. Fishel, "Shared Task on Quality Assessment for Text Simplification", in LREC Workshop & Shared Task on Quality Assessment for Text Simplification (QATS), The International Conference on Language Resources and Evaluation, Portorož. 2016.
[59] Sulem, E., O. Abend, and A. Rappoport, "BLEU is Not Suitable for the Evaluation of Text Simplification", CoRR, 2018.
[60] Sulem, E., O. Abend, and A. Rappoport, "Simple and Effective Text Simplification Using Semantic and Neural Methods", CoRR, 11.10.2018.
[61] Surya, S., A. Mishra, A. Laha, P. Jain, and K. Sankaranarayanan, "Unsupervised Neural Text Simplification", CoRR, 2018.
[62] Sutton, J., S.C. Vos, M.M. Wood, and M. Turner, "Designing Effective Tsunami Messages: Examining the Role of Short Messages and Fear in Warning Response", Weather, Climate, and Society, 10(1), 2018, pp. 75–87.
[63] Temnikova, I., S. Vieweg, and C. Castillo, "The Case for Readability of Crisis Communications in Social Media", in WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web. 2015.
[64] Templier, M., "A Framework for Guiding and Evaluating Literature Reviews", Communications of the Association for Information Systems(37), 2015, pp. 112–137.
[65] V., S. and D. Meurers, On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition, 2012.
[66] Vajjala, S. and D. Meurers, "Assessing the relative reading level of sentence pairs for text simplification", in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden. 2014.
[67] Vajjala, S. and D. Meurers, "Readability-based Sentence Ranking for Evaluating Text Simplification", CoRR, 2016.
[68] Vu, T., B. Hu, T. Munkhdalai, and H. Yu, "Sentence Simplification with Memory-Augmented Neural Networks", CoRR, 2018.
[69] Wagner Filho, J.A., R. Wilkens, and A. Villavicencio, "Automatic Construction of Large Readability Corpora", in Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, Osaka, Japan. 2016.
[70] Walaski, P., Risk and Crisis Communications: Methods and Messages, John Wiley & Sons, Incorporated, New York, U.S.A., 2011.
[71] Wang, T., P. Chen, K. Amaral, and J. Qiang, "An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification", CoRR, 2016.
[72] White, K., Effective Communication: Independent Study, Federal Emergency Management Agency, 2005.
[73] Xu, W., C. Callison-Burch, and C. Napoles, "Problems in Current Text Simplification Research: New Data Can Help", Transactions of the Association for Computational Linguistics, 3(4), 2015, pp. 283–297.
[74] Yaneva, V., C. Orasan, R. Evans, and O. Rohanian, "Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities", in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Copenhagen, Denmark. 2017.
[75] Zhang, X. and M. Lapata, "Sentence Simplification with Deep Reinforcement Learning", CoRR, 2017.
[76] Zhang, Y., Z. Ye, Y. Feng, D. Zhao, and R. Yan, "A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification", CoRR, 2017.
[77] Zhao, S., R. Meng, D. He, S. Andi, and P. Bambang, "Integrating Transformer and Paraphrase Rules for Sentence Simplification", CoRR, 2018.
[78] Zheng, J. and H. Yu, "Assessing the Readability of Medical Documents: A Ranking Approach", JMIR medical informatics, 6(1), 2018, e17.