Replicability under Near-Perfect Conditions – A Case-Study from Automatic Summarization

Margot Mieskes
University of Applied Sciences Darmstadt
Germany
margot.mieskes@h-da.de

Abstract

Replication of research results has become more and more important in Natural Language Processing. Nevertheless, we still rely on results reported in the literature for comparison. Additionally, elements of an experimental setup are not always completely reported. This includes, but is not limited to reporting specific parameters used or omitting an implementational detail. In our experiments based on two frequently used data sets from the domain of automatic summarization and the seemingly full disclosure of research artifacts, we examine how well results reported are replicable and what elements influence the success or failure of replication. Our results indicate that publishing research artifacts is far from sufficient, and that publishing all relevant parameters in all possible detail is crucial, but often neglected, making the situation in automatic summarization only near-perfect.

1 Introduction

Replicability is gaining more and more attention in the NLP world with dedicated workshops[1], replication checklists[2] etc. While this improves the situation considerably, and the availability of research artifacts is improving, there is still the question if replicability is possible if all artifacts necessary are available. Additionally, often results from the literature are cited, but it is far from clear whether the reported results are obtained experimentally (by re-implementing or re-running a particular method) or also cited. One domain where the availability of research artifacts is almost perfect, is Automatic Summarization. Standard benchmark data sets published in the course of various shared tasks are available, the evaluation method is well known, its implementation is available and resulting data submitted to shared tasks have also been made available by the organizers. Therefore, it should be straightforward to replicate results reported by the organizers of the shared task, as well as results reported in the literature.

This would hardly be a submission to a workshop on insights from negative results if things were that easy. Normally, successfully replicating previous results would just appear as one or more number in a table used for comparison. But our results indicate that despite this near-perfect conditions, reporting and replicating results is far from straightforward. Based on a literature review and experiments in replicating results we show the discrepancies that occur both in cited results, as well as when experiments are replicated.

Our contributions are therefore a closer look and comparison of reported results from the domain of automatic summarization and results from replicated experiments and factor benefitting or hindering complete replication.

2 Replication in NLP

Experiments in reproducing results in the NLP domain such as those presented by Fokkens et al. (2013) are still quite rare. One reason is, that when undertaking such projects, "sometimes conflicting results are obtained by repeating a study"[3].

Fokkens et al. (2013) report, that their experiments on two tasks in NLP are difficult to carry out and to obtain meaningful results. Preprocessing, experimental setup, versioning, system output, and system variation cause experimental variation according to the authors.

The 4Real workshop[4] focuses on the "the topic of the reproducibility of research results and the citation of resources, and its impact on research

[1] examples are https://lrec2020.lrec-conf.org/en/reprolang2020/ and https://reprogen.github.io

[2] https://2021.naacl.org/calls/reproducibility-checklist/

[3] https://sciencebasedmedicine.org/science-based-medicine-101-/

[4] http://4real.di.fc.ul.pt/

Proceedings of the Third Workshop on Insights from Negative Results in NLP, pages 165 - 171
May 26, 2022 ©2022 Association for Computational Linguistics

integrity". Their call for papers asks for submissions of "actual replication exercises of previous published results" (Branco et al., 2016). Results from this workshop suggest that reproducing experiments gives additional insights, and is therefore beneficial for the researchers as well as for the community (Cohen et al., 2016).

Horsmann and Zesch (2017) present a study on the replication of results in the context of Part-of-Speech tagging and whether LSTMs really work as well as the literature suggests. The results are mixed and show that the replicability depends on parameters such as tagset complexity.

Crane (2018) looks into the area of Question Answering and finds that "Source code without a reproducible environment does not mean anything". The author presents a set of experiments to show, that different parameters can lead to different results, similar in magnitude to those reported in the literature.

Dror et al. (2017) give a more general overview on this issue, as they perform a replicability study on various NLP tasks. They find that the increasing amount of evaluation data sets is a two-edged sword and only beneficial if the data reflects a variety of linguistic phenomena and are heterogeneous at least with respect to language or domain. Otherwise, showing that results are valid on one data set is probably sufficient.

Other authors look into the availability of research artifacts (i.e. (Mieskes, 2017; Wieling et al., 2018) who found that a large proportion of research artifacts are not available. A recent study by Belz et al. (2021) systematically looked into the replicability of various publications from the NLP domain, finding, that only approx. 14 % of the examined publications were replicable.

3 Automatic Summarization

Fokkens et al. (2013), Crane (2018) and others observe that re-implementation does not guarantee the reproducibility of the reported results, but rather a range of parameters cause differences between reported results and replicated results. Therefore, we focus on available data, systems and differences due to the evaluation method.[5]

The **DUC 2002** data set is used for an evaluation on Single-Document Summarization (SDS). It contains over 500 documents from 59 thematic clusters. The target length of the summaries is 100 words. The **DUC 2004** data set is used for the evaluation on the Multi-Document Summarization (MDS) task. It contains 500 documents from 50 thematic clusters. The length restriction was set to 665 bytes, which, for English, also results in a length of 100 words.

For both data sets the organizers of the shared task published reference summaries as well as submitted summaries. Furthermore, the evaluation results are available as well. Lin (2004) introduced an automatic evaluation metric, which became the standard both for subsequent shared tasks, as well as for automatic summarization in general. ROUGE has a range of parameters, which have to be set prior to running the evaluation. Several of these parameters are not binary, which results in a extensive parameter space. Graham (2015) gives details on these parameters and the resulting issues.

Both data sets that have been widely used in the past 15 to 20 years and therefore provide a reasonable basis for our analysis, which contains three parts: First, we will look into results reported in the literature and we aim to replicate those reported results. Second, we use available summarization methods out of the box or retrain them and evaluate the results. Third, we use a data set published by Hong et al. (2014) to replicate their results.

In our experiments, we stick as close as possible to the description offered in the cited publications and cite the results given.

3.1 Single Document Summarization (SDS)

Table 1 lists the ranking for DUC 2002 both based on the officially released results[6], as well as three examples from the literature: Lloret and Palomar (2010); Mihalcea and Tarau (2004) and Barrera and Verma (2011). Table 2 additionally shows results reported in these three papers. We experiment with various settings for ROUGE, relying on parameters reported in the literature.

We specifically focus on the stopword and stemming parameters, as we observe that they result huge differences in the results – marked as "Stopwords" and "Stemmed" in the table. "Both" indicates that stopwords were filtered and stemming was applied. Both tables (1 and 2) show that there is quite some discrepancy between the rankings

[5] Please note, that we do not report all publications that cite the same results, but rather highlight the differences.

[6] S19 and S27 are very close together and the error bars as published in `https://www-nlpir.nist.gov/projects/duc/pubs/2002slides/overview.02.pdf` do not allow for an exact distinction between the two.

reported officially and those in the literature. The comparison between the official results and the results in the literature might not be quite appropriate, as the official evaluation has not been done using ROUGE and while ROUGE has shown high correlation with human judgements, the ranking does not necessarily match exactly. The situation is somewhat different for the three reported rankings, which have all been done using ROUGE, as can be seen in Table 2.

Loret	Barrera	Mihalcea	official
S28	S28	S27	S19
S21	S19	S31	S27
S19*	S21*	S28	S28
–	S29*	S21	S21
–	S23*	S29*	S31

Table 1: Ranking as listed in the literature; * did not beat the baseline according to the source paper.

Some systems (i.e. S31) do not even occur in all three reported rankings. A closer look at the reported and replicated ROUGE-scores show that they vary considerably. We also observe that applying stopword filtering gives the worst results, while applying stemming gives the highest results, which are also similar to results reported by Mihalcea and Tarau (2004, 2005) and Barrera and Verma (2011). Applying both stopword filtering and stemming gives results that are in a similar range to those reported by Lloret and Palomar (2010). It is interesting to note, that in all four papers the baseline is reported differently: 0.4779 (Barrera and Verma, 2011), 0.4599 (Mihalcea and Tarau, 2004), 0.4799 (Mihalcea and Tarau, 2005) and 0.4113 (Lloret and Palomar, 2010). As only Lloret and Palomar (2010) note the parameters for the evaluation[7] this is the only experiment we could replicate in detail. But differences remain. It is interesting to see that while Mihalcea and Tarau (2004) also experimented with stemming and stopword filtering, they report the best results when using the basic settings, while our results are highest when stemming is applied, whereas stopword filtering gives the worst results.

3.2 Multi-Document Summarization (MDS)

For the MDS scenario the situation is somewhat better as ROUGE has been used in the official evaluation as well. The best system was identified as S65 and there is no discrepancy we could find in the literature regarding this.

Citation	S28	S21	S19
Mihalcea and Tarau (2004)	0.4703	0.4683	na
stemmed	0.4890	0.4869	na
stemmed/no stopwords	0.4346	0.4222	na
Mihalcea and Tarau (2005)	0.4890	0.4869	na
Lloret and Palomar (2010)	0.4278	0.4149	0.4082
Barrera and Verma (2011)	0.4781	0.4754	0.4552
Stemmed	0.473	0.467	0.452
Stopwords	0.395	0.380	0.379
Both	0.421	0.406	0.404

Table 2: Evaluation results for systems in DUC 2002 based on reports from the literature and based on our own replication with various parameter settings.

basic	Stemmed	Stopwords	Both
0.35909	0.38317	0.27068	0.30595

Table 3: Results for various preprocessing parameters for the output for S65 from DUC 2004.

Table 3 presents our results for evaluating S65 with various preprocessing parameters. As with the DUC 2002 data, stemming the resulting summaries give the best results, while the basic parameters only give the second best results.

Citation	ROUGE-1
Original	0.38224
Yih et al. (2007)†	0.305
Alguliev et al. (2012)	0.3822
Ryang and Abekawa (2012)	0.3827
Manna et al. (2012)†	0.3913
Rioux et al. (2014)†	0.3828
Ren et al. (2016)†	0.3788
Wang et al. (2017)†	0.3762

Table 4: Results on S65 as reported by the organizers (Original) and in various publications ever since. † indicates that parameters have been reported in the publication.

Table 3 presents the results for S65 as officially reported and various results found in the literature, which show a considerable range. When running ROUGE on the available data with various parameter settings we observe that the results also vary considerably, similar to the SDS scenario. Comparing the results in Table 3 to those officially published and reported in the literature (Table 4) we observe that applying stemming gives results close to what has been officially reported. Applying both stemming and stopword filtering our results are close to those reported by Yih et al. (2007). As indicated, most of the cited papers also report the evaluation parameters. A closer look at these parameters shows that although there are some differences, the parameters affecting ROUGE-1 are the same, ex-

[7] -n 2 -m 2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -l 100 -d

cept for Rioux et al. (2014), where `-l 250` was used. This allows summaries to be longer than 100 words, which could have a considerable effect on the ROUGE scores. Ren et al. (2016) do not set any length parameter, which means that the summaries are evaluated in their full length. Ren et al. (2016) presents a summarization method that ensures a final length of 100 words. And in all cases, stemming was applied, but no stopword filtering. Taking this into account, our results are similar to those originally reported, but also to those reported by Alguliev et al. (2012), Ryang and Abekawa (2012) and Rioux et al. (2014), where longer summaries were considered.

3.3 Re-run Summarization Methods

For the 2004 MDS data we perform two additional experiments. First, we use MEAD which has successfully participated in various shared tasks on automatic summarization. Second, we follow instructions to retrain and run an SVM-based summarization method and compare our evaluation with the reported results.

MEAD can be downloaded[8] and used for summarization. Therefore, we use the code as is to summarize the DUC 2004 data. Table 5 shows the results found in the literature. Preprocessing has a considerable influence on the results, as with no preprocessing we only achieve R-1 = 0.31 and the best result is R-1 = 0.349. This is still lower than the reported results, which are considerably higher and as with previous experiments, vary considerably. Unfortunately, only Hong et al. (2014) report the parameters used, but nevertheless, our results are considerably different.

Citation	Result
Erkan and Radev (2004a) (added features)	0.38304
Erkan and Radev (2004b)	0.3758
Alguliev et al. (2012)	0.3673
Hong et al. (2014)†	0.3641
re-run	0.3494

Table 5: Results for MEAD on DUC 2004 (MDS) data. † indicates that parameters have been reported in the publication.

SVM We retrain the SVM introduced by Sipos et al. (2012), following the guidelines provided[9]. This included all relevant packages and detailed instructions on how to train the SVM model, which data has been used and how the resulting model was applied to the data. Table 6 shows our results and the result reported in the original publication. We observe that the results are similar to each other and the confidence interval (CI) indicates, that the results do not significantly differ.

Sipos et al. (2012)	re-train & eval (95% CI)
0.4066	0.3995 (0.3883–0.4117)

Table 6: Results for Sipos et al. (2012) re-evaluation on DUC 2004 data.

Summary Data The final experiment builds on data introduced by Hong et al. (2014), which contains summaries for a range of methods.[10] The authors give the parameters used for evaluation and results for R-1, but also for ROUGE-2 (R-2) and ROUGE-4 (R-4). Table 7 shows the results as originally reported (O) and as replicated (R).[11] Comparing the results, we can see some differences and out of 36 values 22 do not match exactly (marked in italics). Out of these 22 only 8 differ by more than 0.01 points (marked in bold). For CLASSY 04 we see a difference of 0.04 in R-1 and for KL we see a difference of 0.03 in R-2.

System	R-1	R-2	R-4
LexRank (O)	35.95	7.47	0.82
LexRank (R)	**35.97**	**7.49**	0.82
Centroid (O)	36.41	7.97	1.21
Centroid (R)	36.41	*7.98*	1.21
FreqSum (O)	35.30	8.11	1.00
FreqSum (R)	35.30	*8.10*	*0.99*
TsSum (O)	35.88	8.15	1.03
TsSum (R)	*35.89*	8.15	1.03
KL (O)	37.98	8.53	1.26
KL (R)	**38.00**	**8.56**	1.26
CLASSY 04 (O)	37.62	8.96	1.51
CLASSY 04 (R)	**37.66**	*8.97*	1.51
CLASSY 11 (O)	37.22	9.20	1.48
CLASSY 11 (R)	**37.20**	*9.21*	1.48
Submodular (O)	39.18	9.35	1.39
Submodular (R)	*39.17*	*9.34*	*1.38*
DPP (O)	39.79	9.62	1.57
DPP (R)	**39.81**	*9.63*	*1.58*
RegSum (O)	38.57	9.75	1.60
RegSum (R)	*38.56*	9.75	*1.61*
OCCAMS_V (O)	38.50	9.76	1.33
OCCAMS_V (R)	38.50	9.76	*1.32*
ICSISumm (O)	38.41	9.78	1.73
ICSISumm (R)	38.41	**9.80**	1.73

Table 7: Original (O) and replicated (R) results for the data set published by (Hong et al., 2014).

[8] http://www.summarization.com/mead/

[9] Unfortunately, the link given in the original publication is not functional anymore.

[10] The link given in the original publication is still functional and provides the data set, as well as the recommended evaluation settings.

[11] Please note that for better comparison we adopt their notation.

4 Discussion

We looked into the question of whether the fact that all necessary research artifacts are available for specific benchmark data sets in automatic summarization allow for a straightforward evaluation and replication. We also looked into results reported in the literature, as often results are cited in subsequent works as baselines or for comparison.

We observed quite severe differences not only in the exact values obtained by running the evaluation, but also in the conclusions drawn from these with respect to the ranking of the system outputs.

We also observed that the results highly depend on the parameters used for evaluation. If **evaluation parameters** and **system output** results are given, results are reproducible, as we were able to show with the data and results presented by Hong et al. (2014). Using their data and the evaluation parameters, our results were almost identical to those reported in the original publication. As only some results differed, it remains open if the observed differences are due to changes on the hardware and/or software level. Also, not all three evaluation metrics differed. As most values were in the range of ±0.1 one assumption is, that this is due to differences in rounding. In order to evaluate this, a more detailed analysis of individual results is required. If the **method used to produce** the summaries has been described in enough detail, it is possible to achieve similar results as we did with work by Sipos et al. (2012).

Despite the seemingly ideal circumstances, we failed to reproduce the results for System 65 in DUC 2004. For the DUC 2002 task we were only partially able to replicate or reproduce results reported in the literature, despite similar circumstances. We could not reproduce results reported in the literature. Also our experiments with MEAD were not conclusive. They showed that depending on the parameters used for evaluation, the results can vary considerably, sometimes even significantly, even though the system implementation is available and the evaluation metric is known.

A closer look at the publications analyzed for this study, we found that only about 40% report the full set of evaluation parameters. Almost 50% of the publications did not mention the evaluation parameters at all.[12] Replicating or even reproducing results for these publications is therefore unnecessarily complicated and involves testing all possible combinations of parameters. As the correct parameter set is unknown in these cases, comparisons are as not as valuable as they could be. Additionally, re-implementations such as py-rouge[13] do not offer all the parameters ROUGE originally offered, making comparisons even harder. Therefore, one of our next steps is to re-evaluate the presented experiments using py-rouge.

More analysis, also in other areas of NLP would be beneficial to strengthen the results of this study. While ROUGE has quite an extensive parameter range, it is negligible compared to modern machine learning approaches and as has been pointed out by Crane (2018) they "often go unreported". Nevertheless, our results highlight a problem that will become more severe the more complicated the methods developed in NLP become: Disclosing all parameters used for creating and evaluating a specific system is crucial. Publishing the algorithms and the resulting data is not enough to ensure replicable results. And even having details about the evaluation procedure (including relevant parameters) does not ensure that results can be replicated and conclusions in line with previous work can be drawn. While this might sound trivial, our results indicate that this is not being done in enough detail to ensure replicability and reproducibility of results.

Acknowledgements

This work was supported by the research center for Digital Communication and Media Innovation (DKMI) of the University of Applied Sciences, Darmstadt.

References

Rasim M. Alguliev, Ramiz M. Aliguliyev, and Makrufa S. Hajirahimova. 2012. Gendocsum + mclr: Generic document summarization based on maximum coverage and less redundancy. *Expert Systems with Applications*, 39:12460–12473.

Araly Barrera and Rakesh Verma. 2011. Automated extractive single-document summarization: Beating the baselines with a new approach. In *Proceedings of the 2011 ACM Symposium on Applied Computing (SAC '11)*, pages 268–269, TaiChung, Taiwan.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of re-

[12] A detailed analysis of this would allow a more reliable quantification of this issue, not only in the context of automatic summarization.

[13] https://github.com/andersjo/pyrouge

producibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, and Khalid Choukri, editors. 2016. *4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*. An LREC 2016 Workshop, Portorož, Slovenia.

Kevin Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. 2016. Reproducibility in Natural Language Processing: A Case Study of two R Libraries for Mining PubMed/MEDLINE. In *4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6–12, Portorož, Slovenia. An LREC 2016 Workshop.

Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Günes Erkan and Dragomir R. Radev. 2004a. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Günes Erkan and Dragomir R. Radev. 2004b. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham. 2015. Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal. Association for Computational Linguistics.

Kai Hong, John M. Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland, 26–31 May 2014, pages 1608–1616. European Language Resources Association (ELRA).

Tobias Horsmann and Torsten Zesch. 2017. Do lstms really work so well for pos tagging? – a replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* Copenhagen, Denmark, 7–11 September 2017, pages 738–747. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out,* Barcelona, Spain, 2004, pages 74–81.

Elena Lloret and Manuel Palomar. 2010. Challenging issues of automatic summarization: Relevance detection and quality-based evaluation. *Informatica*, 34:29–35.

Sukanya Manna, Byron J. Gao, and Reed Coke. 2012. A subjective logic framework for multi-document summarization. In *Proceedings of the 24th International Conference on Computational Linguistics,* Mumbay, India, December 2012, pages 797–808.

Margot Mieskes. 2017. A quantative study of data in the nlp community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* Barcelona, Spain, 25–26 July 2004.

Rada Mihalcea and Paul Tarau. 2005. A language independent algorithm for single and multiple document summarization. In *The Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 19–24, Jeju Island, Korea.

Pengjie Ren, Furu Wei, Zhumin Chen, Jun Ma, and Ming Zhou. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 33–43, Osaka, Japan.

Cody Rioux, Sadid A. Hasan, and Yllias Chali. 2014. Fear the reaper: A system for automatic multi-document summarization with reinforcement learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, October 25-29, 2014, Doha, Qatar*, pages 681–690.

Seonggi Ryang and Takeshi Abekawa. 2012. Framework of automatic text summarization using reinforcement learning. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 256–265.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *Proceedings of the*

13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23–27, 2012, pages 224–233.

Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2017. Affinity-preserving random walk for multi-document summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* Copenhagen, Denmark, 7–11 September 2017, pages 210–220. Association for Computational Linguistics.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Wen-Tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-Document Summarization by Maximizing Informative Content-Words. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* Hyderabad, India, 6–12 January, 2007, pages 1776–1782.