On Evaluation of Automatically Generated Clinical Discharge Summaries

Hans Moen[1], Juho Heimonen[2,5], Laura-Maria Murtola[3,4], Antti Airola[2],
Tapio Pahikkala[2,5], Virpi Terävä[3,4], Riitta Danielsson-Ojala[3,4],
Tapio Salakoski[2,5], and Sanna Salanterä[3,4]

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology, Norway
[2] Department of Information Technology, University of Turku, Finland
[3] Department of Nursing Science, University of Turku, Finland
[4] Turku University Hospital, Finland
[5] Turku Centre for Computer Science, Finland
hans.moen@idi.ntnu.no
{juaheim,lmemur,ajairo,aatapa,vmater,rkdaoj}@utu.fi
{tapio.salakoski,sansala}@utu.fi

Abstract. Proper evaluation is crucial for developing high-quality computerized text summarization systems. In the clinical domain, the specialized information needs of the clinicians complicates the task of evaluating automatically produced clinical text summaries. In this paper we present and compare the results from both manual and automatic evaluation of computer-generated summaries. These are composed of sentence extracts from the free text in clinical daily notes – corresponding to individual care episodes, written by physicians concerning patient care. The purpose of this study is primarily to find out if there is a correlation between the conducted automatic evaluation and the manual evaluation. We analyze which of the automatic evaluation metrics correlates the most with the scores from the manual evaluation. The manual evaluation is performed by domain experts who follow an evaluation tool that we developed as a part of this study. As a result, we hope to get some insight into the reliability of the selected approach to automatic evaluation. Ultimately this study can help us in assessing the reliability of this evaluation approach, so that we can further develop the underlying summarization system. The evaluation results seem promising in that the ranking order of the various summarization methods, ranked by all the automatic evaluation metrics, correspond well with that of the manual evaluation. These preliminary results also indicate that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance.

Keywords: Summarization Evaluation, Text Summarization, Clinical Text Processing, NLP

Copyright ©2014 by the paper's authors. Copying permitted for private and academic purposes.

1 Introduction

With the large amount of information generated in health care organisations today, information overload is becoming an increasing problem for clinicians [1,2]. Much of the information that is generated in relation to care is stored in electronic health record (EHR) systems. The majority of this is free text – stored as clinical notes – written on a daily basis by clinicians about care of individual patients. The rest of the information contained in EHRs is mainly images and structured information, such as medication, coded information and lab values. Towards tackling the problems of information overload, there is a need for (EHR) systems that are able to automatically generate an overview, or summary, of the information in these health records - this applies to both free text and structured information. Such systems would enable clinicians to spend more time treating the patients, and less time reading up on information about the patients. However, in the process of developing such summarization systems, quick and reliable evaluation is crucial.

A typical situation where information overload is frequently encountered is when the attending physician is producing the discharge summary at the end of a care episode. Discharge summaries are an important part of the communication between different professionals providing the health care services and their aim to ensure the continuity of a patients care. However, there are challenges with these discharge summaries as they are often produced late, and the information they contain tend to be insufficient. For example, one study showed that discharge summaries exchanged between the hospital and the primary care physicians is often lacking information, such as diagnostic test results (lacking in 33-63%), treatment progression (lacking in 7-22%), medications (lacking in 2-40%), test results (lacking in 65%), counseling (lacking in 90-92%) and follow-up proposals (lacking in 2-43%) [3]. One reason for this is that, during discharge summary writing process, the physicians tend to simply not have the time to read everything that has been documented in the clinical daily notes. Another reason is the difficulty of identifying the most important information to include in the discharge summary.

Computer-assisted discharge summaries and standardized templates are measures for improving the transfer time and the quality of discharge information between the hospital and the primary care physicians [3]. Furthermore, computer-assisted discharge summary writing using automatic text summarization could improve the timeliness and quality of discharge summaries further. Another more general user scenario where text summarization would be useful is when clinicians need to get an overview of the documented content in a care episode, in particular in critical situations when this information is needed without delay.

Automatic summarization of clinical information is a challenging task because of the different data types, the domain specificity of the language, and

In: E.A.A. Jaatun, E. Brooks, K.E. Berntsen, H. Gilstad, M. G. Jaatun (eds.): Proceedings of the 2nd European Workshop on Practical Aspects of Health Informatics (PAHI 2014), Trondheim, Norway, 19-MAY-2014, published at http://ceur-ws.org

the special information needs of the clinicians [4]. Producing a comprehensive overview of the structured information is a rather trivial task [5]. However, that is not the case for the clinical notes and the free text they contain. Previously, Liu et al. [6] applied the automated text summarization methods of the MEAD system [7] to Finnish intensive care nursing narratives. In this work the produced summaries were automatically evaluated against corresponding discharge reports. The authors found that some of the considered methods outperformed the random baseline method, however, the authors noted that the results were overall quite disappointing, and that further work was needed in order to develop reliable evaluation methods for the task.

We have developed an extraction based text summarization system that attempts to automatically produce a textual summary of the free text contained in all the clinical (daily) notes related to a – possibly ongoing – care episode, written by physicians. *The focus of this paper is not on how the summarization system works, but rather on how to evaluate the summaries it produces.* In our ongoing work towards developing this system, we have so far seven different *summarization methods* to evaluate, including a `Random` method and an `Oracle` method. The latter method representing an upper bound for the automatic evaluation score. Having a way to quickly and automatically evaluate the summaries that these methods produce is critical during method development phase. Thus the focus of this paper is how to perform such automated evaluation in a reliable and cost-effective way.

Automatic evaluation of an extraction based summary is typically done through having a gold standard, or "gold summary", for comparison. A gold summary is typically an extraction based summary produced by human experts [8]. Then one measures the textual overlap, or similarity, between a targeted summary and the corresponding gold summary, using some metric for this purpose. However, we do not have such manually tailored gold summaries available. Instead we explore the use of the original physician-made discharge summaries for evaluation purposes as a means of overcoming this problem. These discharge summaries contain sentence extracts, and possibly slightly rewritten sentences, from the clinical notes. They also typically contain information that has never been documented earlier in the corresponding care episode, which makes them possibly suboptimal for the task of automatic evaluation.

To explore whether this approach to automatic evaluation is viable, we have also conducted *manual evaluation* of a set of summaries, and then compared this to the results from the automatic evaluation. A possible correlation between how the manual and automatic evaluation ranks the summarization methods would mean that further automatic evaluation with this approach can be considered somewhat reliable. In this study, automatic evaluation is mainly performed using the *ROUGE evaluation package* [9]. The manual evaluation was done by domain experts who followed an evaluation scheme/tool that we developed for this purpose. Figure 1 illustrates the evaluation experiment.

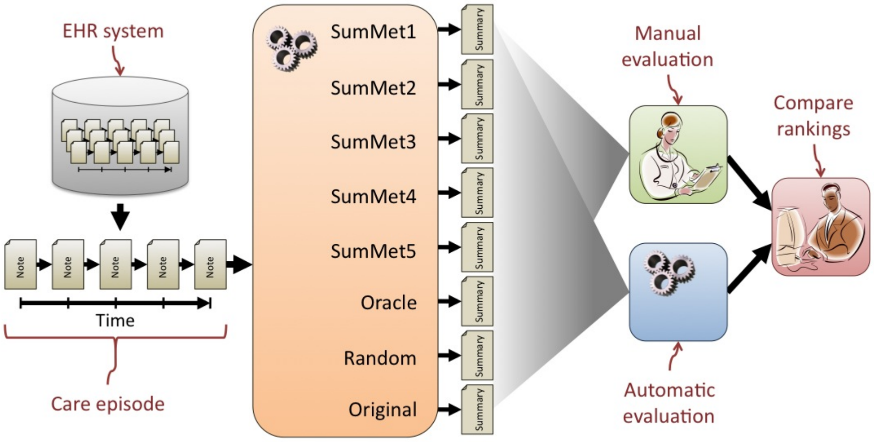

Fig. 1. Illustration of the evaluation experiment.

2 Data

The data set used in this study contained the electronic health records of approximately 26,000 patients admitted to a Finnish university hospital between the years 2005–2009 with any type of cardiac problem. An ethical statement (17.2.2009 §67) and the organisational permission (2/2009) from the hospital district was obtained before collection of this data set.

To produce data suited for automatic summarization, discharge summaries were extracted and associated to the daily notes they summarize. There were two types of discharge summaries: internal (written when the patient is moved to another ward and summarizing the time spent on the given ward) and final (written when the patient leaves the hospital and summarizing the whole stay). Note that a final discharge also summarizes any internal summaries written during the stay.

All notes and discharge summaries were lemmatized at the sentence level using the morphological analyser FinTWOL [10] and the disambiguator FinCG [11] by Lingsoft, Inc.[6]. Stopwords were also removed[7]. The preprocessed corpus contained 66,884 unique care episodes with 39 million words from a vocabulary of 0.6 million unique terms.

The full corpus was utilized in deriving statistics about the language for some of the summarization methods. For the summarization and evaluation experiment, the corpus was narrowed down to the care episodes having I25 (Chronic ischaemic heart disease; including its sub-codes) as the primary ICD-10 code and consisting of at least 8 clinical notes, including the discharge summary. The

[6] http://www.lingsoft.fi

[7] http://www.nettiapina.fi/finnish-stopword-list/

latter condition justifies the use of text summarization. The data were then split into the *training* and *test* sets, containing 159 and 156 care episodes, for the parameter optimization and evaluation of summarization methods, respectively.

3 Text Summarization

All summarization methods used in this study are based on *extraction-based* multi-document summarization. This means that each summary consist of a subset of the content in the original sentences, found in the various clinical notes that the summary is produced from [12]. This can be seen as a specialized type of multi-document summarization since each document, or clinical note, belong to the same patient, and together constitute a connected sequence. In the presented evaluation, seven different summarization methods are used, including `Random` and `Oracle`, resulting in seven different summaries per care episode. The original physician made discharge summary, `Original`, which is used as the *gold summary* for automatic evaluation, is also included in the manual evaluation. For the automatic evaluation, this gold summary is viewed as the perfect summary, thus having a perfect F-score (see Section 4.1). As stated earlier, the focus of this paper is on evaluating the text summaries produced by a summarization system. Thus the description of the utilized summarization system and the various methods used will be explained in more detail in a forthcoming extended version of this work. However, the two trivial control methods, `Random` and `Oracle`, deserves some explanation here.

Random This is the baseline method, which works by composing a summary through simply selecting sentences randomly from the various clinical notes. This method should give some indication of the difficultly level of the summarization task at hand.

Oracle This is a control-method that has access to the gold summary during the summarization process. It basically tries to optimize the ROUGE-N2 F-scores for the generated summary according to the gold summary, using a greedy search strategy. This summarization method can naturally not be used in a real user scenario, but it represents the upper limit for what is possible to achive in score for an extraction based summary, or summarization method, when using ROUGE for evaluation.

The other summarization methods are here referred to as `SumMet1`, `SumMet2`, `SumMet3`, `SumMet4` and `SumMet5`.

For each individual care episode, the length of the corresponding gold summary served as the length limit for all the seven generated summaries. This was mainly done so that a sensible automatic evaluation score (F-score, see Section 4.1) could be calculated. In a more realistic user scenario, a fixed length could be used, or e.g. a length that is based on how many clinical notes the summaries are generated from. Each computer generated summary is run through a post-processing step, where each sentence are sorted according to when they were written.

4 Evaluation Experiment

We conducted and compared two types of evaluation, *automatic evaluation* and *manual evaluation* in order to evaluate the different summarization methods. The purpose of this study is primarily to find out if there is a correlation between the conducted automatic evaluation and the manual evaluation. This will further reveal which of the automatic evaluation metrics that correlates the most with the scores from the manual evaluation. As a result, we would get some insight into the reliability of the selected approach to automatic evaluation. Ultimately this study can help us in assessing the reliability of this evaluation approach, so that we can further develop the underlying summarization system.

In the automatic evaluation we calculated the F-score for the overlap between the generated summaries and the corresponding gold summaries using the ROUGE evaluation package. As gold summaries we used the original discharge summary written by a physician. This gold summary is thus considered to be the optimal summary[8], so we assume that it to always have an F-score of 1.0.

The conducted evaluation can be classified as so called *intrinsic evaluation*. This means that the summaries are evaluated independently of how they potentially affect some external task [8].

4.1 Automatic Evaluation

ROUGE metrics, provided by the ROUGE evaluation package [9] (see e.g. [13]), are widely used as automatic performance measures in the text summarization literature. To limit the number of evaluations, we selected four common variants:

- **ROUGE-N1**. Unigram co-occurrence statistics.
- **ROUGE-N2**. Bigram co-occurrence statistics.
- **ROUGE-L**. Longest common sub-sequence co-occurrence statistics.
- **ROUGE-SU4**. Skip-bigram and unigram co-occurrence statistics.

These metrics are all based on finding word n-gram co-occurrences, or overlaps, between a) one or more *reference summaries*, i.e. gold summary, and b) the *candidate summary* to be evaluated.

Each metric counts the number of overlapping units (the counting method of which depends on the variant) and uses that information to calculate recall (R), precision (P), and F-score (F). The recall is the ratio of overlapping units to the total number of units in the reference while the precision is the ratio of overlapping units to the total number of units in the candidate. The former

[8] This is of course not always the truth from a clinical perspective, but we leave that to another discussion.

describes how well the candidate covers the reference and the latter describes the quality of the candidate. The F-score is then calculated as

$$F = \frac{2PR}{P+R} \ , \tag{1}$$

The evaluations were performed with the lemmatized texts with common stopwords and numbers removed.

4.2 Manual Evaluation

The manual evaluation was conducted independently by three domain experts. Each did a blinded evaluation of the eight summary types (seven machine generated ones plus the original discharge summary) for five care episodes. Hence, the total sum of evaluated summaries per evaluator was 40. All summaries were evaluated with a 30-item schema, or *evaluation tool* (see Table 1). This tool was constructed based on the content criteria guideline for medical discharge summaries, used in the region where the data was collected. So each item correspond to a criteria in this guideline. In this way, items were designed to evaluate the medical content of the summaries from the perspective of discharge summary writing. When evaluating a summary, each of these items were evaluated on a 4-class scale from -1 to 2, where, -1 = not relevant, 0 = not included, 1 = partially included and 2 = fully included. The evaluators also had all the corresponding clinical notes at their disposal when performing the evaluation.

The items in our tool are somewhat comparable to the evaluation criteria used in an earlier study of evaluating neonate's discharge summaries, where the computer generated discharge summaries using lists of diagnoses linked to ICD-codes [14]. However, the data summarized in the aforementioned work is mainly structured and pre-classified data, thus the summarization methods or performance is not comparable to our work.

The evaluators experienced the manual evaluations, following the 30-item tool, to be very difficult and extremely time consuming. This was mainly due to the evaluation tool, i.e. its items, being very detailed and required a lot of clinical judgement. Therefore, for this study, only five patient care episodes and their corresponding summaries, generated by the summarization system, were evaluated by all three evaluators. This number of evaluated summaries are too small for generalization of the results, but this should still give some indication of the quality of the various summarization methods in the summarization system. The 30 items in the manual evaluation tool are presented in Table 1.

4.3 Evaluation Statistics

In order to test whether the differences in the automatic evaluation scores between the different summarization methods were statistically significant, we performed the Wilcoxon signed-rank test [15] for each evaluation measure, and each pair of methods at significance level $p = 0.05$. We also calculated the p-values

Table 1. Items evaluated in the manual evaluation.

Evaluation criteria
Care period
Care place
Events (diagnoses/procedure codes) of care episode
Procedures of care episode
Long-term diagnoses
Reason for admission
Sender
Current diseases, which have impact on care solutions
Effects of current diseases, which have impact on care solutions
Current diseases, which have impact on medical treatment solutions
Effects of current diseases, which impact on medical treatment solutions
Course of the disease
Test results in chronological order with reasons
Test results in chronological order with consequences
Procedures in chronological order with reasons
Procedures in chronological order with consequences
Conclusions
Assessment of the future
Status of the disease at the end of the treatment period
Description of patient education
Ability to work
Medical certificates (including mention of content and duration)
Prepared or requested medical statements
A continued care plan
Home medication
Follow-up instructions
Indications for re-admission
Agreed follow-up treatments in the hospital district
Other disease, symptom or problem that requires further assessment
Information of responsible party for follow-up treatment

for manual evaluation. These were obtained with the paired Wilcoxon test computed over the 30 mean content criteria scores of the 30-item evaluation tool (see Table1). The mean scores were calculated by averaging the manually entered scores of the three evaluators and five care episodes. The -1 values indicating non-relevance were treated as missing values (i.e. they were ignored when calculating the averages). Also here a significance level of $p = 0.05$ was used. The agreement between the three evaluators was investigated by calculating the intraclass correlation coefficient (ICC) for all manually evaluated summaries using the two-way-mixed model.

To identify which of the automatic evaluation metrics that best follows the manual evaluation, Pearson product-moment correlation coefficient (PPMCC) and Spearman's rank correlation coefficient (Spearman's rho) [16] were calculated between the normalized manual evaluation scores and each of the automatic evaluation scores (from Table 2).

5 Evaluation Results

The results from the automatic and the manual evaluations are presented in Table 2. The scores from the automatic evaluation are calculated from the average F-scores from the 156 test care episodes, while the results from the manual eval-

Table 2. Evaluation results, each column are ranked internally by score.

Rank	**ROUGE-N1** F-score	**ROUGE-N2** F-score	**ROUGE-L** F-score	**ROUGE-SU4** F-score	**Manual** (norm_{max})
1	Original 1.0000	Original 1.0000	Original 1.0000	Original 1.0000	Original 1.0000
2	Oracle 0.7964	Oracle 0.7073	Oracle 0.7916	Oracle 0.6850	SumMet2 0.6738
3	SumMet2 0.6700	SumMet2 0.5922	SumMet2 0.6849	SumMet2 0.5841	Oracle 0.6616
4	SumMet5 0.5957	SumMet5 0.4838	SumMet5 0.5902	SumMet5 0.4723	SumMet5 0.6419
5	SumMet1 0.4785	SumMet1 0.3293	SumMet1 0.4717	SumMet1 0.3115	SumMet3 0.5326
6	SumMet4 0.3790	SumMet4 0.2363	SumMet4 0.3725	SumMet4 0.2297	SumMet1 0.5167
7	Random 0.3781	Random 0.2094	Random 0.3695	SumMet3 0.2013	Random 0.5161
8	SumMet3 0.3582	SumMet3 0.2041	SumMet3 0.3521	Random 0.2001	SumMet4 0.5016

uation are the average scores from a subset of five care episodes (also included in the automatic evaluation), all evaluated by three domain experts. The latter scores have all been normalized by dividing the scores of the highest ranking method. This was done in an attempt to scale these scores to the F-scores from the automatic evaluation.

All automatic metrics and the manual evaluation agreed in terms of what summarization method belongs to the top three, and the bottom four. When calculating significance levels for the automatic evaluations for the five highest ranked methods, the differences were always significant. However, in several cases the differences between the three lowest ranked methods, those being SumMet4, SumMet3 and Random, were not statistically significant. These results are in agreement with the fact that all the evaluation measures agreed on which the five best performing methods were, whereas the three worst methods are equally bad, all performing on a level that does not significantly differ from the approach of picking the sentences for the summary randomly.

For the manual evaluation, the original discharge summaries scored significantly higher than any of the generated summaries. Furthermore, the summaries produced by the Oracle, SumMet2 and SumMet5 methods were evaluated to be significantly better than those produced by the four other methods. Thus, the manual evaluation divided the automated summarization methods into two distinct groups, one group that produced seemingly meaningful summaries, and the other that did not work significantly better than the Random method. The divi-

[8] The original discharge summary is of course not a product of any of the summarization methods.

Table 3. PPMCC and Spearman's rho results, indicating how the scoring by the automatic evaluation metrics correlates with the normalized manual evaluation scores.

Evaluation metric	PPMCC (p-values)	Spearman's rho (p-values)
ROUGE-N1	0.9293 (0.00083)	0.8095 (0.01490)
ROUGE-N2	0.9435 (0.00043)	0.8095 (0.01490)
ROUGE-L	0.9291 (0.00084)	0.8095 (0.01490)
ROUGE-SU4	**0.9510** (0.00028)	**0.8571** (0.00653)

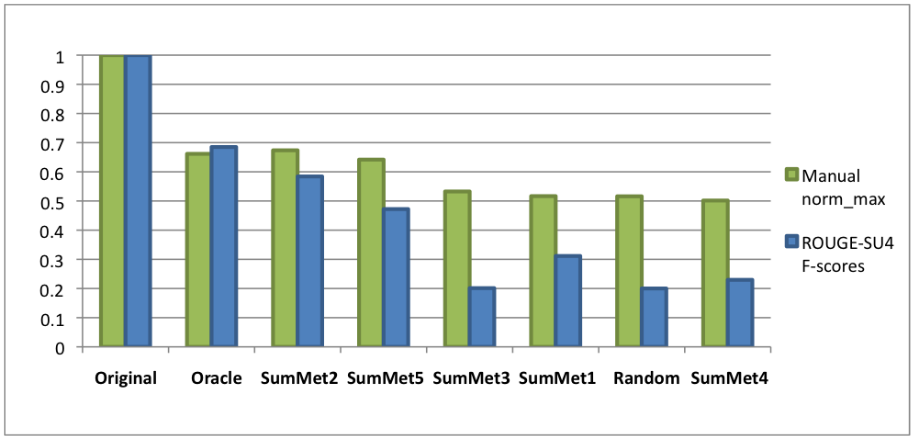

Fig. 2. Graph showing the evarage manual scores (norm$_{max}$), calculated from five care episodes (evaluated by three domain experts), and average F-scores by ROUGE-SU4, calculated from 156 care episodes. The order, from left to right, is sorted according to descending manual scores.

sion closely agrees with that of the automated evaluations, the difference being that in the manual evaluation also `SumMet1` ended up in the bottom group of badly performing summarization methods.

In Table 3 are the PPMCC and Spearman's rho results, indicating how each automatic evaluation metric correlates with the manual evaluation scores. Spearmans rho is a rank-correlation measure, so it does not find any difference between most of the measures, since they rank the methods in exactly the same order (except ROUGE-SU4, which has a single rank difference compared to others). In contrast, PPMCC measures the linear dependence taking into account magnitudes of the scores in addition to the ranks, so it observes some extra differences between the measures. This shows that ROUGE-SU4 has the best correlation compared to the manual evaluation. Figure 2 illustrates the normalized manual evaluation scores with the ROUGE-SU4 F-scores.

6 Discussion

All automatic evaluation metrics and the manual evaluation agreed that the top three automatic summarization methods significantly outperform the `Random` method. These methods are `SumMet2`, `SumMet5` and `Oracle`. Thus we can with a certain confidence assume that this reflects the actual performance of the utilized summarization methods. `Oracle` is of course not a proper method, given that it is "cheating", but it is a good indicator for what the upper performance limit it.

The reliability of the manual evaluation is naturally rather weak, given that only five care episodes were evaluated by the three evaluators. The findings of the manual evaluation are not generalizable due to the small number of care episodes evaluated. Therefore, more manual evaluation is needed to confirm these findings.

On a more general level, the results indicate that the utilized approach – using the original discharge summaries as gold summaries – is a seemingly viable approach. This also means that the same evaluation framework can potentially be transferred to clinical text in other countries and languages who follow a similar hospital practice as in Finland.

The manual evaluation results show that the various summarization methods are less discriminated in terms of scores when compared to the automatic evaluation scores. We believe that this is partly to blame for the small evaluation set these scores are based on, and also because of the evaluation tool that was utilized. For these reasons we are still looking into ways to improve the manual evaluation tool before we conduct further manual evaluation. It is interesting to see that `SumMet2` is considered to outperform the `Oracle` method, according to the manual evaluators.

6.1 Lessons learned from the manual evaluation

The agreement between the evaluators in the manual evaluation was calculated with the 40 different summaries evaluated by each of the three evaluators. The ICC value for the absolute agreement was 0,68 (95% CI 0,247-0,853, p<0,001). There is no definite limit in the literature on how to interpret ICC values, but there are guidelines that suggest that values below 0.4 are poor, values from 0.4 to 0.59 are fair, values from 0.6 to 0.74 are good and values from 0.75 to 1.0 are excellent in terms of the level of interrater agreement [17]. The agreement between the evaluators in the manual evaluation was good, based on these suggested limits. This means that there were some differences between the evaluations conducted by the three evaluators, which indicates that the criteria used in the 30-item manual evaluation tool allowed this variance, and therefore, the tool with its items need further development. Another aspect is that the evaluators would need more training concerning the use of the criteria and possibly more strict guidelines.

Furthermore, the evaluators reported that the manual evaluation was difficult and very time consuming, due to the numerous and detailed items in the manual evaluation tool. They also reported that the judgement process necessary when evaluating the summaries was too demanding. It became obvious that several of the items in the evaluation tool were too specifically targeting structured information. This means information that is already identified and classified in the health record system, which does not need to be present in the unstructured free text from where the summaries are generated. Examples are 'Care period', 'Care place' and 'Sender'. In the future, a shorter tool, i.e. less items, with stricter criteria and more detailed guidelines for the evaluators is needed. One important property of such a new tool would be, when used by the human evaluators, that good and bad summaries (i.e. summarization methods) are properly discriminated in terms of scoring.

7 Conclusion and Future Work

In this paper we have presented the results from automatic and manual evaluation of seven different methods for automatically generating clinical text summaries. The summary documents was composed from the free text of the clinical daily notes written by physicians related to patient care.

Seven automatic summarization methods were evaluated. For the automatic evaluation the corresponding original discharge summaries were used as gold summaries for doing the automatic evaluation. Among these summarization methods were the control-methods `Random` and `Oracle`. Four ROUGE metrics were used for the automatic evaluation, ROUGE-N1, ROUGE-N2, ROUGE-L and ROUGE-SU4.

The evaluation results seem promising in that the ranking order of the various summarization methods, ranked by all the automatic evaluation metrics, correspond well with that of the manual evaluation. These preliminary results indicates that the utilized automatic evaluation setup can be used as an automated and reliable way to rank clinical summarization methods internally in terms of their performance.

More manual evaluation, on a larger sample of care episodes, is needed to confirm the findings in this study. In this context, more research is needed to make a manual evaluation tool that better discriminates good from bad summaries, as well as being easier to use by evaluators. This preliminary work provided us good insight and ideas about how to further develop the manual evaluation tool, suited for a larger-scale manual evaluation.

As future work, we also plan to conduct so called *extrinsic evaluation* of the summarization methods, meaning that the various summaries produced by the system are evaluated in terms of their impact on clinical work.

8 Acknowledgements

This study was financially supported partly by the Research Council of Norway through the EviCare project (NFR project no. 193022), the Turku University Hospital (EVO 2014), and the Academy of Finland (project no. 140323). The

study is a part of the research projects of the Ikitik consortium (`http://www.ikitik.fi`).

References

1. Hall, A., Walton, G.: Information overload within the health care system: a literature review. Health Information & Libraries Journal **21**(2) (2004) 102–108
2. Van Vleck, T.T., Stein, D.M., Stetson, P.D., Johnson, S.B.: Assessing data relevance for automated generation of a clinical summary. In: AMIA Annual Symposium Proceedings. Volume 2007., American Medical Informatics Association (2007) 761
3. Kripalani, S., LeFevre, F., Phillips, C.O., Williams, M.V., Basaviah, P., Baker, D.W.: Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. Jama **297**(8) (2007) 831–841
4. Feblowitz, J.C., Wright, A., Singh, H., Samal, L., Sittig, D.F.: Summarization of clinical information: A conceptual model. Journal of biomedical informatics **44**(4) (2011) 688–699
5. Roque, F.S., Slaughter, L., Tkatšenko, A.: A comparison of several key information visualization systems for secondary use of electronic health record content. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, Association for Computational Linguistics (2010) 76–83
6. Liu, S.: Experiences and reflections on text summarization tools. International Journal of Computational Intelligence Systems **2**(3) (2009) 202218
7. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic Summarization. Volume 4 of NAACL-ANLP-AutoSum '00., Association for Computational Linguistics (2000) 21–30
8. Afantenos, S., Karkaletsis, V., Stamatopoulos, P.: Summarization from medical documents: a survey. Artificial intelligence in medicine **33**(2) (2005) 157–177
9. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens, S.S., ed.: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain, Association for Computational Linguistics (July 2004) 74–81
10. Koskenniemi, K.: Two-level model for morphological analysis. In Bundy, A., ed.: Proceedings of the 8th International Joint Conference on Artificial Intelligence. Karlsruhe, FRG, August 1983, William Kaufmann (1983) 683–685
11. Karlsson, F.: Constraint grammar as a framework for parsing running text. In: Proceedings of the 13th Conference on Computational Linguistics - Volume 3. COLING '90, Stroudsburg, PA, USA, Association for Computational Linguistics (1990) 168–173
12. Nenkova, A., McKeown, K.: Automatic summarization. Foundations and Trends in Information Retrieval **5**(23) (2011) 103–233
13. Dang, H.T., Owczarzak, K.: Overview of the tac 2008 update summarization task. In: Proceedings of text analysis conference. (2008) 1–16
14. Lissauer, T., Paterson, C., Simons, A., Beard, R.: Evaluation of computer generated neonatal discharge summaries. Archives of disease in childhood **66**(4 Spec No) (1991) 433–436

15. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics **1** (1945) 80–83
16. Lehman, A.: JMP for basic univariate and multivariate statistics: a step-by-step guide. SAS Institute (2005)
17. Cicchetti, D.V.: Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment **6**(4) (1994) 284