Medical Dialogue Summarization for Automated Reporting in Healthcare

Sabine Molenaar(✉), Lientje Maas, Verónica Burriel, Fabiano Dalpiaz, and Sjaak Brinkkemper

Department of Information and Computing Sciences,
Utrecht University, Utrecht, The Netherlands
{s.molenaar,j.a.m.maas,v.burriel,f.dalpiaz,s.brinkkemper}@uu.nl

Abstract. Healthcare providers generally spend excessive time on administrative tasks at the expense of direct patient care. The emergence of new artificial intelligence and natural language processing technologies gives rise to innovations that could relieve them of this burden. In this paper, we present a pipeline structure for building dialogue summarization systems. Our pipeline summarizes a consultation of a patient with a care provider and automatically generates a report compliant with medical formats. Four pipeline components are used to generate a report based on audio input. The outputs of each component are analyzed to determine the most important challenges and issues. The current proof-of-concept, which was applied to eight doctor-to-patient sessions concerning ear infection, shows that automatic dialogue summarization and reporting is achievable, but requires improvements to increase completeness.

Keywords: Dialogue summarization · Automated reporting · Natural language processing · Artificial intelligence · Healthcare

1 Introduction

The introduction of the Electronic Medical Record (EMR) was intended to improve the communication among care providers within and between healthcare institutions. The EMR contains information about patients such as medical history, vital signs and medication among others. In addition, the EMR demands guideline adherence and may, in some uses, provide decision support [7].

While the EMR aims to improve patient care, this may not always be the case. Administrative burden in healthcare is a well-known problem, especially in general practice, psychiatric care, and trauma surgery [12,28]. In the US, a first year resident spends more time with the EMR than with patients [8].

As a solution to these problems, the Care2Report project strives for automated reporting in healthcare [18]. The goal is to automatically generate medical reports of patient-doctor dialogues in compliance with clinical guidelines and

S. Dupuy-Chessa and H. A. Proper (Eds.): CAiSE 2020 Workshops, LNBIP 382, pp. 76–88, 2020.
https://doi.org/10.1007/978-3-030-49165-9_7

without disrupting the current way of working. This is the research framework within which we position this paper.

Starting from the vision and overall architecture of Care2Report [18], we focus here on a detailed study of the dialogue summarization pipeline, which aims to support speech and text processing in healthcare by combining computational linguistics and AI techniques. After a brief description of the pipeline structure, we investigate difference facets of *quality*. We make the following contributions:

1. We study how quality in the pipeline can be measured and which threats can affect quality;
2. We evaluate the quality of the pipeline and its components by analyzing eight reports generated by the proof-of-concept;
3. We identify which threats have affected the quality to provide a basis for further improvement of the pipeline.

The paper is structured as follows. Sect. 2 describes related work. We present the dialogue summarization pipeline in Sect. 3. We describe metrics of and threats to quality in Sect. 4. We report on an analysis of eight medical consultations in Sect. 5. Finally, we present limitations and outline future work in Sect. 6.

2 Related Works

An extensive study on the effect of the EMR on doctor-patient communication [2] revealed several benefits, such as improved understanding by the patient and a positive communication experience with the EMR. However, several concerns were identified, both from the perspective of patients and doctors. In case of the former, patients expressed worries about the doctor potentially getting distracted by the computer during the appointment. The latter mentioned not being able to tend to the patient while interacting with the computer at the same time. In addition, it is reported that doctors spent an estimated 32% of the appointment interacting with the computer (based on an average of six studies). In three studies, patients were found to stop talking whenever the doctor was typing [2]. Another issue is the potential loss of emotional and/or psychosocial elements. Non-verbal communication (e.g., eye contact) is important for sharing emotions between patient and physician and such information may be overlooked if the physician is interacting with the EMR [22]. In summary, challenges arise when medical staff needs to interact with the EMR during direct patient care.

In the healthcare domain, various attempts have been made to automatically generate documents concerned with patient data. Firstly, speech recognition is a prominent approach to reducing time spent on reporting in healthcare, as studies frequently make use of dictating after a consultation [1]. In the Netherlands, however, a mere 1% of medical staff makes use of speech recognition. Reasons for the lack of adoption are, reportedly, interference with doctors' normal way of working, lack of support by hospitals and financial limitations [17]. Secondly,

Klann & Szolovits delivered a proof-of-concept framework that captures the dialogue during a doctor-patient meeting. Their approach covers the entire dialogue, rather than a report of the consultation [14]. More recently, Chiu *et al.* developed and tested a system that transcribes conversations between doctors and patients. Their best model resulted in a word error rate of 18.3% [9]. Again, this system delivers a medical transcription, rather than an EMR update or a report. Finally, the BabyTalk project utilizes a prototype that generates summaries in text, using physiological signals and events performed by medical staff as input. While the prototype proved to be able to generate proper summaries of clinical data, the texts provided by human experts were still superior [21].

Jiang *et al.* [13] discuss the use of AI in healthcare and conclude that both linguistics (through NLP) and AI (via Machine Learning, ML) are used to enrich medical data. In their study, NLP uses human language notes as input and returns a structured version of these notes for the EMR. Then, the EMR data feeds ML algorithms. The summarization pipeline aims to combine the two rather than execute them sequentially: both NLP and ML are used to enrich the data stored in the EMR. Without taking precautions, neural networks may overlook rare outcomes, due to the under-representation of these outcomes in training data [26]. Another possible disadvantage of using (deep) neural networks is that they often lack transparency, which limits their use in the healthcare domain [15].

3 Dialogue Summarization Pipeline

We use the term *dialogue summarization pipeline* to refer to the set of software components required to generate reports using audio input [18]. We define a pipeline as a series of (NLP- and/or AI-enabled) computational components, which transform output from one system into input for another system.

The pipeline combines AI and computational linguistics algorithms to automatically generate reports through the components shown in Fig. 1. Example outputs generated by the components using real-world input (from patient-General Practitioner (GP) consultations) are depicted in boxes with dashed arrows (note that the original audio input was in Dutch and was translated for the purpose of this example). First, a **speech transcription** is made using the audio of the dialogue as input. Subsequently, the **triple extraction** component extracts semantic triples from the transcription. Semantic triples consist of subjects, predicates and objects respectively [4]. A domain-specific example of a triple is: $\langle Ear, hasSymptom, Pain \rangle$. In a separate component, triples are utilized for **ontology population**. The ontology contains domain-specific information, such as clinical guidelines and standards. Once the ontology contains the guidelines for a specific illness, it does not need to be populated for this illness again. Ideally, the ontology will contain all the clinical guidelines and will only be modified if changes to the original guidelines are made.

Thirdly, triples are selected in the **triple matching** component. Extracted triples are selected if they match triples in the ontology. For instance, in Fig. 1, one of the triples is $\langle Patient, has, Earache \rangle$, which can be matched to part of

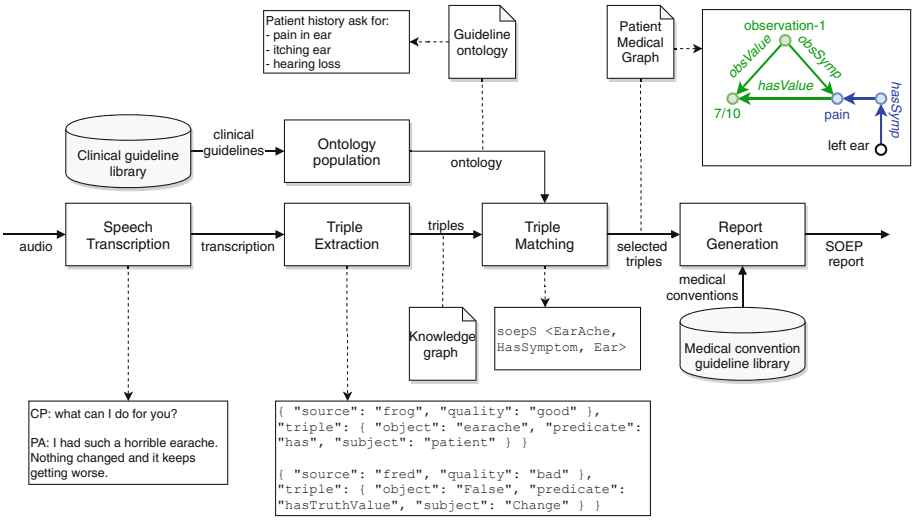

Fig. 1. Dialogue summarization pipeline for reporting in healthcare.

the guidelines shown at the top "*Patient history ask for: pain in ear*". All the triples that are matched in this manner are selected to be included in the report and are stored in a graph. This graph contains an overview of the patient's symptoms, the findings by the GP, the diagnosis and the treatment.

To avoid affecting the way of working of care providers, the report is generated in compliance with medical conventions. For example, GPs in the Netherlands use the SOEP (or SOAP) format, which defines four sections for reporting on a consultation: Subjective (S), Objective (O), Evaluation (E) (or Assessment (A)) and Plan (P) [6]. When the triples are matched, they are also categorized, to ensure they are included in the correct place in the format or report. Finally, the categorized triples are transformed back into natural language by the **report generation** component, to make them easier to read and understand.

4 Quality in the Dialogue Summarization Pipeline

We first describe the quality metrics that can be used for assessing the performance of individual components of a pipeline or of the pipeline as a whole in Sect. 4.1. Then, we discuss how to cope with threats that affect the performance of the dialogue summarization pipeline in Sect. 4.2.

4.1 Measuring Quality in a Pipeline

A pipeline consists of multiple components that are sequentially connected: $(c_1, \ldots, c_n)$. Since each component is an imperfect data processor, the outcomes

of that component will contain error. Intuitively, at each step, additional error is potentially introduced, thereby affecting the quality of the pipeline.

We define quality in terms of information retrieval metrics such as precision, recall, F_β-score, etc. We use the generic term *quality* to refer to one of these metrics, or a combination of them. The choice of the specific quality metrics is domain specific. Given a pipeline $(c_1, \ldots, c_n)$, an input i_1 and a ground truth output gt_n, the **pipeline quality** can be measured by feeding i_1 to c_1, executing in sequence all the components until c_n, and comparing the output against gt_n.

Pipeline quality can be used to assess the quality of a sub-sequence of components $(c_i, \ldots, c_k)$ with $i \geq 1$ and $k \leq n$, by feeding gt_{i-1} as input to c_i, running the sub-pipeline till c_k, and by comparing the output against gt_k. Brought to the extreme, given a sequence of a single component (c_i), we can feed gt_{i-1} to c_i and compare the output against gt_i to measure the **component quality**.

Inspired by Valls-Vargas *et al.* [27], we consider the notion of **error propagation**: given a sub-pipeline $(c_i, \ldots, c_k)$, we feed gt_{i-1} to c_i and run the entire pipeline to obtain some output o', then feed gt_{k-1} to c_k and run that component to obtain another output o'', and consider the quality difference between o'' and o'. Such a difference denotes how much error in c_k is introduced by c_i.

The measures of pipeline quality, component quality and error propagation provide a comprehensive picture of how a pipeline performs, also how the earlier components of a pipeline affect the performance of a later component.

4.2 Handling with Quality Threats

Table 1 lists the major threats that affect the quality of a pipeline's processing. The list is not meant to be exhaustive, but it rather serves as a summary of the challenges to consider in order to maximize quality.

The **speech transcription** component suffers from well known challenges in automated speech recognition [10]. For example, background noise makes it harder for the algorithms to distinguish the voices of the participants in the conversation. Moreover, when multiple voices participate in the conversation, they have to be distinguished [25]. Out-of-vocabulary words pose a challenge, as the algorithm is unable to match the recorded sound to a word within its prior knowledge. Finally, in the context of dialogue summarization, it is reasonable to expect that the people in the conversation may have accents or employ dialect.

Regarding **triple extraction**, the existing challenges arise from computational linguistics. A common problem refers to non-trivial sentence fragments such as compound nouns and phrasal verbs. For example, a phrasal verb such as 'to search for' it not easy to map to a triple; take the sentence 'I was searching my cupboard for my pills' could result either in a triple $\langle I, search, Cupboard \rangle$ or in a triple $\langle I, searchFor, Pills \rangle$. Furthermore, coreference resolution is a well-known issue [19], which refers to how pronouns can be linked to the noun they refer to. The use of pronouns such as 'it' or 'which' are common in medical consultations. Conflicting statements are also a hard to tackle; consider the following dialogue: (i) [patient] 'my hand hurts'; (ii) [doctor] 'you are indicating your finger, so I

Table 1. Overview of the quality threats within the dialogue summarization pipeline.

Component	Threat to quality
Speech transcription	T1: Background noise
	T2: Multiple voices in the conversation
	T3: Out-of-vocabulary words
	T4: Accent, dialect, and spontaneous speech
Triple extraction	T5: Compound nouns and phrasal verbs
	T6: Coreference resolution
	T7: Conflicting statements
Triple matching	T8: Incompleteness of ontology
	T9: Omission of relevant information out of context
	T10: Different categorization of measuring values
	T11: Redundancy of information (synonyms detection)
Report generation	T12: Erroneous triple categorization (SOEP)
	T13: Wrong positioning of information in text

suppose your finger hurts' – in this case, only the doctor's statement should be considered for triplification, but it is hard to reliably do so.

The quality of **triple matching** could also be affected by several threats. One of them is the incompleteness of the ontology. The building and population of the ontology is a component described in [18] that only needs to be done during the development phase of the system and could involve some threats to the triple matching component. If this ontology is incomplete or has been populated for a very specific medical specialization, some concepts could be missing, and therefore the triple matching process could discard relevant information. Also, some relevant information to be included in the final report could be discarded if it belongs to a non-medical domain and it is not included in the ontology or it is not mentioned in the clinical guidelines. For instance, some experience lived by the patient, which could look like an anecdote to omit in the report, but maybe it is the precursor of the disease. Another threat to be taken into account is the way of measuring some values. For example, pain can be measured from 1 to 10, or from soft to strong, or from light to hard, etc. Despite these different categories, *1, soft* and *light* are representing the same level of pain and this could cause errors in the categorization if this is not taken into account during the matching process. Also, some information can be repeated by the care provider or the patient, even using synonyms. This redundancy should be detected while doing the matching to avoid redundancy in the generated report.

During **report generation**, some challenges arise to ensure the quality of the final report. One important part of this component is the categorization of the triples according to the SOEP convention. If a triple is categorized in a wrong section, not only the information of this triple is missing in the corresponding section, but the correctness of the wrongly assigned section is also

affected. Once this categorization is done, the Natural Language Generation process starts. During this process, the relevant triples are ordered (text planning), each triple is converted into a standalone sentence (lexicalization) and some of these sentences are merged into longer ones (aggregation) [18]. Assuring a correct order of the triples in the text planning and aggregation phases is essential to avoid incorrect meanings and interpretations of the information collected during the consultation, which could cause erroneous statements in the final report.

5 Analysis of Automatically Generated Reports

We build a proof-of-concept implementation of our pipeline. In the current implementation, we rely on Google Speech for the speech-to-text transcription. For the triple extraction component, three different triple analyzers are used: Frog [5], FRED [11], Ollie [23]. The former supports Dutch, while the other two only support English, for these the transcriptions are translated from Dutch to English using Google Translate. The ontology was populated using clinical standards[1] that GPs in the Netherlands use to examine and diagnose patients. Matched triples are stored in a custom component, the patient medical graph [18]. Triples are transformed into natural language sentences using an extension of NaturalOWL [3] that supports the Dutch language.

The pipeline was tested on eight real-world consultations concerning external and middle ear infection. Transcriptions of the consultations were provided as input for the system, resulting in automatically generated medical reports, according to the SOEP convention. On average, the reports consisted of 1,132 words, ranging between 568 words (R-4) and 1,767 words (R-6).

We assessed the quality of the reports and of the intermediate results of the components in the pipeline (Fig. 1). Four out of five results of components in the pipeline are analyzed: transcription, triples, selected triples, and the SOEP report. If items are missing in the generated reports, we attempt to determine where these 'went missing' by tracing backward through the intermediate results.

Golden Standard. The SOEP format does not include any metrics with which to measure report quality and completeness. In addition, to the best of our knowledge, medical professionals in the Netherlands do not receive any formal training on how to write such formatted reports. Therefore, in order to determine the quality of the generated reports, we will rely on consultation reports, written by a GP in the SOEP format, which we use as a golden standard [16].

5.1 Report Quality

The eight generated reports were compared to the golden standard. We measure pipeline quality using three metrics: precision, recall and false positives (FPs). The quality of the reports was assessed according to the following process. Firstly, the number of items included in the generated and golden standards

[1] https://www.nhg.org/nhg-standaarden.

was established. Since GPs do not always include full sentences, but also partial phrases or even just words, the term 'item' needs to be defined. In this case, an item can be defined as a word or a sequence of words. Items are separated from each other using conjunctions (e.g., "and") and/or punctuation (e.g., periods, commas). Secondly, precision and recall were calculated. Thirdly, the number of false positives was determined, by counting items that were included in the generated report and items that are incorrectly included (i.e., partial items). An example of the latter is when the generated Plan includes "*paracetamol*", while the golden standard explicitly states "*no paracetamol*".

The number of items included for each section of the SOEP format by the generated (*R-x*) and golden standard (*S-x*) of the consultations (*C-x*) are shown in Table 2. For the generated reports the number of true positives (TPs), FPs and false negatives (FNs) are also shown, respectively.

Table 2. Number of items included for each section of the SOEP format, with TPs/FPs/FNs for the generated reports.

	C-1		C-2		C-3		C-4		C-5		C-6		C-7		C-8	
	R-1	S-1	R-2	S-2	R-3	S-3	R-4	S-4	R-5	S-5	R-6	S-6	R-7	S-7	R-8	S-8
S	1/0/3	4	1/0/8	9	1/0/4	5	0/0/4	4	0/0/4	4	0/0/9	9	0/0/8	8	0/0/8	8
O	0/0/2	2	0/1/6	6	2/0/1	3	1/1/8	9	0/0/4	4	0/0/4	4	1/0/2	3	1/0/1	2
E	0/0/1	1	0/0/2	2	1/0/0	1	1/0/1	2	0/0/2	2	0/0/1	1	0/0/1	1	0/0/1	1
P	0/2/1	1	3/1/5	8	1/1/1	2	1/0/2	3	1/0/5	6	0/2/5	5	0/1/2	2	2/0/0	2

It is apparent that the golden standards tend to consist of more items than the generated reports. The precision, recall, F1-score and number of FPs of each of the generated reports are presented in Table 3.

Table 3. Analysis of relevance and completeness of generated reports using the precision, recall and F-measure.

	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	μ
Precision	0.333	0.667	0.833	0.750	1.000	0.000	0.500	1.000	0.635
Recall	0.125	0.160	0.455	0.167	0.063	0.000	0.071	0.231	0.159
F1-score	0.182	0.258	0.588	0.273	0.112	0.000	0.125	0.375	0.239
FPs	2	2	1	1	0	2	1	0	1.125

Since we consider relevance (precision) and completeness (recall) to be equally important, the F1-score is calculated by assigning equal weight to both. Six out of eight reports achieved a precision score of half or higher, with the average being 0.635, meaning that the majority of the selected items are relevant. The recall score, however, is much lower. On average, 15.9% of the items that are

considered relevant are included, meaning that the majority of relevant items is missing. The Subjective section lacks most items, even though this section often includes a high frequency of items when written by a GP. The reason may be T9, i.e., the omission of out-of-context information (see Sect. 4).

Finally, nearly all reports include FPs. While some of these items are harmless, for example if they provide additional information that is not required, others can lead to inaccurate reporting. In *R-2*, "*antibiotics*" was included as an item, while the golden standard explicitly stated "*in consultation with patient no antibiotics*". **Challenge 1:** The system should be able to recognize negations, in order to provide correct information.

5.2 Pipeline Analysis

We analyze the quality of the outputs of two automated components in the pipeline: triple extraction and triple matching. We omit the transcription component because, in our case, this activity was done manually. We also use the intermediate results to locate where the missing item in the report were lost.

Triple Extraction. Triples are extracted from text using three triple analyzers: FRED (78.3% of extracted triples), Ollie (15.3%) and Frog (6.5%). All extracted triples are checked for their quality and labeled as `good` or `bad` (see Fig. 1). If a triple contains an item of excessive length (e.g., a full sentence instead of one or a few words) and/or does not contain a reference to either the doctor or patient, the triple is given the label `bad`. Only between 0.9% and 2.1% of the triples are labeled `good` (on average 21.1 triples out of 1,367 based on eight reports).

Triple Matching. All matched triples end up in the generated report, however, not all matched triples receive the correct categorization according to the SOEP format (see T12 in Sect. 4). The only erroneous categorization is within the Plan section. Medication mentioned during the consultation is always assigned to the Plan, while sometimes it is part of the Subjective. This part is in essence 'the patient's side of the story' and may also include previously used medication. **Challenge 2:** The location of the item in the transcription can help mitigate these errors, by separating the Subjective and the Plan. In addition, the tense of verbs can be used, since they distinguish past medications from future ones.

Locating Missing Report Items. When compared to the golden standards, the generated reports lacked 106 items. To determine at which point in the pipeline these items were lost (excluded or not identified), each missing item was traced back from the report to the transcript. Out of 106 items, only eleven could not be found in any of the intermediate results in the pipeline, as shown in Table 4. Note that the table does not show how many items were found, but the number of items that are still not identified in the particular output of the pipeline. Percentages are shown as portion of the total number of missing items.

All items that we found in the extracted triples (and yet were not included) had low quality. The analysis controller assesses the quality of the triples and distinguishes `good` and `bad` triples. 50.9% of the missing items were found in

Table 4. Overview of which missing items could not be located in the pipeline.

	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	%	μ
Missing items (total)	7	21	6	15	15	19	13	10	100.0%	13.3
In extracted triples	4	7	2	6	11	9	8	5	49.1%	6.5
In transcription (explicit)	0	3	1	2	4	4	1	0	14.2%	1.9
In transcription (fully)	1	0	0	3	3	2	2	0	10.4%	1.4

the extracted triples, but were all of bad quality. **Challenge 3:** Increasing the number of relevant items requires improving the quality of the triples.

A small portion of the items (14.2%) could only be located implicitly: they can be inferred from the transcript, but are not explicitly mentioned. An example of an implicit item is an explanation or recommendation given by the care provider to the patient. The GP does not announce that they will explain something, but this can be observed when reading the text. Another example is that the GP may state they will perform some medical action on the patient tomorrow, which means that the patient will return for another appointment, but this second appointment is not made explicit. **Challenge 4:** Conversation analysis techniques can be utilized to extract implicit information from transcripts.

The items that could not be located in any of the intermediate results in the pipeline are either observations or decisions made by the GP or gestures. Examples of observations are: (1) whether the eardrum is visible; this is part of the clinical guideline for ear infection, but the doctor makes no utterance about it, (2) if the ear canal is red, which only sometimes is mentioned out loud. Decisions made by the GP mostly refer to the diagnosis and plan. Based on the observations, they conclude a diagnosis which may or may not be communicated to the patient explicitly. Furthermore, the plan is discussed with the patient, but not in full detail. For instance, the GP will explain the patient receives ear drops, but does not specify how many of these drops they should use per day, while this is included in the report. The system, however, does allow the GP to add, modify or remove text in the report if necessary. Finally, it is hard to ascribe gestures to the GP or the patient (see T6 in Sect. 4). Oftentimes, patient mention "*I don't hear anything on this side*" or "*do you want to see the other ear as well*", in combination with pointing the GP is able to tell which ear they are referring to, but based on text alone the system cannot distinguish between left and right. **Challenge 5:** Video input can be used to enhance text that includes a gesture by disambiguating the antecedent to reference pronouns such as "this".

6 Discussion and Future Work

In this paper, we introduced a dialogue summarization pipeline as a series of components that can generate a report of a conversation. The quality of the pipeline was evaluated through eight real medical consultations regarding ear infections. We compared the automatically generated reports, as well as the

intermediate results, to the reports produced by a GP. While the evaluation demonstrates feasibility, it also points out several limitations and challenges.

Limitations. More consultations are needed for a more reliable assessment of the pipeline's quality. However, the list of challenges in Sect. 5 suggest clear improvement points. Furthermore, the golden standards were written by a single GP; since medical students/professionals do not receive any formal training on writing reports using the SOEP format, other care providers might write slightly different reports. Papineni *et al.* referred to these situations as 'stylistic variations' [20]. Finally, some items in the golden standards are not mentioned explicitly and, thus, may not end up in the output of the triple extraction. To ensure that implicitly discussed relevant information is also extracted by the system, conversation analysis techniques are needed. For instance, when the patient and GP agree on a course of action, this is not contained in one sentence and can only be inferred from the dialogue between both [24].

Future Directions. We plan on generating more reports to acquire more data and to measure the effects of error propagation in the pipeline, as discussed in Sect. 4. We will ask more care providers to write reports in SOEP formats in order to be able to randomly select stylistic variations of golden standards to compare the output to. We also intend to include support for additional diseases and ailments by including them in the ontology. The results from our evaluation (the low number of `good` triples), together previous findings showing that over half of the utterances in the speech transcript are not relevant for reporting [18], call for relevance selection algorithms that diminish the amount of unnecessary information that is stored. Furthermore, we intend to implement support for the two additional modalities, video and sensors, as well.

Conclusion. While the generated reports are still imperfect, the proof-of-concept shows that dialogue summarization using the proposed pipeline structure is achievable when sufficient engineering effort is put in optimizing the implementation. When sufficiently improved, our pipeline can help care providers reduce their administrative burden and focus on direct patient care.

References

1. Ajami, S.: Use of speech-to-text technology for documentation by healthcare providers. Nat. Med. J. India **29**(3), 148–152 (2016)
2. Alkureishi, M.A., et al.: Impact of electronic medical record use on the patient-doctor relationship and communication: a systematic review. J. Gen. Intern. Med. **31**(5), 548–560 (2016)
3. Androutsopoulos, I., Lampouras, G., Galanis, D.: Generating natural language descriptions from OWL ontologies: the NaturalOWL system. J. Artif. Intell. Res. **48**, 671–715 (2013)
4. Antoniou, G., Van Harmelen, F.: A Semantic Web Primer. MIT press, Cambridge (2004)
5. Bosch, A.V.D., Busser, B., Canisius, S., Daelemans, W.: An efficient memory-based morphosyntactic tagger and parser for Dutch. LOT Occas. Ser. **7**, 191–206 (2007)

6. Cameron, S., Turtle-Song, I.: Learning to write case notes using the SOAP format. J. Couns. Dev. **80**(3), 286–292 (2002)
7. Campanella, P., et al.: The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur. J. Public Health **26**(1), 60–64 (2015)
8. Chaiyachati, K.H., et al.: Assessment of inpatient time allocation among first-year internal medicine residents using time-motion observations. JAMA Intern. Med. **179**(6), 760–767 (2019)
9. Chiu, C.C., et al.: Speech recognition for medical conversations. arXiv preprint arXiv:1711.07274 (2017)
10. Deng, L., Huang, X.: Challenges in adopting speech recognition. Commun. ACM **47**(1), 69–75 (2004)
11. Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A.G., Draicchio, F., Mongiovì, M.: Semantic web machine reading with FRED. Seman. Web **8**(6), 873–893 (2017)
12. Golob Jr., J.F., Como, J.J., Claridge, J.A.: The painful truth: the documentation burden of a trauma surgeon. J. Trauma Acute Care Surg. **80**(5), 742–747 (2016)
13. Jiang, F., et al.: Artificial intelligence in healthcare: past, present and future. Stroke Vasc. Neurol. **2**(4), 230–243 (2017)
14. Klann, J.G., Szolovits, P.: An intelligent listening framework for capturing encounter notes from a doctor-patient dialog. BMC Med. Inform. Decis. Making **9**(1), S3 (2009)
15. Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S., Tagaris, G.: Deep neural architectures for prediction in healthcare. Complex Intell. Syst. **4**(2), 119–131 (2017). https://doi.org/10.1007/s40747-017-0064-6
16. Liu, C., Talaei-Khoei, A., Zowghi, D., Daniel, J.: Data completeness in healthcare: a literature survey. PAJAIS **9**(2), 75–100 (2017)
17. Luchies, E., Spruit, M., Askari, M.: Speech technology in Dutch health care: a qualitative study. In: HEALTHINF, pp. 339–348 (2018)
18. Maas, L., et al.: The Care2Report system: automated medical reporting as an integrated solution to reduce administrative burden in healthcare. In: Proceedings of HICSS (2020)
19. Ng, V.: Advanced machine learning models for coreference resolution. In: Poesio, M., Stuckardt, R., Versley, Y. (eds.) Anaphora Resolution. TANLP, pp. 283–313. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-47909-4_10
20. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL, pp. 311–318 (2002)
21. Portet, F., et al.: Automatic generation of textual summaries from neonatal intensive care data. Artif. Intell. **173**(7–8), 789–816 (2009)
22. Rathert, C., Mittler, J.N., Banerjee, S., McDaniel, J.: Patient-centered communication in the era of electronic health records: what does the evidence say? Patient Educ. Couns. **100**(1), 50–64 (2017)
23. Schmitz, M., Bart, R., Soderland, S., Etzioni, O., et al.: Open language learning for information extraction. In: Proceedings of EMNLP-CoNLL, pp. 523–534 (2012)
24. Sidnell, J., Stivers, T.: The Handbook of Conversation Analysis, vol. 121. Wiley, Hoboken (2012)
25. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: robust DNN embeddings for speaker recognition. In: Proceedings of ICASSP, pp. 5329–5333. IEEE (2018)

26. Syed, Z., Rubinfeld, I.: Unsupervised risk stratification in clinical datasets: identifying patients at risk of rare outcomes. In: Proceedings of ICML, pp. 1023–1030 (2010)
27. Valls-Vargas, J., Zhu, J., Ontanon, S.: Error analysis in an automated narrative information extraction pipeline. IEEE Trans. Comput. Intell. AI Games **9**(4), 342–353 (2016)
28. Woolhandler, S., Himmelstein, D.U.: Administrative work consumes one-sixth of US physicians' working hours and lowers their career satisfaction. Int. J. Health Serv. **44**(4), 635–642 (2014)