GRACE: Generating Summary Reports Automatically for Cognitive Assistance in Emergency Response

M Arif Rahman,[1] Sarah M. Preum,[1] Ronald Williams,[2] Homa Alemzadeh,[2] John A. Stankovic[1]
[1]Department of Computer Science, [2]Electrical and Computer Engineering
University of Virginia
{mir6zw, preum, rdw, ha4d, jas9f}@virginia.edu

Abstract

EMS (emergency medical service) plays an important role in saving lives in emergency and accident situations. When first responders, including EMS providers and firefighters, arrive at an incident, they communicate with the patients (if conscious), family members and other witnesses, other first responders, and the command center. The first responders utilize a microphone and headset to support these communications. After the incident, the first responders are required to document the incident by filling out a form. Today, this is performed manually. Manual documentation of patient summary report is time-consuming, tedious, and error-prone. We have addressed these form filling problems by transcribing the audio from the scene, identifying the relevant information from all the conversations, and automatically filling out the form. Informal survey of first responders indicate that this application would be exceedingly helpful to them. Results show that we can fill out a model summary report form with an F1 score as high as 94%, 78%, 96%, and 83% when the data is noise-free audio, noisy audio, noise-free textual narratives, and noisy textual narratives, respectively.

1 Introduction

Emergency Medical Service (EMS) responders communicate extensively with many different stakeholders in emergency scenarios to ensure that the correct measures are taken and adverse outcomes are minimized. While communicating, the severity of the scene as well as the condition of the injured patients are often mentioned. State-of-the-art technologies such as omni-directional microphones, noise-canceling microphones, headphones, the global positioning system (GPS) and other devices aid the communication and recovery procedure. Currently, a textual narrative of the scene as well as a summary report for the patients are created afterward. These reports often lack critical details that are collected from the scene in real-time, but forgotten. Research shows that in the USA, 13.6% of the time mistakes are made while inputting information into the summary forms. Mistakes happen in the form of inputting wrong information, forgetting to include a correct piece of information and misplacing data in the wrong field of a form (Burnett et al. 2011). Such manual errors can be attributed to the following factors. First, unfavorable circumstances such as getting a call at 2 AM as well as multitasking activities at the scene create adversarial conditions for the first-responders. Second, as responders try to remember the events from the scene, their recall of the events is often not 100% accurate. Finally, most emergency scenes demand dynamic information flow, such as changing vitals, changing medication dosage, etc. which makes the task of post-incident form filling accurately even more difficult. Discussions with first responders indicate that **automatic** form filling followed by only needing to check the forms would be a tremendous aid in their jobs.

At first, with the availability of accurate transcription tools and the current state of NLP research, this may seem like a simple task. However, this is not true, as many challenges must be overcome. These challenges include:

(i) domain-specific concept extraction that is unique for emergency response when compared to current medical and clinical oriented ontologies, as the specialized vocabulary used by first responders limits the applicability of current solutions;

(ii) semantic inference from EMS data, e.g., negation detection, temporal expression detection, and value association for accurate information extraction;

(iii) minimizing the effects of noisy environments and noisy data, missing data, homophones, and other realistic speech issues on information extraction;

(iv) deep inference of EMS text, including, (a) distinguishing patient-related information from scene and unrelated information in the conversations and (b) chronological ordering of information since the scene is not always narrated linearly.

We developed GRACE (Generating Summary Reports Automatically for Cognitive Assistance in Emergency Response) to solve the above-mentioned challenges. We have collaborated with a regional ambulance agency to get access to 8,000 textual narratives of real EMS scenarios. We also developed 119 simulated audio versions of a subset of the narratives with and without noise to evaluate the variation of the performance of GRACE in presence of noise in speech

data, as most emergency scenes are noisy. Further noise insertion in textual corpus is investigated for the validation of GRACE. The main contributions in our paper are:

- Developed the first NLP based system to address formal documentation or reporting of critical information for emergency response. Our thorough evaluation uses real EMS dataset that includes both textual and speech EMS data. We have explored the applicability of three benchmark NLP clinical information extraction tools for EMS domain, namely, MetaMap (Aronson 2001), cTAKES (Savova et al. 2010), and CLAMP (Soysal et al. 2017). GRACE outperforms these benchmark tools for information extraction for documentation of emergency response events.
- Demonstrated the impact of noise on audio and textual narratives of emergency incidents and developed a resilient form-filling module that performs acceptably under adverse and noisy conditions. Since emergency response is a low-resource domain in terms of availability of realistic information-rich data, we have generated synthetic noisy conversational data with varying degree and types of noise based on real EMS data for evaluating GRACE.
- Resolved some semantic challenges of domain-specific information extraction for EMS documentation, including, negation detection in EMS text and information validation (e.g., vitals) for EMS data under both noise-free and noisy conditions.

2 Related Work

To the best of our knowledge, this is the first work to address the problem of automatic documentation for the EMS domain. Although, there has been a lot of work on developing smart assistants for emergency response, none of those focus on form-filling.

Cognitive and intelligent assistants for emergency response: Montanga et al. (Montagna et al. 2019) present TraumaTracker, a trauma tracking system for documentation. They demonstrate that the accuracy of trauma documentation significantly improves after using TraumaTracker, as the system adds data and information that were not recorded in the paper documentation. But this system is deployed only in the trauma domain, GRACE is more generic and can be used for any medical emergency scenario, if the documentation format is similar. Preum et al. in (Preum et al. 2018), Shu et al. in (Shu et al. 2019) and Lindes et al. in (Lindes, Lonsdale, and Embley 2015) discuss the idea of developing cognitive assistant systems to improve and aid the awareness of first-responders. However, they do not focus on EMS incident report generation or documentation for the patients involved.

Transportation, health and many industry applications have seen different cognitive assistant systems over the years. Authors in (Ha et al. 2014) illustrate a Google glass based assisting system, which is developed to perform context-aware real-time scene interpretation by identifying objects for people suffering from cognitive decline. While the system is useful for this group of people, emergency situations often result in compromising visual capabilities and video signals may not always carry the whole information due to missing angles, and other adverse conditions. Thus, audio data and on-scene conversations are more trustworthy sources for EMS and our module uses them for documentation of patients.

Documentation and reporting tools for emergency response: *ImageTrend*[1] is an increasingly popular tool for documentation, tracking and visualization of EMS information.

Another software *Emergency Department Information Exchange (EDIE)*[2] links all hospital emergency departments by facilitating real-time communication and collaboration. However, both ImageTrend and EDIE, require manual input in the initial phase of data collection which is tedious and prone to errors. GRACE does not require any such effort, as summary reports are automatically generated using the audio data from on-scene EMS conversations.

Clinical information extraction: Different tools exist for extracting information from unstructured clinical texts, including, MetaMap (Aronson 2001), cTAKES (Savova et al. 2010), and CLAMP (Soysal et al. 2017). MetaMap combines natural language processing (NLP) with knowledge-intensive approaches for clinical concept identification and mapping or normalization. The Clinical Text Analysis and Knowledge Extraction System (cTAKES) combines rule-based and machine learning techniques to achieve this. CLAMP is a comprehensive clinical Natural Language Processing (NLP) software that enables recognition and automatic encoding of clinical information in narratives. All three of MetaMap, cTAKES and CLAMP use the Unified Medical Language System (UMLS) to extract medical concepts. One of the main issues with using these tools for EMS documentation or form filling is categorizing the contexts in finer granularity. For example, MetaMap has Concept Unique Identifiers (CUI) and semantic type lists which signify whether a clinical concept is 'Disease' or 'Medication'. But there is no way to differentiate whether the disease or medication is the current condition of the patient or an occurrence from the past. GRACE, on the other hand, uses NLP based heuristics to categorize contexts in finer granularity which is necessary for filling the form. There have been some other works on clinical document summarization and information extraction including (He 2016; Mujjiga et al. 2019). However, they focus only a subset of information relevant for EMS documentation and require significant amount of annotated data, which is not available for the EMS domain.

3 Approach and Solution

Figure 1 shows the overview of our solution. Although there are many tools to extract medical information, categorizing them into specific fields of the EMS form requires further text processing. This requires additional logic and heuristics compared to the state-of-the-art tools. In the following subsections, we describe our solution.

[1] https://www.imagetrend.com/

[2] https://collectivemedical.com/ed-utilization/

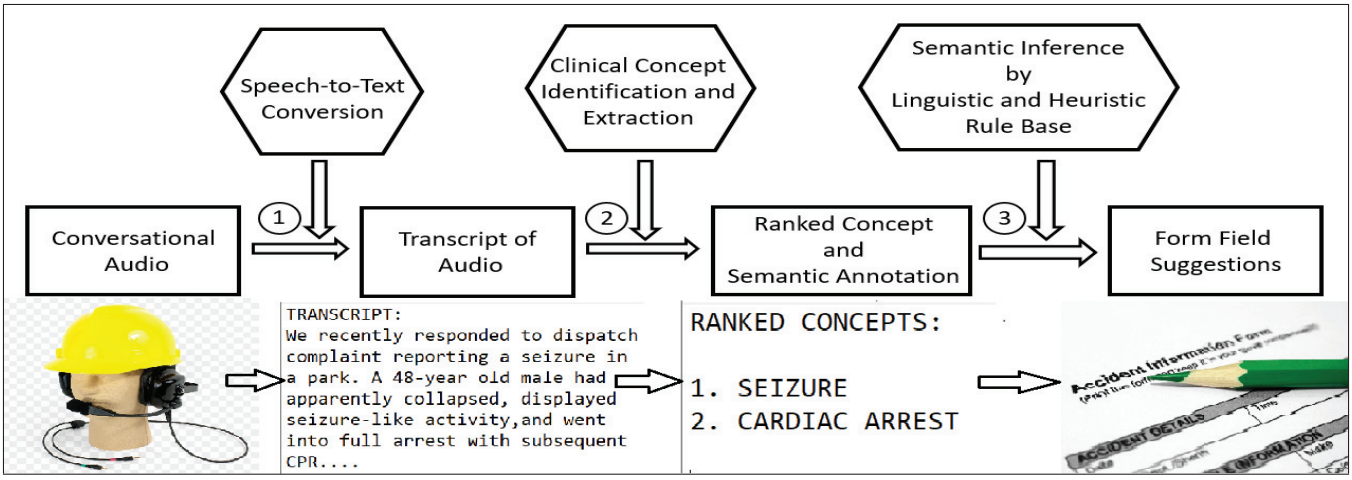


Figure 1: Solution steps for GRACE

Speech-to-text conversion

The first step of our solution is speech-to-text conversion, marked by ① in figure 1. There is a lot of noise in EMS scenes, and the accuracy of transcriptions are significantly affected under such adverse conditions (Preum et al. 2018). Solving this problem is important, but not the subject of this paper. In the experiments in this paper, we consider both accurate transcription and noisy transcription to reflect the potential variations in the performance of off-the-shelf speech recognition tools.

EMS concept extraction

After speech-to-text conversion, EMS concepts are extracted from the converted text in step ② (Figure 1). Ranking of concepts is done using state-of-the-art medical NLP tools, i.e. MetaMap, CLAMP, and cTAKES. In this paper, we used Concept Unique Identifiers (CUIs) to filter concepts, and the ranking of concepts is done by using the confidence scores provided by MetaMap. A threshold for confidence score for each type of concept was defined by training our module with training data. Unless a concept ranked above the threshold, it was discarded.

Since MetaMap supports ranking of concepts and unique identifiers according to the confidence scores, it is used in GRACE for clinical context detection and concept extraction. On top of MetaMap, we use different heuristics and linguistic rules to extract necessary information for fields of the form. cTAKES and CLAMP are used for validating the output of MetaMap. First, all the clinical contexts are filtered through MetaMap to discard scene and non-patient related information. We have derived a minimum threshold of confidence score of 5.00 for each of the concepts to be considered. For some of the concepts though, the threshold score is higher. For example, to detect medication and intervention related information, we keep the threshold to 5.00 to ensure all possible concepts are extracted. But for chief complaint or illness history of the patient, our tests with training data illustrate that a threshold of 10.00 works best by omitting false positives. The clinical concepts above certain confidence score are further checked with cTAKES and CLAMP, to ensure that all the state-of-the-art tools identify those as clinical concepts. Unless two of the tools signify a concept as clinical, we discard them. After filtering out non-clinical concepts, we try to understand the semantic meaning (next step below) of each concept and align them with the fields of the form.

Semantic inference

Understanding semantics in textual corpus is a challenging problem and different techniques for identifying semantics exist in the literature (Mujjiga et al. 2019). For semantic inference (step ③ in figure 1) such as **negation detection** and **value association** for vitals (i.e. blood glucose levels, Glasgow coma score, respiratory, blood pressure, pulse, peripheral capillary oxygen saturation or SPO2, etc.), we use modifier selection tools, dependency parsers, and entity recognizers. Specifically, NegEx (Chapman et al. 2001) and Stanford dependency parser (Cer et al. 2010) are used for negation detection and StanfordNER (Finkel, Grenager, and Manning 2005) is used for associating vitals to their values. However, without punctuation it is quite difficult to understand the context of the narrative. Researchers have identified various methods for adding punctuation in a text corpus (Say and Akman 1996), and recent developments have seen neural network based approaches. Authors in (Tilk and Alumäe 2016) discussed a recurrent bidirectional neural network for missing punctuation. Although this accuracy is not sufficient, we used their online tool to add punctuation into our transcripts as overall performance of GRACE improves afterwards.

4 Experimental Setting

Table 1 summarizes our datasets. We have generated synthetic data by adding relevant noise profiles to original noise-free audios, however some of our audio data originally had background noise. We have also used textual data from our

collaborator, a regional ambulance agency. To train our module, we have randomly selected half of each type of data shown in Table 1, while the other half is used for test purposes. The lengths of the audio files vary from one-minute to four-minutes. The artificial noise was added in continuous and discreet mode, and randomly. The amplitude of noise profiles were as high as the amplitude of the original audio, while the minimum amplitude of noise is half of the main audio. For textual narratives, all 32 annotated versions were randomly chosen and consists of minimum 1,000 words and maximum of 5,000 words. Due to limited and constrained resources, and restrictions in collecting live data in real-world EMS scenarios, we consider our dataset to be sufficient for this research. Also, annotating the dataset by professional EMS personnel is a time-consuming and difficult process. However, we are planning to collect more data from real world EMS training scenes and extend our collaboration with various Advanced Life-Support (ALS) EMS providers to enrich our experiments on this research.

Generating synthetic data

We have used the following five types of data for evaluating GRACE:

(i) EMS narratives: We have 8,000 post-incident narratives of different EMS scenarios from our regional collaborators. These textual corpora were used to determine the accuracy of our system. Since these data is not annotated and the annotation process is expensive in terms of both time and intellectual effort, this task can not be crowd-sourced for reliable and correct annotation. Instead, a small subset of 32 narratives was randomly selected from this data for training and testing purposes.

(ii) Noisy EMS narratives: We utilized different noise-insertion methods in existing research (e.g., (Agarwal et al. 2007)) to insert noise in the textual data mentioned above to validate the robustness of GRACE in presence of textual noise.

(iii) Noise-free audios from the EMS narratives: we have selected 12 test case scenarios from the data we obtained from our regional collaborator (different subset of data when compared to the subset mentioned above) and asked certified EMS responders to simulate a real scene for each of those. There was minimal ambient noise.

(iv) Noisy audio: The same procedure as above was followed, however, there was substantial noise around the scene. The noise was typically people talking in the background, screaming and ambulance siren.

(v) Simulated noisy audio: For the noise-free audio mentioned in the third point above, 8 different types of artificial background noise were inserted with varying degree of intensity. Thus 96 additional synthetic noisy audio data were generated from 12 noise-free audios and 8 noise profiles.

We conducted our experiments according to the form layout from one of our regional collaborators- a local fire response agency. Figure 2 shows the fields in the form. The minimum fields required in a post EMS documentation are locally standardized by *ImageTrend* (mentioned in section 2), and we included all the required fields in our model report. All the textual and audio data mentioned above was manually annotated according to this form layout by two graduate students working on this project, both of them are certified Emergency Medical Technicians (EMT). The annotations were further reviewed by certified EMS personnel to ensure correctness.Since our target is to measure how accurately GRACE can create summary forms, we have selected *Precision, Recall, and F1 Score* as our accuracy metrics.

Table 1: Description of synthesized datasets

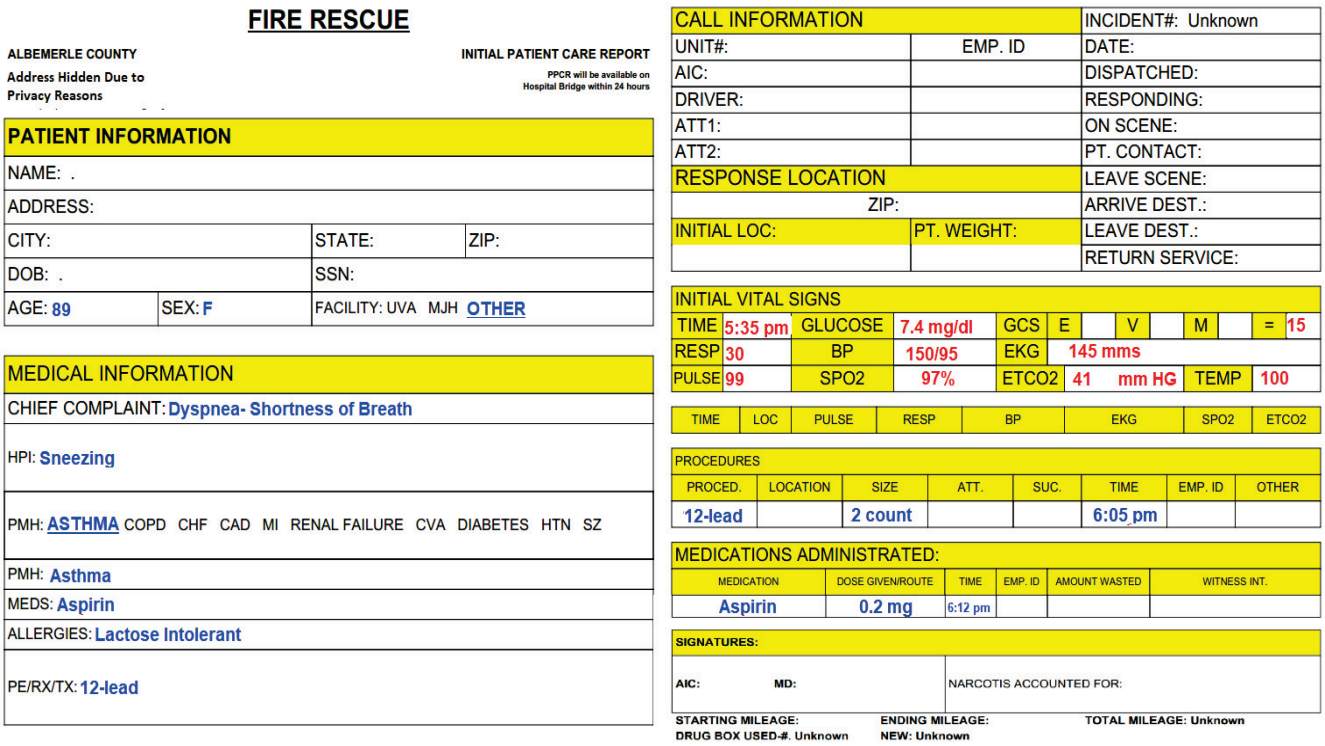
Type	Description	Size
Text	EMS narratives	32
	Noise-inserted EMS narratives	32
Audio	Noisy audio (with ambient noise)	11
	Noise-free audio	12
	Audio with artificially injected noise (using 8 noise profiles)	(12*8) = 96

5 Evaluation

GRACE outputs acceptable accuracy numbers for all the fields in the form shown in Figure 2. Typical fields such as Medication Administration, Vital Signs, and Procedures yield an average F1 score of 0.79, 0.86 and 0.71 for test data that includes noise-free audio, noisy audio, noise-free narratives and noisy narratives. The performances of the medical concept extraction tools (e.g. MetaMap, cTAKES, and CLAMP) are also comparable for these information fields. Due to limitation of space, we omit further details for these fields. Also, because of the paper space limits and because of their difficulties and importance we choose to show results of our negation detection step and the filling in the most important fields, including Chief Complaint, HPI (History of Present Illness), and PMH (Past Medical History).

Performance of negation detection

Although state-of-the-art tools use an enriched set of rule bases for detecting negation in clinical texts and electronic health records, it is difficult to identify if sentences have multiple negations or multiple contexts. For example, transcriptions such as *patient denied having shortness of breath* or *patient denied having lack of chest pain* contains double negations. Multiple negated contexts are also difficult to determine, e.g., "*patient denied having headache, shortness of breath and chest pain*". Another issue is that ill-punctuated transcriptions create lots of false positives in our train and test data while detecting negation. We have experimented with off-the-shelf and state-of-the-art tools such as DEEPEN (Mehrabi et al. 2015), NegEx (Chapman et al. 2001), MetaMap, and cTAKES. The accuracy of each tool is shown in Figure 3, the experiment was done with all of the test data in Table 1. NegEx outperforms all the other tools; the F1 score of NegEx is 0.81 with the highest precision compared to other tools. Although recall of DEEPEN is higher than NegEx (0.77 compared to 0.76 of NegEx), but the precision and F1 score is lower for our data. MetaMap and cTAKES have built-in negation detectors which can be



FIRE RESCUE

ALBEMERLE COUNTY
Address Hidden Due to Privacy Reasons

INITIAL PATIENT CARE REPORT
PPCR will be available on Hospital Bridge within 24 hours

PATIENT INFORMATION		
NAME: .		
ADDRESS:		
CITY:	STATE:	ZIP:
DOB: .	SSN:	
AGE: 89 SEX: F	FACILITY: UVA MJH OTHER	

MEDICAL INFORMATION
CHIEF COMPLAINT: Dyspnea- Shortness of Breath
HPI: Sneezing
PMH: ASTHMA COPD CHF CAD MI RENAL FAILURE CVA DIABETES HTN SZ
PMH: Asthma
MEDS: Aspirin
ALLERGIES: Lactose Intolerant
PE/RX/TX: 12-lead

CALL INFORMATION		INCIDENT#: Unknown
UNIT#:	EMP. ID	DATE:
AIC:		DISPATCHED:
DRIVER:		RESPONDING:
ATT1:		ON SCENE:
ATT2:		PT. CONTACT:
RESPONSE LOCATION		LEAVE SCENE:
ZIP:		ARRIVE DEST.:
INITIAL LOC:	PT. WEIGHT:	LEAVE DEST.:
		RETURN SERVICE:

INITIAL VITAL SIGNS												
TIME	5:35 pm	GLUCOSE	7.4 mg/dl	GCS	E		V		M		=	15
RESP	30	BP	150/95	EKG	145 mms							
PULSE	99	SPO2	97%	ETCO2	41	mm HG	TEMP	100				

TIME	LOC	PULSE	RESP	BP	EKG	SPO2	ETCO2

PROCEDURES							
PROCED.	LOCATION	SIZE	ATT.	SUC.	TIME	EMP. ID	OTHER
12-lead		2 count			6:05 pm		

MEDICATIONS ADMINISTRATED:					
MEDICATION	DOSE GIVEN/ROUTE	TIME	EMP. ID	AMOUNT WASTED	WITNESS INT.
Aspirin	0.2 mg	6:12 pm			

SIGNATURES:	
AIC: MD:	NARCOTIS ACCOUNTED FOR:

STARTING MILEAGE:
DRUG BOX USED-#. Unknown
ENDING MILEAGE:
NEW: Unknown
TOTAL MILEAGE: Unknown

Figure 2: Sample fields in patient summary report form, filled fields colored in blue/red demonstrate the output of GRACE

used solely for detecting negative phrases, but they perform poorly; their F1 score is 0.44 and 0.49, respectively. Since NegEx performs the best, we adapt NegEx for GRACE. Additional customization is done on top of NegEx by adding to the existing rule base and incorporating heuristics for detecting multiple negated contexts in a sentence.

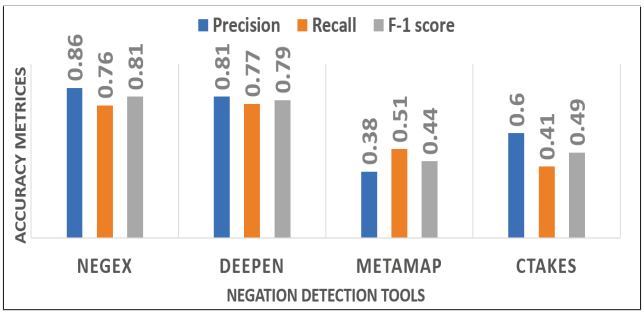


Figure 3: Baseline comparison for negation detection

Accuracy of critical medical information

Chief Complaint The chief complaint (CC) of the patient is challenging to define as there are multiple clinical contexts in the narrative. Medical information extraction tools provide different tags for chief complaint, e.g. "sign or symptom" by cTAKES; "findings", "sign or symptom" and "injury or poisoning" by MetaMap; or"problem" class in CLAMP. But these tags could relate to any of the contexts of other fields in the form also, such as past medical history, history of present illness, allergies and so on. On top of the tools used, hypothesis developed in GRACE detects the most likely candidate for chief complaint from the contexts in the transcription. Figure 4 summarizes the accuracy of our findings for chief complaint, and also demonstrates the comparison with the state-of-the-art tools. For the clarity of the figure and due to space limitations, we show the F-1 scores only. The tags mentioned above were used for each of the tools to extract the chief complaint candidates. These results are compared with the ground-truth data annotated by a real EMS responder.

We apply different heuristics and keyword identification for determining the chief complaint of the patient. Our investigation with EMS transcripts reveal that the chief complaint of the patient is generally mentioned at the beginning of scene description. Lack of correct punctuation causes difficulty in understanding the semantic meaning, thus we apply the punctuation insertion mechanism discussed in (Tilk and Alumäe 2016), after the speech-to-text conversion step. The resultant narratives are filtered for clinical concepts by MetaMap, cTAKES and CLAMP; we only select the clinical concepts that are found up to first three sentences. These clinical concepts have higher probability of holding the information of chief complaint of the patient. GRACE seeks for any mention of phrases like *"The patient is complaining of"* or *"Chief complaint is"* or *"The patient is suffering from"*, and if found then finds which clinical concept(s)

are related to that phrase using dependency parsers. The first two clinical concepts with highest confidence scores (determined by MetaMap, cTAKES and CLAMP) are selected as chief complaints unless such phrases are mentioned explicitly. If at least one common concept does not exist in the output of all three tools, we leave the field empty for post-scene manual input by first-responders with a highlighted remark to draw their attention.

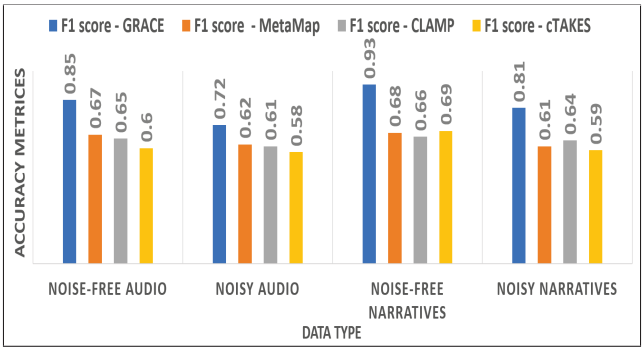


Figure 4: GRACE outperforms the state-of-the-art clinical information extraction tools for detecting Chief Complaint from each EMS dataset used in our evaluation.

The implication of the result in Figure 4 is two-fold. First, state-of-the-art tools are far off from defining clinical information in finer granularity. Although the concepts in concern are detected fairly accurately (acceptable precision), but the false positives and false negatives are too high (poor recall). MetaMap, CLAMP and cTAKES has an average F1 score of 0.65, 0.64 and 0.62 respectively for noise-free audio, noisy audio, noise-free narratives and noisy narratives while determining chief complaints. GRACE uses additional logic and filter to narrow down the possible results and achieves an average F1 score of 0.83. Second, many of the concepts are previous symptoms or past history, but they were detected as probable chief complaint by the tools. When using only the tags mentioned above, the tools return a lot of clinical concepts, most of which are effects of the chief complaint or related to the development of current conditions of the patient. GRACE, on the other hand, uses additional heuristics, ranking, and semantic inference to distinguish the clinical concepts, and selects chief complaint with better accuracy. Our understanding is that assuming the chronological development of patient's clinical condition in the transcription plays an important role in increased F1 score of GRACE, 0.85, 0.72, 0.93 and 0.81 for noise-free audio, noisy audio, noise-free narratives and noisy narratives, respectively. Most of the information transcribed in the middle or later sections of the audio data do not contribute to the chief complaint; information in the beginning holds all the true positives.

History of Present Illness (HPI) and Past Medical History (PMH) History of present illness (HPI) and past medical history (PMH) are very important information to understand patients' condition and the development of the symptom. Empirically, there are certain keywords and phrases which first-responders use to signify HPI and PMH, for example *"Patient has been feeling stomach ache for two days"* or *"She took pregnancy-related pills two months ago"*. Our heuristics use state-of-the-art NLP tools to understand the difference, and determine possible candidates for both of these fields in the form. The significance of detecting correct information in these two fields are very important, as the range of candidates span from clinical concepts to daily activities which might be linked with the current condition of the patient. Past information related to allergies are also critical, because many of our false positives are caused due to misclassification of this information, and interchanged content in these fields. Our heuristics only rely on specific keywords for this part, however we use an entity recognizer and different NLTK classifiers to separate the related information. Figure 5 and 6 summarize our findings for HPI and PMH respectively, comparison with other tools is irrelevant as there is no specific tag or semantics provided by these tools to identify the two categories. One important thing to mention here is that our module is tested on data which do not have any time-stamps, we assume that chronological ordering of development of patient's symptoms is maintained while transcribing. Average F1 scores for HPI and PMH are 0.71 and 0.70. This is due to the inability of GRACE to understand the context at times due to lack of proper punctuation and noise in transcriptions. Within sentence boundaries, transcribing multiple symptoms which relate to different fields of the form adds to the challenge. One important thing to mention for all fields of the form is that no specific keyword or verbalization was predefined while generating the synthetic data. It is our understanding that explicit mentioning of the context and better noise-canceling techniques can improve the accuracy of these fields.

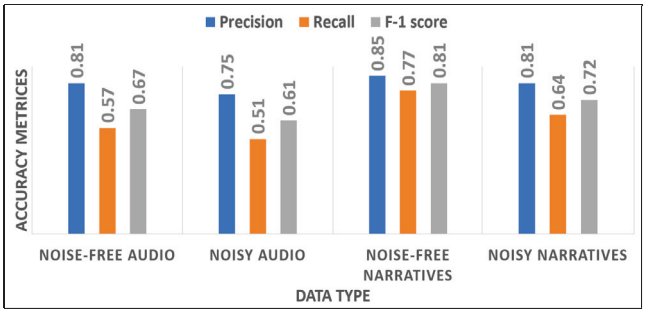


Figure 5: Accuracy of GRACE for detecting HPI

6 Conclusion

This paper addresses the problem of automatic summary report generation for patients involved in an EMS scenario. Using simulated audio data from the scene and conversations from first-responders, we show that our solution can generate an initial summary report by filtering and identifying relevant EMS information and context. We are the first to show that such documentation can be done with an F1 score as high as 94%, 78%, 96%, and 83% when the data is noise-free audio, noisy audio, noise-free textual narratives, and noisy textual narratives, respectively. Due to approval

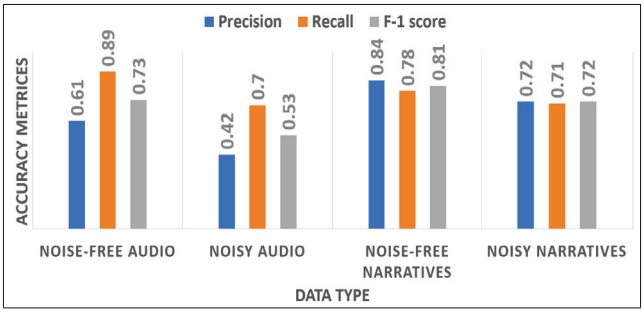


Figure 6: Accuracy of GRACE for detecting PMH

issues, we are yet to test our system in real-world EMS scenarios, but we are planning to deploy the system in EMS training soon. Our solution is not robust to all kinds of error and noise at the moment; however we claim that GRACE is very helpful for first-responders as it provides them with an initial draft of the summary of an injured patient, which can further be modified later manually if needed, to finalize post EMS scene documentation. The EMS responders do not have to completely depend on their memory for the task; and even though the accuracy is not perfect, the first-responders will highly benefit by the automate initial draft. In the future we plan to highlight missing interventions and critical inconsistencies detected from the conversation regarding patient's clinical condition. We also aim to develop more a generic and scalable approach by considering multi-patient and multi-responder scenes, and by applying machine learning techniques, respectively.

Acknowledgments. This work was supported by the award 60NANB17D162 from the U.S. Department of Commerce, National Institute of Standards and Technology (NIST).

References

Agarwal, S.; Godbole, S.; Punjani, D.; and Roy, S. 2007. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 3–12. IEEE.

Aronson, A. R. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, 17.

Burnett, S. J.; Deelchand, V.; Franklin, B. D.; Moorthy, K.; and Vincent, C. 2011. Missing clinical information in nhs hospital outpatient clinics: prevalence, causes and effects on patient care. *BMC health services research* 11(1):114.

Cer, D. M.; De Marneffe, M.-C.; Jurafsky, D.; and Manning, C. D. 2010. Parsing to stanford dependencies: Trade-offs between speed and accuracy. In *LREC*. Floriana, Malta.

Chapman, W. W.; Bridewell, W.; Hanbury, P.; Cooper, G. F.; and Buchanan, B. G. 2001. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, 105. American Medical Informatics Association.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, 363–370. Association for Computational Linguistics.

Ha, K.; Chen, Z.; Hu, W.; Richter, W.; Pillai, P.; and Satyanarayanan, M. 2014. Towards wearable cognitive assistance. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 68–81. ACM.

He, Y. 2016. Extracting topical phrases from clinical documents. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Lindes, P.; Lonsdale, D. W.; and Embley, D. W. 2015. Ontology-based information extraction with a cognitive agent. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Mehrabi, S.; Krishnan, A.; Sohn, S.; Roch, A. M.; Schmidt, H.; Kesterson, J.; Beesley, C.; Dexter, P.; Schmidt, C. M.; Liu, H.; et al. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of biomedical informatics* 54:213–219.

Montagna, S.; Croatti, A.; Ricci, A.; Agnoletti, V.; Albarello, V.; and Gamberini, E. 2019. Real-time tracking and documentation in trauma management. *Health informatics journal* 1460458219825507.

Mujjiga, S.; Krishna, V.; Chakravarthi, K.; and Vijayananda, J. 2019. Identifying semantics in clinical reports using neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9552–9557.

Preum, S. M.; Shu, S.; Ting, J.; Lin, V.; Williams, R.; Stankovic, J.; and Alemzadeh, H. 2018. Towards a cognitive assistant system for emergency response. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems*.

Savova, G. K.; Masanz, J. J.; Ogren, P. V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K. C.; and Chute, C. G. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5):507–513.

Say, B., and Akman, V. 1996. Current approaches to punctuation in computational linguistics. *Computers and the Humanities* 30(6):457–469.

Shu, S.; Preum, S.; M Pitchford, H.; D Williams, R.; Stankovic, J.; and Alemzadeh, H. 2019. A behavior tree cognitive assistant system for emergency medical services. In *Intelligent Robots and Systems (IROS)*.

Soysal, E.; Wang, J.; Jiang, M.; Wu, Y.; Pakhomov, S.; Liu, H.; and Xu, H. 2017. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* 25(3):331–336.

Tilk, O., and Alumäe, T. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech*, 3047–3051.