Spoken Medical Prescription Acquisition Through a Dialogue System on Smartphone: Perspective of a Healthcare Software Company

Ali Can Kocabiyikoglu[†], François Portet[⋆], Jean-Marc Babouchkine[†], Hervé Blanchon[⋆]
[†] Calystene SA, 38320 Eybens, France
[⋆]Univ. Grenoble Alpes, CNRS, Grenoble INP-LIG F-38000 Grenoble France
{a.kocabiyikoglu, jm.babouchkine}@calystene.com
{francois.portet, herve.blanchon}@imag.fr

Abstract

Industrial Medical Practice Management Software (PMS) have appeared in health institutions to reduce medication errors which affect several million people worldwide each year. However, practitioners must enter information manually into PMS which decreases the time devoted to care. In this paper, we describe the approach and some experiments of the implementation of an initial spoken dialogue system in this low-resourced domain in an industry-oriented setting. The main objective is to provide a natural language interface as an alternative to typing prescriptions in PMS. We highlight some of the difficulties of using deep-learning systems in an industrial context and discuss how these systems could be used while enabling full traceability. To overcome the lack of annotated speech data, we present a way to generate aligned data for machine-learning systems and discuss the limitations of using artificial data generation. We report on the findings of human evaluation conducted on the initial prototype with 2 medical experts and 2 naive users and discuss the results of each module of the dialogue system from an industrial perspective.

Keywords: Spoken Dialogue Systems, Natural Language Understanding, Health Informatics, Natural Language Processing

1. Introduction

Mobile applications, internet of things, big data, and digital health are affecting permanently the healthcare domain and all of its actors. In this transformation process, health information systems are becoming increasingly complex as a result of the diversity of the digital tools available to patients and healthcare professionals. After the earlier dialogue systems that simulated conversations between doctor-patient (Weizenbaum, 1966; Colby et al., 1971) and then that focused on commercial interactions such as advanced traveler information systems (Bobrow et al., 1977; Price, 1990), there has been an increasing interest in dialogue systems developed for health-related purposes. While Calystene SA is proposing healthcare informatics solutions for hospitals since 1992 in France, following this trend, we also have taken interest in dialogue systems as a part of our R&D program.

If most industrial dialogue systems have been based on some sort of expert rule-based models (Wallace, 2009), recent research on dialogue systems focused on neural approaches for dialogue systems (Wen et al., 2017; Ultes et al., 2017; Williams et al., 2017a) and led to end-to-end dialogue systems trainable by using a dataset of textual human dialogues (Ultes et al., 2017; D'Haro et al., 2020). The availability of deep learning frameworks and large datasets allowed the training of such systems. Hence, for task-oriented dialogue systems, instead of creating domain-specific semantics for each new task, these systems learn an intermediate semantic representation from data that is supposed to fit better the needs of the task.

However, from an industrial perspective, the goal is also to build agile, predictable, sustainable and scalable software. Hence, statistical approaches must provide control over the pipeline. Especially for the healthcare domain, without using any intermediate control mechanism, inferring directly an output from an input representation can have serious consequences.

Developing a dialogue system in the health care domain implies specific requirements, in particular, to meet the standards of security and confidentiality of health data, but also the standards of healthcare software conformity and effectiveness relating to the use of medical practice software. In this paper, we share our experience of the modeling process of a hybrid dialogue system that uses recent approaches for NLU and dialogue management while incorporating rule-based software PMS which would allow a control mechanism for the medical domain.

2. Overview of the Dialogue System

A classical dialogue architecture is based on a modular system where each component is responsible for a specific task (Williams et al., 2016). The different components of a dialogue system are the following: automatic speech recognition (ASR), spoken language understanding (SLU), dialogue state tracking (DST), dialogue policy, natural language generation (NLG) and text-to-speech (TTS). Since our objective is to deliver a prototype as a mobile application, we have to take into consideration several aspects regarding the human-computer interaction and other details that are not related to the Natural Language Processing (NLP) domain.

We approach the medical prescription understanding problem as a dialogue task in which the utterance initiated by the user must be understood, disambiguated and completed through goal-oriented dialogue. This way, missing information about the prescription could be completed through dialogue turns. When connected to a patient profile, prescription assistance software could warn the practitioner for potential adverse drug events (ADE) or adverse drug interactions, etc. The example described Figure 1 illustrates this strategy.

(1) **Spoken Language Understanding**
Prescriber: Metoprolol 200 mg one-half tablet once a day at night during dinner

(2) **Disambiguation and Information Filling:**
(METOPROLOL 200 mg, coated tablets, extended-release, route oral) (freq-ut: everyday,freq-startdate: immediately)

(3) **Requesting precision from the prescriber:**
System: Please specify the duration of the prescription
Prescriber: For one week

(4) **Proposition of a structured prescription:**
METOPROLOL 200 mg, tablets, route of administration oral. One-half tablet at night during dinner, starting from today for 1 week. Do you confirm?

(5) **Checking for drug interactions and patient history:**
System: Contraindication detected, the patient's file shows that the patient has had arthritis. Do you want to add this prescription to the patient's file?
Prescriber: No

Figure 1: Overview of the general approach

3. Challenges & Approach

Most of the dialogue systems created for health purposes focus on preventive health and medical data collection of patients, especially in the context of mental health problems (Fitzpatrick et al., 2017) which raises a lot of concerns in terms of privacy, ethics and effectiveness of the proposed solutions.

3.1. Low-resourced domain

One of the major challenges for biomedical NLP is the lack of freely available datasets for developing machine learning algorithms. There are very few datasets that are used in the NLP domain which are composed of natural language medical prescriptions. For example, the challenges i2b2, which took place in 2009, involved the extraction of medication information from electronic health records (Uzuner et al., 2010). As a part of this competition, 696 health records were released with 17 documents annotated by medical experts and 251 documents annotated by the scientific community.

Another dataset that is widely used in the medical domain is the MIMIC-III corpus which is an extension of MIMIC-II providing a massive amount of data also used by the NLP community (Johnson et al., 2016). However, this dataset is not annotated medical prescription-wise.

Another issue in NLP is that for languages other than English, the situation is even worse since the only paper we found was (Deléger et al., 2010), who applied techniques used for the i2b2 Shared Task to a French dataset extracted from 17,412 French EHRs. This dataset cannot be made available. We are not aware of any speech dataset related to medical prescriptions that would be available to the community.

For a spoken dialogue system, a corpus of dialogues is required for training and evaluating the systems. Since we were not able to find any corpus for this task, we approached the dialogue modeling by using interactive learning (Bocklisch et al., 2017). Regarding the SLU aspect, our approach for generating initial training data was to extract medical prescriptions from a medical textbook and to complement it with prescriptions generated by a context-free grammar. Although the artificial data generation techniques lack naturalness, it is capable of providing a large coverage (cf. (Kocabiyikoglu et al., 2019)).

3.2. ASR and NLU Quality in the Pipeline

The most frequent use of speech recognition in the clinical context has been speech recognition software for dictation (Kumah-Crystal et al., 2018). However, most of the earlier ASR systems were abandoned quickly due to recognition errors (Blackley et al., 2019). Recent systems transcribe conversational medical speech with around 20% Word Error Rate (WER) (Edwards et al., 2017; Chiu et al., 2017). Most of these systems are using cloud platforms which raises questions about the privacy of the data, especially knowing the fact that when proposing a solution to a client, the client is not necessarily aware of how the data is processed.

For a modular dialogue system, since the output of a module is feed as input into another module, errors appearing in the earlier modules of the system is propagated through other components and result in poor performances (Li et al., 2017). This is why the ASR stage and NLU stage must be as error-free as possible. Even though ASR systems are prone to errors, a spoken dialogue system can benefit from human intelligence by implementing implicit repetitions and visual confirmation of information on the smartphone application. Also, for a given utterance when the confidence level of the system is low, dialogue systems could implement fallback actions such as explicit confirmation to make sure that recognized utterances are validated by the user.

However, creating an ASR system for health-related purposes is a complex process and requires good quality conversational speech data. For demonstration purposes, we have started using the cloud-based speech recognition system available on Android smartphones. A recent study on major ASR services on medical conversational data shows that Google ASR produces on average 40% WER on healthcare-related conversational speech (Kodish-Wachs et al., 2018). For medical prescriptions, the entries are more clear and concise which allowed the use of an ASR cloud engine for demonstration purposes. However, in order to scale the dialogue system at the production-level, one should take into account the performance of ASR and the impact on the NLU system. Ideally, before the scaling process, it would be interesting to test toolkits such as Kaldi (Povey et al., 2011) which can run the ASR system locally even on Android platform, or explore end-to-end SLU system (Desot et al., 2019).

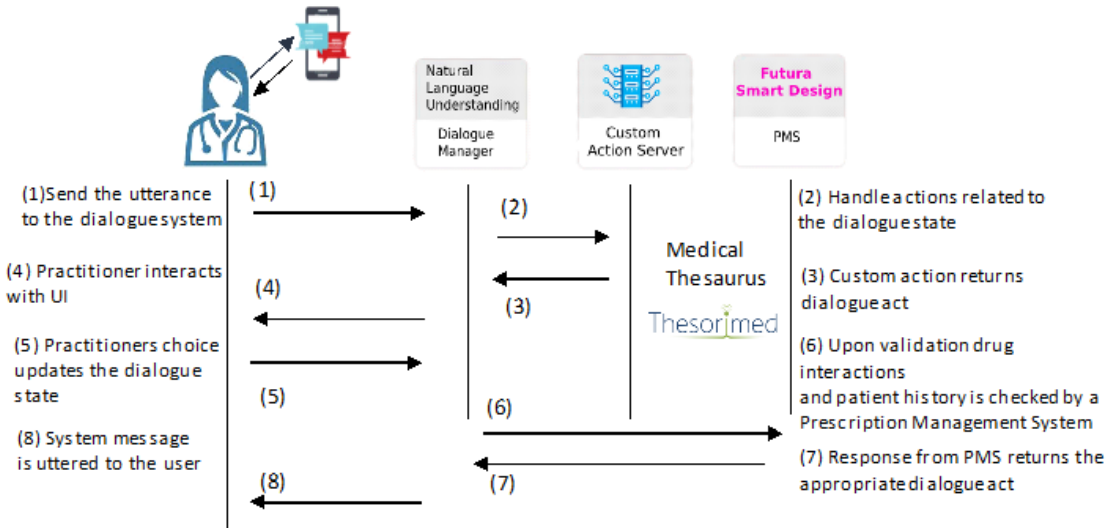

Figure 2: Pipeline of dialogue servers and external services

3.3. Medical Components of the Pipeline

Healthcare applications are subject to strict conformity checks often required by governments. In the USA, the HIPAA (Edemekong and Haydel, 2019) requires all healthcare facilities to implement strict rules to protect the confidentiality and integrity of patient information. In a similar way, in France, medical applications should be certified by the National Authority for Health (HAS).

One of the crucial aspects of conformity is to provide a seamless flow of information with medical knowledge bases. Healthcare databases are updated regularly (daily or weekly) and provide information about drugs and their interactions. For example, in the case of refusal of a marketing authorization from the government, medical applications should be capable of modifying the applications by disabling the prescription of some drugs.

Our PMS, Futura Smart Design® is certified by HAS and is used at over 50 health care facilities in France. Our knowledge base is based on Thesorimed®, a large database on drugs, which is updated regularly by the French National Health Insurance (Ameli). We are planning two levels of verification to ensure a secure information validation process. In order to do so, the processing pipeline involves different levels of verification using several medical knowledge bases. The overview of the different components that incorporates expert knowledge is illustrated Figure 2.

At first (1), when a prescription intent is inferred from an utterance, slot-fillers are extracted (2) to be associated with common dispensation codes (UCD) of drugs (3). If no drugs could be associated with the extracted semantic frames, the system suggests restarting the prescription process. If there is more than one potential drug corresponding to semantic frames, the user is provided with a list of drugs to choose and the process continues until all missing information is inferred (4 and 5). In the second stage, when the necessary prescription information is complete, we plan to send this structured data to PMS (6). For a given patient file, PMS handles the validation process of the prescription and give information about drug interactions, patient allergies, and so on (7 and 8). Until the prescriber validates the prescription or cancels it.

3.4. Training Procedure and Traceability

Over the last decade, systems based on deep neural networks surpassed other methods in most of the existing literature in many fields. Availability of large datasets, reduction of costs of cloud computing platforms and continuously evolving deep learning frameworks allowed this change especially in academic studies. Compared to academia, the adoption of fully statistical methods has been somewhat slower in industry. The reason is once a dialogue system is deployed, it should be a part of continuous integration (CI) and continuous deployment (CD) at any point. However, the process for developing, deploying, and continuously improving deep neural networks is more complex compared to traditional software (Sculley et al., 2015).

In domains where the output of a system can have serious consequences, a system should be fully traceable. In case of an error, one should be able to pinpoint easily the source of an error. Decoupling tasks in a modular system allow this partly by creating logs of each process.

In industry, it's not uncommon to 'hack' a solution by manually modifying behavior in software or correcting an entry in the database when there is a specific need coming from a client. A neural network is often viewed as a black box in the sense that while it can approximate any function, its structure does not provide any insight into the function being approximated. Thus, quick-fix solutions and even adjusting the system according to new data require the retraining of the whole system. Their behavior is often complex and hard to predict, harder to test, explain, and maintain (Sculley et al., 2015). A hybrid approach tackles this problem by allowing to train/adapt only the module that is concerned by the changes (Williams et al., 2017b). We are using the freely available Rasa X [1] toolset which allows this by providing a simple web interface that allows modifying the NLU, dialog scenarios and the domain definition from

[1] http://www.rasa.com/

	Task Sucess Rate	Average Dialogue Turns	NLU (f-measure)	WER (ASR)	Drug Association Rate (on TP)	Average Time Elapsed (on success)
medical experts	45%	1.56	0.75	3.40%	0.62	30 seconds
naive users	16.6%	1.54	0.43	17.35%	0.65	35 seconds

Table 1: Results of the human evaluation of the dialogue system

a web interface.

4. Experimental Results

In order to have early feedback about the prototype, we performed a human-based experiment. This experiment had a double objective: collect speech dialogue corpus in French using the dialogue system and evaluate the dialogue modeling which extends our previous work of the evaluation of our NLU system (Kocabiyikoglu et al., 2019). For the evaluation process, we have contacted two medical experts and two naive users for prescribing medicine using our mobile dialogue system.

Prescribing medicine is not an easy task for a naive user, even when the information to utter is provided. Therefore, we have prepared two procedures, one for medical experts and another for naive users. To avoid reading behavior from the experts, they were given a textbook of therapeutics which presents clinical cases and for which medical prescription is presented in a non-natural language way (B. Gay, 2009). Thus they had to abstract the prescription before uttering it. For naive users, however, another therapeutics textbook for students in medicine was given (Perrot, 2015). All prescriptions information was explicitly presented so they did not need medical knowledge. Hence, reading behavior could not be avoided.

In order to challenge the system and obtain various examples, the textbook prescriptions were ranked according to their complexity. Each participant had to make 10 medical prescriptions using the mobile application a headset and a microphone in a silent room. For medical experts, they had to try at least two challenging examples. In total, 40 dialogue were collected from 2 medical experts and 2 naive users with 10 prescriptions each. The implementation process does not include the interactions with the PMS and focused on the prescription and drug association.

The results of the evaluation are summarized in table 1. The global task success rate (ratio of validated prescriptions) which describes if the prescription has been completed or not is low for both medical experts and naive users.

It can be seen that for medical experts, the ASR Word Error Rate (WER) is very good while the NLU stage exhibits a fair f-measure of 0.75 which stays in line with our previous study on our NLU performance (Kocabiyikoglu et al., 2019). The picture is far less good regarding naive users.

A behavior which is common for both type of users is the low dialogue turn (about 1.5). In fact, the dialogue stopped quickly because the system often responded "drug not found". This is illustrated the low score of the drug association rate (about 40% of error). Another problem was due to the difficulty of recognizing frequency (e.g., every week) and duration (e.g., for the next two weeks). These elements have been reported as been difficult to extract in the i2b2 challenges as well (Uzuner et al., 2010) and are due to the lack of training data of intermediate interactions such as precising the duration or the dosage of the prescription.

In case the drug is associated correctly, the prescription process takes around 20-30 seconds which is quite reasonable with respect to the typing procedure.

Overall, this quick evaluation shows that although some components give satisfying performances the overall pipeline is not robust enough for the expert (the target user of the product). The evaluation of the modular architecture permits to identify the necessary improvement such as the NLU component (needs to be trained with examples for precising duration and frequency of the prescription) the drug identification as well as the dialogue management. Indeed, the system has been trained using cooperative scenarios. However, a good number of dialogues entered a fallback loop because of the lack of uncooperative scenarios.

The evaluation also showed that including non-expert in the process emphasizes the difference in behavior and language with the expert. Indeed, the poor performance of the naive users was mainly due to some formulations that were less technical and more familiar which differed significantly from most of the training examples. Although very limited in size, this evaluation shows the importance of performing experiments including target users.

5. Conclusion

This paper presents an approach to make oral medical prescription possible for an industrial prototype through dialogue on a smartphone. We discussed the trade-off between fully statistical end-to-end systems and the need for control, maintenance, traceability and privacy in a real industrial setting. An initial working pipeline prototype has been developed using a mixture of inference models acquired by machine learning, expert system and professional knowledge bases. The evaluation of this initial prototype showed the importance of keeping a modular architecture to identify the components that need improvement and emphasized the dependence of machine learning techniques on data; data that are very often unavailable in an industry setting. This calls for more research in machine learning to develop methods that could benefit both from expert knowledge and a reduced amount of data.

References

B. Gay, P.-L. Druais, P. A. T.-D. (2009). *Thérapeutique en médecine générale*. apnet.

Blackley, S. V., Huynh, J., Wang, L., Korach, Z., and Zhou, L. (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the American Medical Informatics Association*, 26(4):324–338.

Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H., and Winograd, T. (1977). Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

Bocklisch, T., Faulkner, J., Pawlowski, N., and Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.

Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., et al. (2017). Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

Deléger, L., Grouin, C., and Zweigenbaum, P. (2010). Extracting medication information from French clinical texts. In *MEDINFO 2010*, pages 949–953, Amsterdam.

Desot, T., Portet, F., and Vacher, M. (2019). SLU for voice command in smart home: comparison of pipeline and end-to-end approaches. In *IEEE Automatic Speech Recognition and Understanding Workshop*, Singapore.

D'Haro, L. F., Yoshino, K., Hori, C., Marks, T. K., Polymenakos, L., Kummerfeld, J. K., Galley, M., and Gao, X. (2020). Overview of the seventh dialog system technology challenge: Dstc7. *Computer Speech & Language*, page 101068.

Edemekong, P. F. and Haydel, M. J. (2019). Health insurance portability and accountability act (hipaa). In *StatPearls [Internet]*. StatPearls Publishing.

Edwards, E., Salloum, W., Finley, G. P., Fone, J., Cardiff, G., Miller, M., and Suendermann-Oeft, D. (2017). Medical speech recognition: reaching parity with humans. In *International Conference on Speech and Computer*, pages 512–524. Springer.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Kocabiyikoglu, A. C., Portet, F., Blanchon, H., and Babouchkine, J.-M. (2019). Towards spoken medical prescription understanding. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

Kodish-Wachs, J., Agassi, E., Kenny III, P., and Overhage, J. M. (2018). A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In *AMIA Annual Symposium Proceedings*, volume 2018, page 683. American Medical Informatics Association.

Kumah-Crystal, Y. A., Pirtle, C. J., Whyte, H. M., Goode, E. S., Anders, S. H., and Lehmann, C. U. (2018). Electronic health record interactions through voice: a review. *Applied clinical informatics*, 9(03):541–552.

Li, X., Chen, Y.-N., Li, L., Gao, J., and Celikyilmaz, A. (2017). Investigation of language understanding impact for reinforcement learning based dialogue systems. *arXiv preprint arXiv:1703.07055*.

Perrot, S. (2015). *Thérapeutique pratique 2015*. MEDLINE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

Price, P. (1990). Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511.

Ultes, S., Barahona, L. M. R., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gasic, M., et al. (2017). Pydial: A multi-domain statistical dialogue system toolkit. In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78.

Uzuner, Ö., Solti, I., and Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Wallace, R. S. (2009). The anatomy of a.l.i.c.e. In Robert Epstein, et al., editors, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pages 181–210. Springer Netherlands.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *EACL 2017*, pages 438–449, Valencia, Spain.

Williams, J., Raux, A., and Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Williams, J. D., Asadi, K., and Zweig, G. (2017a). Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *ACL 2017*, pages 665–677, Vancouver, Canada.

Williams, J. D., Asadi, K., and Zweig, G. (2017b). Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274*.