Towards Spoken Medical Prescription Understanding

Ali Can Kocabiyikoglu, François Portet, Hervé Blanchon, Jean-Marc Babouchkine

Towards Spoken Medical Prescription Understanding

Ali Can Kocabiyikoglu
Calystene SA, 38320 Eybens, France
a.kocabiyikoglu@calystene.com

François Portet, Hervé Blanchon
Univ. Grenoble Alpes, CNRS, Grenoble INP
LIG F-38000 Grenoble France
{francois.portet, herve.blanchon}@imag.fr

Jean-Marc Babouchkine
Calystene SA, 38320 Eybens, France
jm.babouchkine@calystene.com

***Abstract*—Prescription Management Systems (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 propose to provide a Natural Language interface to the PMS so that practitioners can record their prescriptions orally through mobile devices at the point of care. We briefly describe the overall approach and focus on the Natural Language Understanding process which was approached through slot-filling. To deal with the paucity of data and the imbalanced class problem, we present a method to artificially generate medical prescriptions. Experiments on the artificial and a realistic dataset with several state-of-the-art NLU systems show that the method makes it possible to learn competitive NLU models and opens the way to experiments on speech corpora.**

***Index Terms*—natural language understanding; spoken dialogue systems; medical computing; prescription management systems**

I. Introduction

For a few decades, Hospital Information Systems (HIS) have become the objective of hospitals in many countries [1]. HIS makes it possible to have a centralized database and software platform to deal with the discharge of patients, the given treatments, drugs, pharmacy details and so on. General adoption of HIS is seen as a way to achieve better management of activities and to improve efficiency and care. One of the tasks HIS is particularly useful for is the support for medical prescriptions. Indeed, it is reported that, in the USA, medication errors are experienced by 1.5 million patients per year [2]. A large number of these errors can be reduced by using Information Technology [3]. This is why Prescription Management Systems (PMS) have appeared in health institutions to reduce the number of medication errors during the prescribing/transcribing/administering process. PMS must go through a number of certification processes to ensure software compliance with minimum requirements for security, adequacy, and efficiency of prescriptions [4]. Such systems not only decrease the number of prescription errors but also have a direct impact on the management of drug stocks and the preparation and administration of treatments. However, HIS and PMS are fully efficient when all the medical information is entered digitally into the system. As a consequence, this has increased the amount of time nurses and physicians must spend entering information.

In this paper, we propose to provide a Natural Language interface to the Prescription Management Systems (PMS). This interface would enable practitioners to record their prescriptions orally or through free-text, a form of interaction closer to their usual practice. Such utterance would then be analyzed by Natural Language Understanding to send structured data to the PMS which would validate or not the prescription. In the long-term vision of the approach, the feedback from the PMS would be handled through dialogue. Such kind of interface would have many advantages. It would enable practitioners to use natural language interaction so that they do not have to learn new complex software interfaces in new care centers. Clinicians could use their own smartphone so that they get quickly familiar with the apps while enabling faster identification than standard login. Mobile interfaces would also enable clinicians to prescribe at the point of care which would save time and would enable better mobility [5]. Once mastered, voice-based PMS could save time for practitioners to concentrate on medical care. Furthermore, a dialogue based prescription system can report missing drugs in pharmacy and warn about adverse drug effects so that the practitioner can adapt the prescription in real time. Moreover, the dialogue policy could include the own care center policy about available drugs and treatment as well as best practice.

Extracting medical-related information from prescriptions in natural language is a challenging task. Indeed, it involves not only extracting correct medication name and dosage but also implicit medical indications as well as pharmaceutical remarks. For instance, in the example of Figure 1, the route of administration of the drug (*oral*) was not stated explicitly, however, from the given contextual information (*capsules*) the oral route can be inferred. Medication names could be ambiguous and new drug names the system has never seen could be challenging [6]. Furthermore, medical prescriptions contain non-exhaustive information about the drug. Hence, associating a prescribed drug to an explicit nomenclature and its national drug code requires disambiguation of implicit information.

In this work, we intend to address the problem of spoken medical prescription understanding by using a slot-filling

978-1-7281-0984-8/19/$31.00 ©2019 IEEE

(1)	amoxicilline	500	mg	2	gélules	2	fois	par	jour	pendant	huit	jours
	inn	d-dos-val	d-dos-up	dos-val	dos-uf	rhythm-perday	O	O	O	O	dur-val	dur-ut
	amoxicillin	500	milligrams	2	capsules	twice		a	day	for	8	days

Fig. 1. Example of a medical prescription with aligned slot labels.

approach coupled with a goal-oriented dialogue system which would allow requesting precision from the practitioner in order to acquire a valid prescription for the latest e-prescribing regulations. In this paper, we focus on the NLU part of the problem. To summarize, the paper brings the following contributions:

- The analysis of the medical prescription from an NLU perspective and a possible semantic description.
- A methodology to learn deep NLU models with low amount of data.
- Experiments on realistic data with several state-of-the-art systems.

This introduction is followed by a short review of the state of the art in Section II. The overall approach is briefly sketched and the NLU method detailed in Section III. Section IV describes how the data shortage problem has been addressed and the evaluation with several state-of-the-art models. The paper ends with a short conclusion and description of further work.

II. Related Work

Initial work on medical prescription processing was mostly performed from clinical free-text narratives. Early systems such as MedLEE [7] or MetaMap [8] were rule-based combining pattern-matching rules with external resources (UMLS, clinical databases). In the last decade, the i2b2 Shared Task on Medication Extraction [9] fostered the community on the task of prescription extraction from clinical texts. The task was to extract information about a medication including name, frequency, dosage, duration as well as reasons for medication. The challenge showed that while the highest-performing system [10] used machine learning classifiers, it was also dependent on handwritten rules as most of the other participants. More recently, several methods have been proposed to enhance the extraction capability of medical prescriptions from clinical texts. Most of them rely on Conditional Random Fields models [11]–[13]. For instance, [13] reported an increase in performances with respect to the 2009 results on i2b2 Shared Task datasets using CRF and word embeddings. CRF was the state-of-the-art model for NLU until the emergence of Deep Learning.

Handwritten rules are still a technique employed despite the time-consuming task of writing them [14]. This is because of some particular entities such as *duration*, and *reason* exhibits a high variability which is difficult to capture with machine learning using a few amounts of data. This data shortage is still a current problem since accessible prescription databases are rare. For English, the i2b2 dataset [15] is composed of 696 Electronic Health Records (EHR) written in English. Another corpus is the MIMIC-III dataset [16] which is a follow up of MIMIC-II released in 2010. The data covers 38,597 distinct adult patients and 49,785 hospital admissions. However, EHRs are not annotated with the medication information.In non-English languages, the situation is even worse, since the only paper we found was [17], who applied techniques used for the i2b2 Shared Task to a French dataset extracted from 17,412 French EHRs. Their rule-based system led to similar result with the French corpus than with the i2b2 English one. However, this experiment was performed in 2010 and, to the best of our knowledge the dataset was not distributed. Regarding voice processing, a recent study on ASR [18] reveals that although ASR assisted documentation has become increasingly common in clinical settings its effect on production and quality led to mixed results and stay in line with previous studies on the subject [19].

Although there is a large body of work on automatic biomedical information extraction from clinical texts [20] and on dialogue systems for health care [21], [22], work about automatic processing of oral medical prescriptions are rare. In fact, the only product we were able to find was FreePharma™ cited in a chapter in 2006 [23]. It was described as being able to extract medical prescriptions from speech captured from a PDA. However, no technical details are provided and the reference points to a dead URL. Another related work is the Mobi-Dev European project [5] which aimed at providing the next generation of mobile devices for clinicians at the point of care. In this projects' website, it is noted that PDAs enabled with a natural language recognition system linked to hospital information systems. However, despite our best efforts, we did not any scientific publications related to the project.

From this short state of the art, it appears that voice-based medical prescription understanding has been under-studied in the NLP community. Furthermore, it seems that there is a lack of datasets in non-English language and that the task has not been investigated using recent Deep Learning techniques.

III. Method

A. General Approach

Since many pieces of information may not be present in oral prescriptions or wrongly recognized, we approach the oral medical prescription understanding problem as a dialogue task in which the utterance initiated by the user must be understood, disambiguated and completed trough goal oriented dialogue. The example described in Figure 2 illustrates this strategy. In the first step, an utterance is analyzed (1). The route of administration and the frequency of the prescription is not explicitly stated, however, from the given contextual information, the correct nomenclature of the drug could be matched and some temporal slots could be inferred as shown in (2). To comply with the latest e-prescribing regulations,

(1) **Utterance:**

amoxicilline	500	mg	2	gélules	2	fois	par	jour	pendant	huit	jours
inn	d-dos-val	d-dos-up	dos-val	dos-uf	rhythm-perday	O	O	O	O	dur-val	dur-ut
amoxicillin	500	milligrams	2	capsules	twice		a	day	for	8	days

(2) **Disambiguation and Information Filling:**
(AMOXICILLIN 500 mg, capsules, route oral) (freq-ut: everyday,freq-startdate: immediately)

(3) **Requesting precision from the prescriber:**
System: At what time of the day should the patient take the two capsules?
Prescriber: One capsule in the morning, one at night.

(4) **Proposition of a structured prescription:**
AMOXICILLIN 500 mg, capsules, route of administration oral. One capsule in the morning, one capsule at night, starting from today for 8 days.

(5) **Checking for drug interactions and patient history:**
System: Contraindication detected, the patient has an allergy for penicillin. Do you want to add this drug to the prescription?
Prescriber: Abort

Fig. 2. Example dialogue

further information could be requested from the prescriber through dialogue as shown in (3). Once all of the requested information is provided, the prescription could be uttered to the practitioner to be confirmed explicitly as shown in (4). Finally, the validation process of HIS could warn the practitioner for contraindications and patient background as in example (5).

In this paper, we focus on the first step aiming at performing NLU from the utterance using the *slot-filling* approach. Slot-filling consists in extracting the overall *intent* of an utterance and identifying the most important elements called *slots*. The intent reflects the intention of the speaker while the slots can be defined as the entities and relations in the utterance which are relevant for the given task [24]. For instance, the utterance of (1), is composed of 8 slots carrying crucial meaning about the prescription. *Amoxicillin* is the active substance of the prescription and 500 is the value and $milligram$ the unit of the strength of the drug. The expression *twice a day* describes the rhythm instructed by the practitioner, which is too vague with respect to e-prescribing regulations. This shows that additional information should be requested from the practitioner. We propose to handle this process by dialogue.

Since the prescription system is expected to be used on the ward, many utterances may be recorded by mistake or for completing or canceling the current medical prescription. Hence, to handle this different usage, the system must also infer the intent of the user. In this work, two intents will be considered (*prescription* and *non-prescription*) and which will be extended in the future. Next section details the semantics of the slots used for the NLU task.

B. Semantics of medical prescriptions

Slot filling approaches must have a clear semantic domain definition which should be strongly linked with the application domain for a high interoperability. To address this issue, there has been a great effort in the health informatics domain for creating a common nomenclature which would allow information exchange flawlessly between information systems. In the US, a standardized drug nomenclature has been developed by the National Library of Medicine [25]. Furthermore, there are various semantic web ontologies proposed for representing medical concepts of a drug prescription [26], [27].

However, these ontologies and standards are established for HIS and does not cover variations we could encounter in a natural language formulation. For instance, in the i2b2 data challenge, 6 slots were defined (*medication, dose, mode, frequency, duration* and *reason*) for which text expressions had to be associated with. But these slot definitions are not sufficient for PMS. Thus, we extended this list with other concepts related to drug dosage conditions, pharmaceutical remarks and temporal expressions used in various medical specialties. This definition is schematized in Figure 4 of the appendix.

To deal with PMS requirements, we broke down high-level slots into finer granularity slots. This definition was built on the formal definition of a certified PMS system [4], on thesauri of the domain, on analyses of real medical prescriptions and interactions with an expert in medical prescriptions. For example, two slots were defined for drug strength: value of the dosage and unit of the dosage. In this way, using a slot-filling approach, missing or wrong information could be modified through dialogue.

At the end of the process, we also defined two intents: *medical prescription* and *None* for none-prescription; 39 slot-labels and; 269 slot-values to describe the medical prescription domain.

Distinct slot-values for each international non-proprietary name (*INN*) –such as *paracetamol*– and commercial drug brand (*drug*) were not considered to reduce the complexity and allow the system to adapt to new coming drug brands. A complete list of slot-values for each INN and drug was not considered.

C. NLU methods

While early slot-filling systems were rule-based [28], modern methods are data-driven. Conditional random fields [29] have recently been superseded by deep neural networks, including basic RNNs [30], Bi-directional LSTM RNN encoder-decoders [31], Attention-based RNNs [32] and Attention based CNNs [33]. Most approaches treat slot-filling as sequence labeling, attaching a slot to each word in the input utterance. However, other approaches are possible, such as treating it as a dependency parsing task [33]. While intent detection has traditionally been seen as a separate task from slot-filling [34], since both tasks are highly correlated, much recent work performs slot-filling (sequence labeling) and intent detection (sequence classification) simultaneously. Such work includes Tri-CRF [29], which extends the linear sequence labeling CRF with a node to represent the dialogue act, and Att-RNN [32], which extends the slot-filling encoder-decoder RNN with an extra intent decoder.

In this work, we consider four different approaches to address the slot-filling task: Rasa NLU, Tri-CRF, Att-RNN, and seq2seq NLU. Rasa NLU[1], an open-source tool for building NLU pipelines, does not predict a sequence of slots for each input word, but rather a set of slot-labels and slot-values associated with different segments of the input. Tri-CRF from [29], [35] is an extension of a linear chain Conditional Random Field (CRF). Linear CRFs model the conditional probability distribution of the output label sequence, given the input sequences (sentences): each observed word x_t in a sequence is conditionally dependent on its corresponding *unobserved* label y_t. The Tri-CRF extends this model by adding an intent z for which each slot y_t (and also potentially each word x_t) is dependent on the overall sentence intent z.

The Attention RNN (Att-RNN) model [32], is a recurrent encoder-decoder architecture for simultaneous intent detection and slot labeling. The encoder is a bi-directional LSTM RNN which takes as input the sequence of words in an utterance, with one word x_t input at each time step. Output t each time step is the hidden state h_t of the bidirectional RNN.This final hidden state is then passed to separate decoders and is used to initialize their initial hidden states. These two decoders are the intent decoder and the slot-label decoder. At each decoding time step, the decoder outputs a slot prediction.

Like the Tri-CRF models, the Att-RNN only predicts intent and slot-labels, not both slot-labels and value-labels. Consequently, two models must be trained, one to predict intent and slot-labels and one for slot-values. Furthermore, all these systems need *aligned* data such as exemplified in Figure 3.

However, acquiring such aligned dataset is very time consuming and slows down the development of NLU systems to new application domains where there is a paucity of data. Furthermore, some domain-dependent implicit information may not be possible to align with the input. This is why we also implemented the seq2seq model from [36] using an encoder-decoder architecture which can be learned from unaligned data. Its architecture is similar to the Att-RNN one except it does not have a specific intent classifier. Contrary to Att-RNN the intent is treated as another slot as shown in the example below:

$$
\begin{aligned}
&medical\_prescription \\
&\quad (inn = inn = Amoxicillin \\
&\quad d-dos-val = d\_dos\_val\_numeric = 500 \\
&\quad d-dos-up = milligram = mg \\
&\quad dos-val = dos\_val\_numeric = 2 \\
&\quad dos-uf = capsule = capsules \\
&\quad rhythm-perday = rhythm\_perday = 2 \\
&\quad dur-val = dur\_val\_numeric = 8 \\
&\quad dur-ut = day = days)
\end{aligned}
$$

Fig. 3. Semantic alignment of a prescription

```
intent [ prescription ], d-dos-up
[ milligram ], dos-uf [ capsule ]...
```

Although models learned on unaligned data usually lead to poorer performances than the aligned ones, they are able to exploit more real-world datasets since their predictions are not restricted to a word-level horizon.

IV. Experiment

A. Datasets

As previously mentioned, we are not aware of any drug prescription data-set fit for NLU being made available in French. However, there is a large number of prescription textbooks for medical studies. Thus, to get access to realistic prescriptions, our strategy was to automatically extract 832 drug prescriptions from the "*Le Guide des Premières Ordonnances*" [37] textbook that has been bought by the authors. The book has been digitized to a pdf version from which prescriptions were automatically extracted. The 832 textual prescriptions were then pre-annotated using the semantics defined in Section III-B using a set of regular expressions. All these annotations were then manually checked and corrected by one of the authors. Then, a random subset of the annotated prescriptions was iteratively checked by a trained physician until convergence. All the prescriptions were annotated with the `drug prescription` intent.

To strengthen the learning and to evaluate the NLU modules in realistic conditions, French non-prescription utterances were also considered. Indeed, it is frequent that speech recording on smartphones can capture colloquial speech. Hence, a robust NLU approach must be able to distinguish true prescription intents from other situations. We used the ESLO corpus of conversational French speech [38]. ESLO is an adequate corpus since, similarly to spontaneous speech, it contains frequent disfluencies, repetitions, revisions, and restarts. From the ESLO2 corpus, 832 speech utterances were extracted and annotated with the `None` intent.

Despite this data collection, the amount of data is still too small for machine learning. Furthermore, as any realistic data collection, prescriptions are more biased towards some specific

[1] `https://rasa.ai/products/rasa-nlu/`

slots than others. Table I exhibit the frequency of some slots in the textbook.

TABLE I
EXCERPT OF FREQUENCY OF SLOTS IN THE TEXTBOOK

Frequent slots	Rare slots
drug: 816	inn : 17
rhythm-perday: 427	rhythm-rec-val: 5
d-dos-form: 125	d-dos-form-ext:5
dur-val: 190	re-val: 1

This situation is well known in machine learning and is called class imbalance. Several techniques exist to deal with this situation, from loss weighting to data augmentation using DNN [39]. However, loss weighting does not solve the data paucity and stochastic data augmentation needs an initial set of data which is still too important in our case. Hence, to solve the paucity of data and the class imbalance at the same time, we set up a simple generation technique.

We defined a feature-based context-free grammar for the medical prescription domain. Top-level rules of the grammar include different parts of the prescription described in III-B, whereas the terminals of the grammar are triplets of keyword, slot-label, and slot-value. The generator produces prescriptions containing under-represented slots by dynamically converting slot-labels to top-level expansion rules of the grammar.

Using half of the prescriptions acquired from [37] as our initial training data, the slot-label distribution is computed to identify candidate slots which are under-represented in the training set. Until a balanced distribution of slots is reached, random candidates are iteratively chosen to be produced as prescriptions. Finally, drug information is extracted from the French public drug database[2] in order to fill the triplets before adding the full prescription to the training data.

At the end of the process, the textbook and ELSO corpus were divided into training (50%), development (12%) and test (38%) sets. The training set was then complemented using the corpus generator to provide a balanced dataset. Table II summarizes the distribution of drug prescriptions by data sources for machine learning.

TABLE II
FINAL DISTRIBUTION OF TRAIN, DEVELOPMENT AND TEST SETS IN NUMBER OF PRESCRIPTIONS FOR THE THREE SOURCES

Corpus	Train	Dev	Test
TextBook	417	99	316
Artificial	3034	0	0
ESLO	417	99	316
Total	3868	198	632

B. Models training

Tri-CRF and Att-RNN predict the intent and slot-labels simultaneously while RASA requires training 2 separate models – one for the intent prediction, and – another for slot-label prediction. Moreover, neither Tri-CRF nor Att-RNN is able to learn slot-labels and slot values at the same time. Hence, two separate models are needed to perform label and value prediction. On the contrary, the seq2seq model predicts intent, slot-labels, and slot-values in only one model.

For the Tri-CRF model, to reduce training time, we pruned low-probability intents ($< 0.1\%$) and initialized the weights using the pseudo-likelihood (for 30 training iterations). Training proceeded for 200 iterations.

In our implementation of Att-RNN, the input words are first passed to a 128-unit embedding layer. The bi-directional LSTM encoder and decoder are each a single layer of 128 units. Training is performed using stochastic gradient descent (SGD) with a batch size of 16, using gradient clipping at a norm of 5.0, dropout with a keep-probability of 0.5 and training was allowed to continue for 30,000 training steps. We selected the trained model with the highest F1 score on the slot labeling task on the validation set. For Tri-CRF and Att-RNN, two models are trained, one to predict intent and slot-labels (Att-RNN-Labels) and one for slot-values (Att-RNN-Values). The Rasa configuration, 'spacy_sklearn', uses a linear chain CRF to classify slot-labels and a lookup table to determine slot-values. Separately, the model uses a linear SVM based on pre-trained word-embeddings to classify intents.

The seq2seq models is a recurrent encoder-decoder architecture with attention which uses a single layer bi-directional LSTM encoder and decoder with 128 sized encoder, decoder and embeddings layer. Learning was performed using SGD with batch size of 16. Learning rate was set to 0.0001 with dropout applied to all examples in the batch with a rate of 0.5. The vocabulary size was set to 10.000. Training continued until 10.000 iterations in total and the best model on the dev set was selected.

Since hand-crafted rules are still competitive in the medical prescription domain, a deterministic finite state automaton was implemented by regular expressions as a baseline model. A set of regular expressions for each slot-label defined in III-B was written and the annotation was performed by passing all of the regular expressions one by one on the utterance.

C. Evaluation

The overall results of the different systems are presented in Table III. For each task (intent and slot-label prediction), the F-measure was computed from the precision and recall. For the intent, this was easily performed since it is equivalent to a binary classification problem. For slot-label and slot-value classification, a sequence of predicted labels is produced per example. A prediction is considered as a true positive if the predicted label is equivalent to the true label. A false positive is considered when an incorrect prediction of a true label is made by the model whereas a false negative is considered when the model fails to predict the true label. Finally, a true negative is considered when the model does not perform a prediction correctly. Then, F-measure was computed as the average of precision and recall over the classes (macro average) as well as when taking the class weight into account (weighted macro average).

[2] http://base-donnees-publique.medicaments.gouv.fr

TABLE III
INTENT AND SLOT F-MEASURE IN THE TEST SET FOR THE DIFFERENT NLU MODELS

Model	Intent	Slot-label	Slot-label (w)
Baseline	-	0.61	0.47
RASA	0.97	0.67	0.62
Tri-CRF	0.97	**0.93**	**0.71**
Att-rnn	**0.99**	0.82	0.67
Seq2Seq	0.97	0.70	0.45

For intent classification, all of the models show similar results. This is not surprising since only two classes of intent were considered. Att-RNN exhibits near perfect intent classification. Regarding slot-label prediction, models trained with aligned data led to better results than the seq2seq model learned with unaligned data. Best overall F-measure score was performed by the triangular CRF model. This contradicts recent results comparing the performance of similar models [40]. However, this can be explained by the fact that CRF is more able to be learned from lower amount of data than seq2seq models. When considering the weighted F-measure, the ranking stays the same but RASA is far less affected by class weight than the other models. The weighted metrics of seq2seq model is similar to our baseline model.

The baseline model low performance is due to a low recall.

Similarly, Rasa and seq-to-seq models exhibit a global low recall. For tri-CRF, despite having generated a nearly balanced training set, most confusions appear between *inn* and commercial brand names *drug*. *d-dos-form* is also sometimes confused with *dos-uf* and *max-unit-uf*. Att-RNN shows a similar pattern of confusion but to a greater extent.

Models have been confronted with another difficulty which was due to the temporal expressions related to prescriptions duration. Furthermore, since the test examples were extracted from a textbook for students, some drugs are presented with alternatives.

D. Subsequent analyses

To evaluate the impact of the artificial dataset, the Tri-CRF model was trained without the artificial data. As a result, for the slot prediction, the recall decreased by 0.11 point and the F-measure (w) to .62. This result shows that the artificial dataset does enable to cover more cases during the training and thus decreases the number of confusions.

Furthermore, to get more insight into the NLU process for medical prescriptions, we have conducted two subsequent analyses using our best model (Tri-CRF): quantitative analysis and qualitative analysis. The quantitative analysis consisted in checking the influence of utterance size on model performances. In fact, as prescription gets longer, it conveys more information and prone to NLU errors. For this reason, we divided the test corpus into two parts: prescriptions shorter (resp. longer) than the average prescription length. Our divided test corpus has 186 short prescriptions and 130 long prescriptions.

The average length of short prescriptions is 41 ranging from 7 to 60 character length whereas 86 examples composed longer prescriptions which varied in length from 61 to 188 characters. Longer prescriptions include complex formulations and more slots which could be challenging for the model.

Short prescriptions overall weighted f-measure performance of 0.74 confirms that the system is slightly better when dealing with shorter prescriptions. For longer prescriptions, the model shows 0.71 of F-measure.

The qualitative analysis consisted of examining each slot prediction made by the Tri-CRF model for 20 manually chosen prescriptions in order to see the difficulties of the system in a finer granularity level. Out of 264 labels, the model failed to predict 18 labels. Two of the misinterpretations were due to the confusion between drug and inn labels. Other prediction errors were mainly due to formulations absent from the training data and confusion of temporal expressions.

V. DISCUSSION AND FURTHER WORK

Automatic analysis of natural language medical prescriptions has been investigated with some success from EHR [9], [12], [13]. However, this work reports on the more ambitious task of extracting medical prescription information for a voice-based PMS. This needs a far larger set of slots which in turn complexifies the extraction. To study this problem, we proposed a clear and exhaustive semantics for medical prescriptions as well as a method to deal with the paucity of available datasets. The experiments undertaken with state-of-the-art systems show that a Tri-CRF approach is the model which is able to benefit the most from a low amount of data to reach the greatest performances. This is in-line with recent work that showed that CRF is still the best model for the i2b2 task [12], [13]. Deep Neural Networks such as Att-RNN is competitive but suffer from the lack of available data.

It can be seen that the artificial generation of corpus enabled to learn sensible models to counterbalance the lack of data and the erratic distribution of slots. However, there are still many challenges to address.

The work reported in this paper was limited to two intents. However, there are many other intents that must be considered for medical prescriptions such as biological analysis demands, radiological examinations requests, etc. Moreover, in a dialogue setting, more intent related to the dialogue management must be considered such as confirmation, correction, advice, etc.

To address these challenges, future work includes the definition of a protocol to acquire real voice-based medical prescriptions in the context of a dialogue with clinicians. Such acquisition would enable to support the study and the development of dialogue systems for medical applications.

REFERENCES

[1] F. Lau, C. Kuziemsky, M. Price, and J. Gardner, "A review on systematic reviews of health information system studies," *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 637–645, 2010.

[2] P. Aspden, J. Wolcott, J. L. Bootman, and L. R. Cronenwett, "Committee on identifying and preventing medication errors: preventing medication errors," *Institute of Medicine National Academy Press, Washington, DC*, 2006.

[3] A. Agrawal, "Medication errors: prevention using information technology systems," *British journal of clinical pharmacology*, vol. 67, no. 6, pp. 681–686, 2009.

[4] Haute Autorité de Santé, "Référentiel de certification par essaie de type des logiciels d'aide à la prescription en médecine ambulatoire," Haute Autorité de Santé, Tech. Rep., september 2016.

[5] R. Altieri, F. Incardona, H. Kirkilis, and R. Ricci, "Mobi-dev: Mobile devices for healthcare applications," in *M-Health: Emerging Mobile Health Systems*, R. S. H. Istepanian, S. Laxminarayan, and C. S. Pattichis, Eds. Boston, MA: Springer US, 2006, pp. 163–175. [Online]. Available: https://doi.org/10.1007/0-387-26559-7_11

[6] T. Hamon and N. Grabar, "Linguistic approach for identification of medication names and related information in clinical narratives," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 549–554, 2010.

[7] C. Friedman, "A broad-coverage natural language processing system." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2000, p. 270.

[8] A. R. Aronson and F.-M. Lang, "An overview of metamap: historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.

[9] Ö. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 514–518, 2010.

[10] J. Patrick and M. Li, "A cascade approach to extracting medication events," in *Proceedings of the Australasian Language Technology Association Workshop 2009*, 2009, pp. 99–103.

[11] R. R. Slavescu, C. Maşca, and K. C. Slavescu, "Automatic extraction of structured information from drug descriptions," in *Mining Intelligence and Knowledge Exploration*, A. Groza and R. Prasath, Eds., 2018, pp. 21–31.

[12] C. Tao, M. Filannino, and Özlem Uzuner, "Prescription extraction using crfs and word embeddings," *Journal of Biomedical Informatics*, vol. 72, pp. 60 – 66, 2017.

[13] ——, "Fable: A semi-supervised prescription information extraction system," in *AMIA Annual Symposium proceedings*, 2018, pp. 1534–1543.

[14] C.-C. Lu, J. Leng, G. W. Cannon, X. Zhou, M. Egger, B. South, Z. Burningham, Q. Zeng, and B. C. Sauer, "The use of natural language processing on narrative medication schedules to compute average weekly dose," *Pharmacoepidemiology and drug safety*, vol. 25, no. 12, pp. 1414–1424, 2016.

[15] Ö. Uzuner, I. Solti, F. Xia, and E. Cadag, "Community annotation experiment for ground truth generation for the i2b2 medication challenge," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 519–523, 2010.

[16] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[17] L. Deléger, C. Grouin, and P. Zweigenbaum, "Extracting medication information from French clinical texts," in *MEDINFO 2010*, Amsterdam, 2010, pp. 949–953.

[18] S. V. Blackley, J. Huynh, L. Wang, Z. Korach, and L. Zhou, "Speech recognition for clinical documentation from 1990 to 2018: a systematic review," *Journal of the American Medical Informatics Association*, vol. 26, no. 4, pp. 324–338, 2019.

[19] S. Basma, B. Lord, L. M. Jacks, M. Rizk, and A. M. Scaranelo, "Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription," *American Journal of Roentgenology*, vol. 197, no. 4, pp. 923–927, 2011.

[20] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *International journal of medical informatics*, 2019.

[21] L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau *et al.*, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.

[22] T. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," *Journal of biomedical informatics*, vol. 39, no. 5, pp. 556–571, 2006.

[23] M. C. Dos Santos, F. Montyne, and C. Dhaen, "Medical natural language processing enhancing drug ordering and coding," in *M-Health: Emerging Mobile Health Systems*, R. S. H. Istepanian, S. Laxminarayan, and C. S. Pattichis, Eds. Boston, MA: Springer, 2006, pp. 147–161.

[24] G. Tur and R. De Mori, *Spoken Language Understanding Systems for Extracting Semantic Information from Speech*. Wiley, 2011.

[25] S. Liu, W. Ma, R. Moore, V. Ganesan, and S. Nelson, "Rxnorm: prescription for electronic drug information exchange," *IT professional*, vol. 7, no. 5, pp. 17–23, 2005.

[26] A. Khalili and B. Sedaghati, "Semantic medical prescriptions–towards intelligent and interoperable medical prescriptions," in *2013 IEEE Seventh International Conference on Semantic Computing*. IEEE, 2013, pp. 347–354.

[27] A. Grando, S. Farrish, C. Boyd, and A. Boxwala, "Ontological approach for safe and effective polypharmacy prescription," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 291.

[28] Y. Wang, L. Deng, and A. Acero, "Semantic frame-based spoken language understanding," in *Spoken language understanding: systems for extracting semantic information from speech*, G. Tur and R. De Mori, Eds. Wiley, 2011.

[29] M. Jeong and G. G. Lee, "Triangular-chain conditional random fields," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1287–1302, 2008.

[30] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, and others, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 3, pp. 530–539, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?id=2876380

[31] A. Bapna, G. Tur, D. Hakkani-Tur, and L. Heck, "Sequential dialogue context modeling for spoken language understanding," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017.

[32] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Interspeech 2016*, 2016.

[33] L. Huang, A. Sil, H. Ji, and R. Florian, "Improving slot filling performance with attentive neural networks on dependency structures," *arXiv:1707.01075 [cs]*, 2017. [Online]. Available: http://arxiv.org/abs/1707.01075

[34] Q. H. Tran, I. Zukerman, and G. Haffari, "A hierarchical neural model for learning sequences of dialogue acts," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain, 2017, pp. 428–437.

[35] M. Jeong and G. G. Lee, "Multi-domain spoken language understanding with transfer learning," *Speech Communication*, vol. 51, no. 5, pp. 412–424, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639309000028

[36] A. Mishakova, F. Portet, T. Desot, and M. Vacher, "Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes," in *The 1st International Workshop on Pervasive Computing and Spoken Dialogue Systems Technology (PerDial 2019)*, Kyoto, Japan, 2019.

[37] Éditions de Santé, Ed., *Le guide des premières ordonnances*, 2009th ed. Éditions de Santé, 2008.

[38] N. Serpollet, G. Bergounioux, A. Chesneau, and R. Walter, "A large reference corpus for spoken French: Eslo 1 and 2 and its variations," in *Proceedings from Corpus Linguistics Conference Series, University of Birmingham*, 2007.

[39] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 452–457.

[40] T. Desot, S. Raimondo, A. Mishakova, F. Portet, and M. Vacher, "Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments," in *21st International Conference on Text, Speech and Dialogue TSD 2018*, Brno, Czech Republic, 2018.

APPENDIX A
DOMAIN DEFINITION

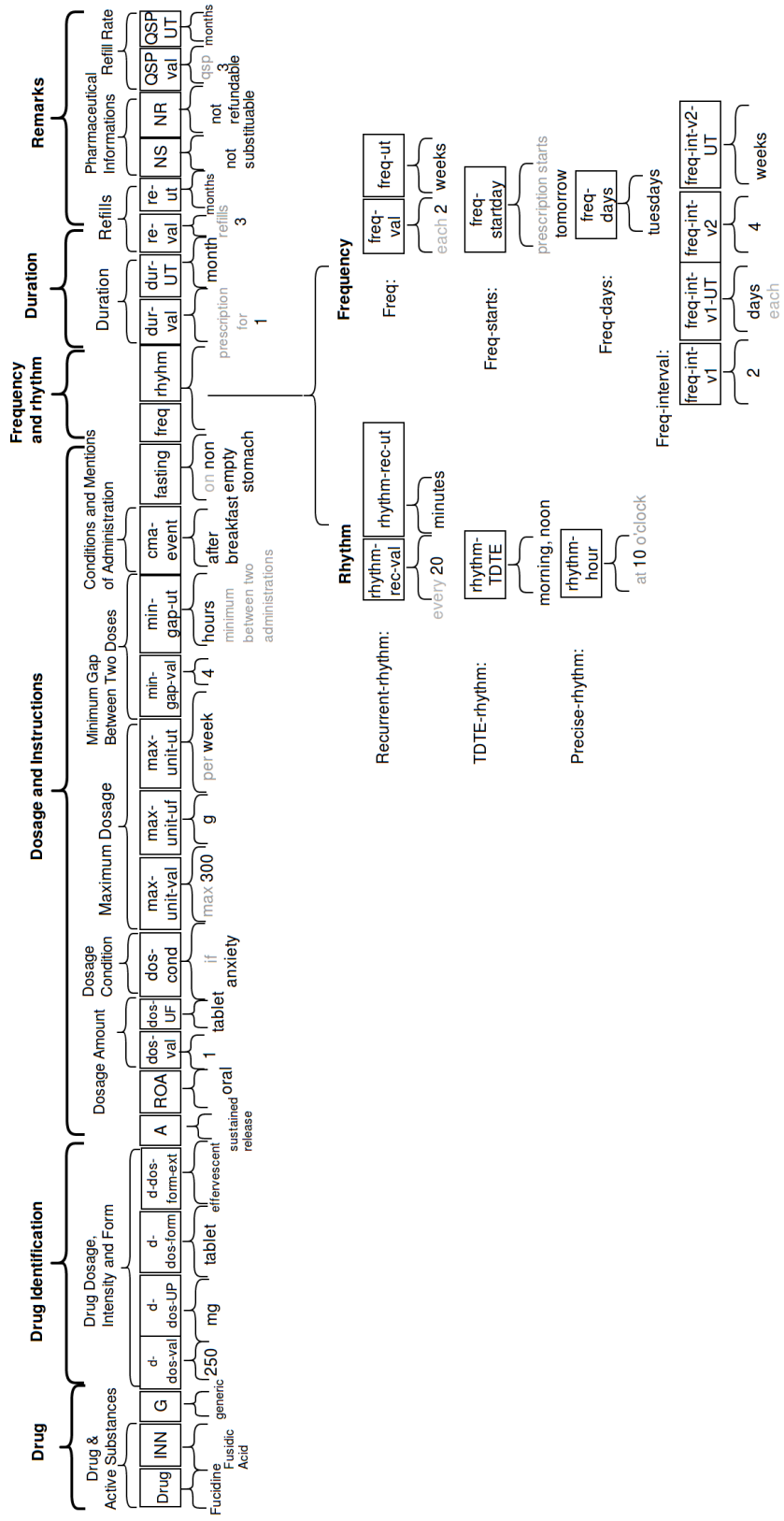

Fig. 4. Slot-label definitions of the prescription domain with example values