Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments

Thierry Desot, Stefania Raimondo, Anastasia. Mishakova, François Portet, Michel Vacher

Towards a French Smart-Home Voice Command Corpus: Design and NLU Experiments

Thierry Desot[1], Stefania Raimondo[1,2], Anastasia Mishakova[1], François Portet[1], and Michel Vacher[1]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] University of Toronto, Toronto, ON M5S 3H7, Canada

Abstract. Despite growing interest in smart-homes, semantically annotated *large* voice command corpora for Natural Language development (NLU) are scarce, especially for languages other than English. In this paper, we present an approach to generate customizable *synthetic* corpora of semantically-annotated French commands for a smart-home. This corpus was used to train three NLU models – a triangular CRF, an attention-based RNN and the Rasa framework – evaluated using a small corpus of real users interacting with a smart home. While the attention model performs best on another large French dataset, on the small smart home corpus the models vary performance across to intent, slot and slot value classification. To the best of our knowledge, no other French corpus of semantically annotated voice commands is currently publicly available

Keywords: natural language understanding, Corpora and Language Resources, Ambient Intelligence, Voice-user interface

1 Introduction

Smart-homes with integrated voice-user interfaces (VUI) can provide in-home assistance to aging individuals, allowing them to retain autonomy [1]. However, speech can only be effectively used to interact with a home automation system if its semantics are properly understood. Since users tend to deviate from a predefined set of voice commands [2,3,4], placing restrictions on their vocabulary and syntax is unrealistic and prohibitive. Instead, we must train a robust Natural Language Understanding (NLU) model on a well-balanced voice command corpus with user intent, slot label and slot value annotations. But, the removal of such constraints is a huge bottleneck for NLU and would necessitate a massive dataset.

In this paper, we present a customizable domain-specific corpus generator as an alternative to a large manually annotated data set. It can be developed quickly without the cost of manual semantic annotations, and is easily adaptable to new smart-home settings. For performance evaluation, a real smart-home corpus has been acquired from a limited set of users. This part is presented in Section 4. To validate the approach, three state-of-the-art NLU models were trained on the synthetic dataset and evaluated on the real smart-home dataset to show

how the trained models perform in realistic conditions. This part is presented in Section 5. The paper ends with a conclusion and an outlook on future work.

2 Related Work

While early slot-filling systems were rule-based [5], modern methods are data-driven. Conditional random fields [6], have recently been replaced by deep neural networks, including basic RNNs [7], Bi-directional LSTM RNN encoder-decoders [8], Attention-based RNNs [9] and Attention based CNNs [10]. 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 [10], template matching, used by the Sweet-Home system [11], or string matching as in [12]. Another approach is the extension of limited domain-specific corpora with synonyms and syntactic replacement [13]. [12] focuses on statistical decision-making using contextual information. While intent detection has traditionally been seen as a separate task from slot-filling [14], 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 [6], which extends linear sequence labeling CRF with a node to represent the dialogue act, and Att-RNN [9], which extends the slot-filling encoder-decoder RNN with an extra intent decoder. The Cassandra system [15] performs NLU via neural networks, using an LSTM for intent prediction and deep networks to identify slot locations and slot types. These simultaneous approaches are the most relevant to the work described here.

Since slots and intents are typically domain-dependent, new domains cannot benefit from models trained on massive, well-studied corpora. In response, some work has targeted cross-domain prediction [16,17], including the Tri-CRF model [16] mentioned above. In this work, we take a third approach: without a large domotic corpus as a starting point, we develop an artificial, automatically generated corpus to bootstrap our models as outlined in the following sections.

3 Method

3.1 Task, Intent and Slot Definition

Two main challenges for NLU in the smart-home environment are syntactic and linguistic variability; and underspecified commands. For the ambiguous command "turn on the fan", the NLU must identify the correct fan in the home based on the user's current location and activity. The NLU must also identify the same intent from a more syntactically complex utterance such as "can you turn on the fan". Similarly, "a bit more" following the command "raise the blinds a bit" must be *inferred* to be a request to repeat the previous action. This syntactic variability and underspecified commands make NLU development a daunting task. For the current version of our artificial corpus, we focused on understanding commands *without context* with one intent per utterance, while still tackling the issue of syntactic and linguistic variability.

Table 1. Examples of NLU annotated voice commands.

Sentence	Intent and slots
"are the lights upstairs on?"	CHECK_DEVICE_GROUP(DEVICE=light="lights", LOCATION-FLOOR=1="upstairs", device-setting=on="on")
"call the doctor"	CONTACT(PERSON-OCCUPATION=doctor="doctor")
"what time is it?"	GET_WORLD_PROPERTY(WORLD_PROPERTY=time="time")
"hey, can you hit the light?"	SET_DEVICE(ACTION=change="hit", DEVICE=light="the light")

The semantics of our artificial corpus were defined and developed around an existing smart-home Amiqual4Home (`https://amiqual4home.inria.fr`) as described in more detail in section 4. The resulting artificial corpus contains seventeen slot categories and eight intent classes. Intents are divided into four main categories: `contact` which allows a user to place a call; `set` to make changes to the state of objects in the smart-home; `get` to query the state of objects as well as properties of the world at large and `check` to check the state of an object.

The slot labels are divided into eight categories: the `action` to perform, the `device` to act on, the `location` of the device or action, the `person` or `organization` to be contacted, a device `component`, a device `setting` and the `property` of a location, device, or world. Table 1 provides representative examples of the annotated voice commands, used in a flat slot-filling approach. Different from previous work *slot-label prediction* is combined with *slot-value* prediction for passing the required information to the decision making unit.

4 Data

The semantics for intents and slots, defined in section 3.1, were used to automatically generate artificial data as well as to annotate a real dataset.

4.1 Artificial Corpus Generation

The core of the corpus generator is a feature-based generative grammar, built around an open-source NLTK python library. Feature-respecting top-down grammar generation was added. Unification functions limit feature propagation between rules to only those features which are explicitly specified in order to avoid conflicting features. The grammar defines intents (section 3.1) as a composition of their possible constituents, with fine-grained constraints on generation. For a rule that defines the slots of the intent `set_device` and can generate the command "open the door", the Slot_action has the feature ACTION whereas Slot_device has the feature ALLOWABLE_ACTION. Both those features are set to the same variable value which makes sure we only generate phrases with an allowable action to a particular device. Subsequent rules, contain other linguistic features such as gender and number agreement. Furthermore domain constraints are defined for object location in the smart-home. Unification of features disallows nonsensical utterances such as "light on the dishwasher in the bedroom".

Fig. 1. Instrumented kitchen.

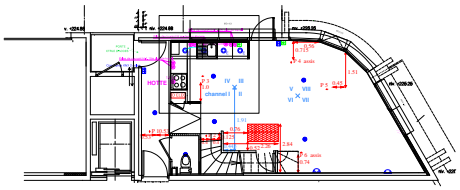

Fig. 2. Ground floor: kitchen and living room.

Furthermore, syntactical variation was added to the grammar rules, as for instance French interrogative constructions with the particle ("est-ce que"). The resulting vocabulary comprises 207 word types. Counting only lemmas, it contains 23 nouns denoting devices and 23 verbs denoting actions. The grammar generates about 28,000 phrases, each annotated with an intent and slots.

4.2 VocADom Real Dataset Acquisition

The real dataset was recorded with users interacting in realistic conditions in Amiqual4Home. This 87m^2 smart-home with a kitchen, living room, bedroom and bathroom, is equipped with home automation systems, multimedia devices, and microphone arrays (Fig. 1 and 2). A control room centralizes remote monitoring, recording of sensors and control of the home devices. Eleven participants uttered voice commands while performing scripted activities of daily living for about one hour of recording per participant. In the first half of the experiment, participants uttered unrestricted voice commands; in the second half, voice commands were restricted by a pre-defined grammar. Using a wizard-of-Oz approach, out-of-sight experimenters enacted user commands, acting as a 'perfect' NLU system. The VocADom corpus includes about twelve hours of audio signal and logs from the automation system. Speech was manually transcribed and 1,650 utterances, annotated with intents and slots were used as the validation dataset for our experiments. Sentences without intent class were excluded.

For comparison purposes, we make use of the Port-Media dataset [18] of French-language tourist information and ticket reservations for the 2010 Avignon music festival. It is of the same size as the artificial data and rich in terms of slot and value labels. The dataset contains natural utterances of 140 speakers in a simulated telephone booking task with slot and value label annotations. A comparison between the Port-Media corpus, the synthetic and VocADom datasets is provided in Table 2.

5 Experiment

To evaluate the synthetic smart-home corpus, we examine performance of state-of-the-art NLU models trained on the artificial corpus and tested on the real corpus. We chose a Triangular Conditional Random Field model (Tri-CRF), a

Table 2. NLU datasets

Dataset	# intents	# slots	# values	Avg. # values/slot	# utterances
Port-Media*	4	32	450	13.6 ± 21.3	20260
Synthetic data	8	17	57	3.35 ± 4.6	28000
VocADom	7	12	46	3.91 ± 3.7	4610

*4 intent classes were extracted based on the manually labeled slots in Port-Media

neural network with attention (Att-RNN), and one open-source commercial tool, Rasa, as a baseline. For comparison, we also evaluated performance of the models on the Port-Media dataset.

5.1 Tri-CRF, Att-RNN and Rasa-NLU Models

The Tri-CRF model from [6,16] 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 label y_t is also conditionally dependent on the previous label y_{t-1}. 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. 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.

The Attention RNN (Att-RNN) model from [9], is a recurrent encoder decoder architecture for simultaneous intent detection and slot labeling. 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 10,000 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).

Rasa NLU (`https://rasa.ai/products/rasa-nlu/`), an open-source tool for building NLU pipelines, is used as a baseline. Unlike Tri-CRF and the Att-RNN, Rasa 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. The used Rasa configuration is 'spacy_sklearn' with 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 embeddings are drawn from the spacy language model 'fr_depvec_web_lg', trained using word2vec on text data from Wikipedia, OpenSubtitles and Wikinews. The final vocabulary contains 1,184,651 words and the embeddings are vectors of length 300.

Table 3. Learning results on Port-Media and VocADom datasets

Model	Precision	Recall	F1
Att-RNN-Intent	**97.56**	**97.56**	**97.56**
Tri-CRF-Intent	96.42	96.43	96.36
Rasa-Intent	92.20	92.52	92.26
Att-RNN-Labels	**95.96**	**96.36**	**96.11**
Tri-CRF-Labels	95.31	95.74	95.39
Rasa-Labels	95.17	94.22	94.16
Att-RNN-Values	**94.85**	**95.73**	**95.08**
Tri-CRF-Values	92.01	93.49	92.32
Rasa-Values	93.94	93.73	93.34

(a) NLU performances on Port-Media development dataset.

Model	Precision	Recall	F1
Att-RNN-Intent	**93.77**	**90.28**	**91.30**
Tri-CRF-Intent	84.11	79.47	76.36
Rasa-Intent	90.48	71.39	76.57
Att-RNN-Labels	69.19	66.24	66.09
Tri-CRF-Labels	77.28	52.65	60.64
Rasa-Labels	**85.72**	**73.54**	**79.03**
Att-RNN-Values	43.02	30.51	35.00
Tri-CRF-Values	51.33	25.52	33.51
Rasa-Values	**68.56**	**56.73**	**61.95**

(b) NLU performances on VocADom dataset.

5.2 Results

Standard metrics were used for both intent classification and slot-labeling: precision, recall and F1-score. For slot-label and slot-value classification, metrics are calculated by comparing labels for words across all examples. The Port-Media and the artificial datasets are randomly divided into a training set of 90% and a development set of 10%.

Table 3.a reports results on the Port-Media dataset. The performance is only reported for the development set. Att-RNN provides the highest performances for the three tasks, with Tri-CRF competitive in all but the slot-value tasks. Both outperform Rasa. These results demonstrate the level of performance that can be achieved on real data, similar to real smart-home data. Accuracies on the artificial data development set are quasi-perfect, due to the very homogeneous nature of the synthetic corpus. Results of the artificial training data on the real VocADom validation dataset of 1,650 utterances are provided in Table 3.b. Overall performances on VocADom are worse than on Port-Media and particularly bad for slot-label and slot-value prediction. However, the high intent prediction accuracies on Port-Media are biased due to the high frequency of 'none' intents in the corpus, the low number of intent classes (4) and overlap between slots and intents. The results of slot-filling on VocADom are unsatisfactory. Errors are more randomly distributed over several categories and therefore more difficult to analyze. This is probably due to the significantly higher syntactic and lexical variation in the VocADom real dataset. Repetitions, disfluencies and interjections (ex."euh") result in utterances that are syntactically different from the artificial dataset. The 3-gram artificial language model perplexity on the real corpus is of 134 (without the $< s >$ tag). The OOV rate is also high, with 206 words not occurring in the artificial dataset.

Contrary to TRi-CRF and Att-RNN, Rasa performs well on slot labeling as it uses a word embedding layer which allows it to deal with the high number of OOV words indicating that artificial data generation benefits from external resources. Compared to results on the manually annotated Port-Media corpus, the poorer slot-filling results indicate that the automatic slot-label generation algorithm of the synthetic corpus can still be improved. Detailed results for Att-RNN predictions are given in Table 4.

Table 4. Detailed results for Att-RNN

Intent	F1
check_device	76.47
check_device_group	71.69
set_device	97.09
set_device_group	88.65
set_room_property	70.59

(a) intents

Slot	F1
action	62.03
device	84.06
device-setting	10.42
location-room	66.90
room-property	70.00

(b) slots

Value	F1
close	72.81
light	80.83
lower	17.53
open	67.51
turn off	41.51
turn on	68.07

(c) values

As hypothesized by [6] and [9], the joint approaches of the Tri-CRF and Att-RNN outperform Rasa's SVM-based intent classification on the VocADom dataset. This also shows that the synthetic corpus does contain enough information to train isolated intent models to be applied on real data.

6 Conclusion

In this paper, we address the lack of smart-home NLU training corpora by building a customizable automatic corpus generator for the smart-home domain. The corpus was evaluated by training two state-of-the-art models which were tested on a small but real smart-home dataset and compared to our baseline. Comparison of the models allowed us to pinpoint the artificial corpus or the models as main source of prediction errors. Both the Tri-CRF and the Att-RNN performed well on the large real Port-Media dataset and on the artificial voice command dataset. However lower performance on the small real smart-home dataset demonstrates difficulty handling its increased naturalness, vocabulary and syntactic variation. Both corpora are intended to be made available to the community. Future research aims to increase the level of naturalness of the generated corpus, using a joint approach inserting generic language models into the task-specific learning phase, taking into account context and history of commands (see [8]) and a simultaneous prediction of intents, slots and values. Such compound models have, at the time of writing, not been previously explored in the slot-filling literature and would be a useful and novel contribution to the field.

Acknowledgments

This work is part of the VocADom project founded by the French National Research Agency (Agence Nationale de la Recherche) / ANR-16-CE33-0006.

References

1. Portet, F., Vacher, M., Golanski, C., Roux, C., Meillon, B.: Design and evaluation of a smart home voice interface for the elderly — Acceptability and objection aspects. Personal and Ubiquitous Computing **17**(1) (2013) 127–144

2. Takahashi, S.y., Morimoto, T., Maeda, S., Tsuruta, N.: Dialogue experiment for elderly people in home health care system. In: Text, Speech and Dialogue, Berlin, Heidelberg (2003) 418–423
3. Möller, S., Gödde, F., Wolters, M.: Corpus analysis of spoken smart-home interactions with older users. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. (2008)
4. Vacher, M., Caffiau, S., Portet, F., Meillon, B., Roux, C., Elias, E., Lecouteux, B., Chahuara, P.: Evaluation of a context-aware voice interface for ambient assisted living: Qualitative user study vs. quantitative system evaluation. ACM Trans. Access. Comput. **7**(2) (2015) 5:1–5:36
5. Wang, Y., Deng, L., Acero, A.: Semantic frame-based spoken language understanding. In: Spoken language understanding: systems for extracting semantic information from speech. Wiley (2011)
6. Jeong, M., Lee, G.G.: Triangular-chain conditional random fields. IEEE Transactions on Audio, Speech, and Language Processing **16**(7) (2008) 1287–1302
7. Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., others: Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **23**(3) (2015) 530–539
8. Bapna, A., Tur, G., Hakkani-Tur, D., Heck, L.: Sequential dialogue context modeling for spoken language understanding. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. (2017)
9. Liu, B., Lane, I.: Attention-based recurrent neural network models for joint intent detection and slot filling. In: Interspeech. (2016) 685–689
10. Huang, L., Sil, A., Ji, H., Florian, R.: Improving slot filling performance with attentive neural networks on dependency structures. arXiv:1707.01075 [cs] (2017)
11. Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., Bonnefond, N.: The sweet-home speech and multimodal corpus for home automation interaction. In: The 9th edition of the Language Resources and Evaluation Conference (LREC). (2014) 4499–4506
12. Chahuara, P., Portet, F., Vacher, M.: Context-aware decision making under uncertainty for voice-based control of smart home. Expert Systems with Applications **75** (2017) 63–79
13. Manishina, E., Jabaian, B., Huet, S., Lefèvre, F.: Automatic corpus extension for data-driven natural language generation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016. (2016)
14. Tran, Q., Zukerman, I., Haffari, G.: A hierarchical neural model for learning sequences of dialogue acts. (2017) 428–437
15. Dumitrescu, S.D.: Cassandra smart-home system description. In: 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). (2017) 1–6
16. Jeong, M., Lee, G.G.: Multi-domain spoken language understanding with transfer learning. Speech Communication **51**(5) (2009) 412–424
17. Bapna, A., Tur, G., Hakkani-Tur, D., Heck, L.: Towards zero-shot frame semantic parsing for domain scaling. arXiv:1707.02363 [cs] (2017)
18. Lefèvre, F., Mostefa, D., Besacier, L., Estève, Y., Quignard, M., Camelin, N., Favre, B., Jabaian, B., Rojas-Barahona, L.M.: Leveraging study of robustness and portability of spoken language understanding systems across languages and domains: the PORTMEDIA corpora. In: LREC. (2012) 1436–1442