Dialogue History Integration into End-to-End Signal-to-Concept Spoken Language Understanding Systems

Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato de Mori, Yannick Estève

▶ **To cite this version:**

Natalia Tomashenko, Christian Raymond, Antoine Caubrière, Renato de Mori, Yannick Estève. Dialogue History Integration into End-to-End Signal-to-Concept Spoken Language Understanding Systems. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr 2020, Barcelona, Spain. pp.5, 10.1109/ICASSP40776.2020.9053247 . hal-02551760

HAL Id: hal-02551760
https://hal.science/hal-02551760v1

Submitted on 4 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:2002.06012v1 [cs.CL] 14 Feb 2020

DIALOGUE HISTORY INTEGRATION INTO END-TO-END SIGNAL-TO-CONCEPT SPOKEN LANGUAGE UNDERSTANDING SYSTEMS

Natalia Tomashenko[1], *Christian Raymond*[2], *Antoine Caubrière*[3,1], *Renato De Mori*[1,4], *Yannick Estève*[1]

[1]LIA - Avignon Université - France
[2]INSA Rennes/IRISA - Rennes, France
[3]LIUM - Le Mans Université - France
[4]McGill University - Montreal, Québec, Canada

ABSTRACT

This work investigates the embeddings for representing dialog history in spoken language understanding (SLU) systems. We focus on the scenario when the semantic information is extracted directly from the speech signal by means of a single end-to-end neural network model. We proposed to integrate dialogue history into an end-to-end signal-to-concept SLU system. The dialog history is represented in the form of dialog history embedding vectors (so-called *h-vectors*) and is provided as an additional information to end-to-end SLU models in order to improve the system performance. Three following types of h-vectors are proposed and experimentally evaluated in this paper: (1) *supervised-all* embeddings predicting bag-of-concepts expected in the answer of the user from the last dialog system response; (2) *supervised-freq* embeddings focusing on predicting only a selected set of semantic concept (corresponding to the most frequent errors in our experiments); and (3) *unsupervised* embeddings. Experiments on the MEDIA corpus for the semantic slot filling task demonstrate that the proposed h-vectors improve the model performance.

***Index Terms*—** End-to-end models, spoken language understanding (SLU), dialog history, h-vectors, semantic slot filling (SF)

1. INTRODUCTION

The task of spoken language understanding (SLU) system is to detect fragments of semantic knowledge in speech data. Popular models are made of frames describing relations between entities and their properties [1–3]. The SLU system instantiates a predefined set of frame structures called concepts that can be mentioned in a sentence or a dialogue turn. Concept mentions express dialogue acts (DA), intents, domain knowledge, and frame properties often represented by slots, identified by entity names, and slot filler values identified by mention types. Concept mentions are difficult to characterize in terms of words or characters. They may be localized by head words or short word sequences called concept supports. For example, word spans can be hypothesized to be mentions of concepts, while entire sentence can be considered for hypothesizing dialogue acts. Unfortunately, mentions may be ambiguous because their word spans may express more semantic constituents, be incomplete or be affected by errors of an automatic speech recognition (ASR) system. These difficulties can be alleviated by considering certain head words, word spans, or a sentence as a seed for hypotheses generation and using additional context for providing predictions useful for constraining instantiation decision. An example of additional distant context used so far is a representation of dialogue history made of embeddings of sentences preceding the sentence or dialogue turn to be interpreted [4–12].

A problem that has not yet thoroughly investigated is to select what to embed and how. Some popular corpora used so far (e.g. ATIS [13]) do not have explicit sentence history. In this case, the only context to pay attention to is the sentence to be interpreted. If some history information is available, then distant contexts for DA and concepts may be different. Specific contexts for DA have been proposed in [14, 15]. For concepts, the selection of distant contexts may depend on the complexity of the application semantic domain. For example, the French MEDIA corpus [16] has concepts of reference, relative time, locations, prices, logical conjunction and disjunction that are expressed by short semantically ambiguous words, which are often difficult to recognize, requiring knowledge of a semantic context called state of-the world to reduce the perplexity. Furthermore, the problem of deciding the type of embedding is also relevant as made evident in recent published papers [17–22].

In this paper, we investigate the use of different types of dialog history representation, extracted with or without supervision, and their impact on the performance of an end-to-end signal-to-concept neural network.

Noticeable approaches for reducing uncertainty in concept detection automatically extract relevant information from dialogue history [4, 5, 8]. Considering the concern expressed in [7] and prior knowledge, we propose to focus on types of history contents starting by considering the previous system turn that contains semantically unambiguous information. In fact, the sequence of words in the system turn is generated by a semantic model whose goal is to reach a commit state for performing a transaction. Furthermore, using the train set, it is possible to compute prediction probabilities of user enunciated concepts, given the system enunciated concepts. The most likely predicted concepts can thus be used for reducing interpretation uncertainty in the following user turn.

The rest of the paper is organized as follows. Section 2 presents an architecture of an end-to-end signal-to-concept model and the proposed way of integration of dialog history ebmeddings (to which we refer as *h-vectors*) into this model. Section 3 introduces different ways to represent the dialog history. Sections 4 describes the experimental setup and results. Finally, the conclusions are given in Section 5.

This work was supported by the French ANR Agency through the ONTRAC and AISSPER projects, under the contracts ANR-18- CE23-0021-01 and ANR-19-CE23-0004-01, and by the RFI Atlanstic2020 RAPACE project.

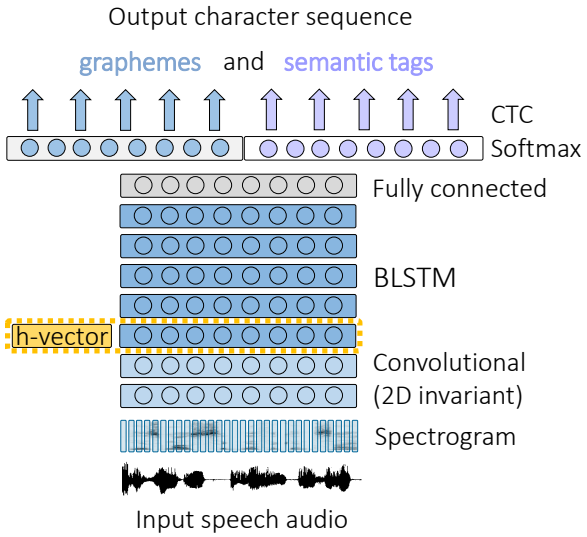


Fig. 1: End-to-end concept-to-semantic deep neural network model architecture. H-vectors represent dialog history embeddings vectors.

2. END-TO-END SIGNAL-TO-CONCEPT NEURAL ARCHITECTURE

Nowadays there is a growing research interest in end-to-end systems for various SLU tasks [23–31]. In this work, similarly to [26, 29], end-to-end training of signal-to-concept models is performed through the recurrent neural network (RNN) architecture and the connectionist temporal classification (CTC) loss function [32] as shown in Figure 1. A spectrogram of power normalized audio clips calculated on 20ms windows is used as the input features for the system. As shown in Figure 1, it is followed by 2D-invariant (in the time and-frequency domain) convolutional layers, and then BLSTM layers. A fully connected layer is applied after BLSTM layers, and the output layer of the neural network is a softmax layer. The model is trained using the CTC loss function. H-vectors are appended to the outputs of the last (second) convolutional layer, just before the first recurrent (BLSTM) layer.

The outputs of the network consist of the two subsets: (1) outputs to represent the words (graphemes of a corresponding language, a *space* symbol to denote word boundaries, and a *blank* symbol), and (2) outputs to represent semantic concepts types and a closing symbol for semantic tags. We have several symbols corresponding to semantic concepts (in the text these characters are situated before the beginning of a semantic concept, which can be a single word or a sequence of several words) and a one tag corresponding to the end of the semantic concept, which is the same for all semantic concepts.

In order to improve model performance, we integrate dialog history information in form of h-vectors into the model as shown in Figure 1. Each h-vector is calculated from the last dialog system response as described further in Section 3.

H-vectors are appended to the outputs of the last (second) convolutional layer, just before the first recurrent (BLSTM) layer. In this paper, for better initialization, we first train a model using *zero vectors* of the same dimension (all values are equal to 0) instead of h-vectors. Then, we use this pretrained model and finetune it on the same data but with the real h-vectors. This approach was inspired by [33], where the idea of using zero auxiliary features during pretraining was implemented for language models, and by [29], where it was used for i-vectors. In our preliminary experiments this type of pretraing demonstrated better results than direct model training with h-vectors, hence we use it in the experiments presented in this paper.

3. DIALOG HISTORY REPRESENTATION

The MEDIA corpus is a French corpus of spoken human/machine dialogues dedicated to hotel booking [16]. Recently, it has been shown that this corpus is one of the current most challenging corpora for slot filling (SF) task [34] due to its complexity. In this dataset, a human/machine dialogue is composed of 15 utterances from the user on average, and the same number from the system.

For this work, we decided to use as history information, the previous system prompt as it provides most of the time a good evidence of what the user answers. The goal is to help the main system to predict concept tags, hence our aim is to encode the previous system prompt into an embedding that contains useful information to achieve this objective.

3.1. Embedding with supervision

A first h-vector type is produced using a bidirectional gated recurrent unit (GRU [35]) network to analyse the system prompt and produce a vector of embedding that is the input to a decision layer whose objective is to predict the bag of concepts of the future user answer, illustrated in Figure 2a. The bag of concepts is represented by a vector whose size is the number of unique concepts (slots) in the application, the concepts that appears in the next user intervention are set to one. The output layer is thus a multiclass multi-output sigmoid layer and the network is trained using a binary cross-entropy loss. As the turn in the dialog itself may identify some useful statistics, a very short part (2%) of the h-vector is reserved to encode the dialog turn itself. Obviously, predicting the presence or absence of all the concepts of the next user answer from the previous system prompt is not possible and the network may overfit.

3.2. Embedding with no supervision

Another solution that is more straightforward to train is to use a recurrent autoencoder to encode the prompt into a single h-vector. This h-vector is obtained by a symmetric neural network using a forward GRU in the encoder and decoder part, the output is a softmax layer (size is the vocabulary of the system) whose objective is to reconstruct the input prompt, illustrated in Figure 2b.

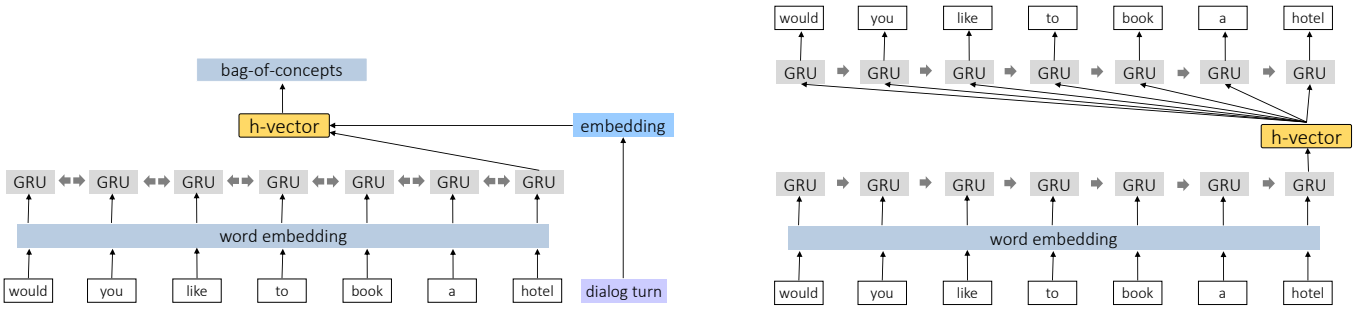
4. EXPERIMENTS

4.1. Data

Several publicly available corpora have been used for experiments (see Table 1).

Table 1: Corpus statistics for ASR and SF tasks.

Task	Corpora	Size, hours
ASR train	EPAC [36], ESTER 1,2 [37] ETAPE [38], REPERE [39] DECODA [40], MEDIA [16] PORTMEDIA [41]	404.6
SF train	MEDIA (train)	15.8
SF dev	MEDIA (dev)	1.6
SF test	MEDIA (test)	4.6



(a) Supervised (b) Unsupervised (autoencoder)

Fig. 2: Supervised and unsupervised architectures for h-vector extraction

4.1.1. ASR data

In this paper, the ASR data (audio speech files with text transcriptions) are used for transfer learning as described in Section 4.3. The corpus for ASR training is composed of corpora from various evaluation campaigns in the field of automatic speech processing for French, as shown in Table 1. The EPAC [36], ESTER 1,2 [37], ETAPE [38], REPERE [39] contain transcribed speech in French from TV and radio broadcasts. These data were originally in the microphone channel and for experiments in this paper were downsampled from 16kHz to 8kHz, since the test set for our main target task (SF) consists of telephone conversations. The DECODA [40] corpus is composed of dialogues from the call-center of the Paris transport authority. The MEDIA [16,42] and PORTMEDIA [41] are corpora of dialogues simulating a vocal tourist information server.

4.1.2. SF data

The MEDIA French corpus, dedicated to semantic extraction from speech in a context of human/machine dialogues, is used in the current experiments (see Table 1). The corpus has manual transcription and conceptual annotation of dialogues from 250 speakers. It is split into the following three parts [43]: (1) the training set (720 dialogues, 12K sentences), (2) the development set (79 dialogues, 1.3K sentences, and (3) the test set (200 dialogues, 3K sentences). A concept is defined by a label and a value, for example with the concept *date*, the value *2001/02/03* can be associated [16, 43]. The MEDIA corpus is related to the hotel booking domain, and its annotation contains 76 semantic concept tags: *room number, hotel name, location, date, room equipment*, etc.

4.2. H-vector extraction

We produced three different types of h-vectors: two types of h-vectors using the neural architecture trained in a supervised way to predict the bag of MEDIA concepts:

- *supervised-all h-vectors*. To extract these h-vectors, we trained a model as described in Section 3.1. The accuracy of the model to predict the next bag of concepts is 45% on the train and 26% on the test dataset. The model has 30.382 parameters.
- *supervised-freq h-vectors*. This version has been trained with a bag of the four history concepts that have been observed in the train and development set to predict concepts that are frequently misrecognized. This version tends to overfit with around 60% of accuracy on the train and only 16% on the test. The model has 23.918 parameters.

The third type of embeddings is trained in an unsupervised way:

- *unsupervised h-vectors*. These h-vectors are produced by the autoencoder architecture as described in Section 3.2. The autoencoder has 246.270 parameters, and the accuracy in the reconstruction is 52% on the train and 48% on the test.

Jointly trained word embedding is of size 10 while the dimension of h-vectors equals to 100 in all experiments. The described architectures for h-vectors were implemented using the *Keras* framework [44].

4.3. Signal-to-concept models

The neural architecture is inspired by the *Deep Speech 2* [45] for ASR. The two major differences in comparison with the original architecture are the following. First, we integrated dialog history into this system based on dialog history embedding vectors (*h-vectors*) as shown in Figure 1 and proposed in Section 3. Second, in this paper, the task is SF, therefore the output sequence besides the alphabetic characters also contains special characters corresponding to the semantic tags [26, 29].

A spectrogram of power normalized audio clips calculated on 20ms windows is used as the input features for the system. As shown in Figure 1, input features are spectrograms. They are followed by two 2D-invariant (in the time and-frequency domain) convolutional layers[1], and then by five 800-dimensional BLSTM layers with sequence-wise batch normalization. A fully connected layer is applied after BLSTM layers, and the output layer of the neural network is a softmax layer. The model is trained using the CTC loss function [32]. We used the *deepspeech.torch* implementation[2] for training baseline models, and our modification of this implementation to integrate dialog history embedding vectors.

In this work, we performed experiments with two types of models: (1) models that are trained directly on the target task using the MEDIA corpus dataset and (2) models that are trained using the transfer learning paradigm. Transfer learning is performed from the ASR task as described in [29].

[1] With parameters: kernel size=(41, 11), stride=(2, 2), padding=(20, 5)

[2] https://github.com/SeanNaren/deepspeech.pytorch

For transfer learning experiments, we first trained an ASR model on the ASR data (described in Section 4.1.1) using a similar end-to-end model architecture as we used for the SLU model. The difference is in the text data preparation and output targets. For training ASR systems, the output targets correspond to alphabetic characters and a *blank* symbol, while for slot filling task, we used additional targets corresponding to the semantic concept tags and one tag corresponding to the end of a concept. Then, we changed the softmax layer in this model by replacing the targets with the SF targets and continue training on the corpus annotated with semantic tags (Section 4.1.2).

4.4. Results

Performance was evaluated in terms of *concept error rate* (CER)[3] and *concept value error rate* (CVER)[4] on the MEDIA test dataset.

In the first series of experiments, we trained a baseline model and models with different types of h-vectors described in Section 4.2. Results for these models are given in Table 2. All the models in this table are trained directly on the MEDIA training corpus. The first line shows the baseline result for the end-to-end signal-to-concept model. The other three lines (#2,3,4) correspond to the models trained with dialog history integration and differ from each other in the way the dialog history is represented in the form of h-vectors. We can observe, that all types of h-vectors provide an improvement over the baseline model for both metrics CER and CVER. The best result (line #4) is obtained for *supervised-all* h-vectors and corresponds to 12.5% of relative CER reduction and to 11.9% of CVER reduction in comparison with the baseline model.

Table 2: Slot filling performance results on the MEDIA test dataset for the baseline model and models trained with different types of dialog history embedding vectors. Results are given in terms of CER and CVER metrics (%); ΔCER and ΔCVER (%) denote relative error reduction for CER and CVER correspondingly in comparison with the baseline model (#1).

#	h-vector type	CER	ΔCER	CVER	ΔCVER
1	no (baseline)	39.2	-	53.0	-
2	unsupervised	35.8	8.7	47.6	10.2
3	supervised-freq	35.9	8.4	48.2	9.1
4	supervised-all	**34.3**	**12.5**	**46.7**	**11.9**

It was shown in [29], that transfer learning can significantly improve the performance of end-to-end SLU models. In this work, we are also interested in exploring the proposed approach for more accurate models trained using the transfer learning paradigm. For this purpose, we trained two models using transfer learning from the ASR task as proposed in [29] and described in Section 4.3. Results for these models are presented in Table 3. The first line corresponds to a baseline model. The second line demonstrates the result for the model trained with the best type of dialog history embedding vectors (*supervised-all*) chosen according to our first series of experiments. We can see that h-vectors continue to provide an improvement in performance over the stronger baseline: 7.7% of relative CER reduction and 6.3% of relative CVER reduction.

[3]CER is defined as the ratio of the total number of deleted, inserted and confused concepts and the total number of concepts in reference utterances.

[4]CVER, in comparison to CER, takes into account concept/value pairs instead of only concepts.

Table 3: Slot filling performance results on the MEDIA test dataset for the baseline model and the best model trained with *supervised-dall* type of dialog history embedding vectors. Models for SF are trained using **transfer learning** from an ASR model.

#	h-vector type	CER	ΔCER	CVER	ΔCVER
1	no (baseline)	23.5	-	30.0	-
2	supervised-all	**21.7**	**7.7**	**28.1**	**6.3**

5. CONCLUSIONS

In this paper, we have proposed a novel way of integration of the dialog history information into end-to-end signal-to-concept SLU models by means of using so-called *h-vectors*. We have proposed different types of h-vectors and investigated their effectiveness for end-to-end SLU using as an example the semantic slot filling task. Experiments on the MEDIA corpus demonstrated that using h-vectors improves the slot filling model performance by about 813% of relative CER reduction, and by about 6-12% of relative CVER reduction. The best result was obtained using *supervised-all* h-vectors predicting bag-of-concepts representations of the user's answer from the last system response.

6. REFERENCES

[1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.

[2] L. Shen, Y. Feng, and H. Zhan, "Modeling semantic relationship in multi-turn conversations with hierarchical latent variables," in *ACL*, 2019.

[3] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental transformer with deliberation decoder for document grounded conversations," in *ACL*, 2019.

[4] Y.-N. Chen, D. Hakkani-Tür, G. Tür, J. Gao, and L. Deng, "End-to-end memory networks with knowledge carryover for multi-turn spoken language understanding.," in *Interspeech*, 2016.

[5] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *NAACL-HLT*, 2018.

[6] Z. Zhao, S. Zhu, and K. Yu, "A hierarchical decoding model for spoken language understanding from unaligned data," in *ICASSP*, 2019.

[7] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio, "Do neural dialog systems use the conversation history effectively? an empirical study," in *ACL*, 2019.

[8] R. Goel, S. Paul, and D. Hakkani-Tür, "Hyst: A hybrid approach for flexible and accurate dialogue state tracking," in *Interspeech*, 2019.

[9] V. Vukotic, C. Raymond, and G. Gravier, "A step beyond local observations with a dialog aware bidirectional GRU network for Spoken Language Understanding," in *Interspeech*, 2016.

[10] M. Henaff, J. Weston, A. Szlam, A. Bordes, and Y. LeCun, "Tracking the world state with recurrent entity networks," *arXiv preprint arXiv:1612.03969*, 2016.

[11] M. Korpusik and J. Glass, "Dialogue state tracking with convolutional semantic taggers," in *ICASSP*, 2019, pp. 7220–7224.

[12] H. Lee, J. Lee, and T.-Y. Kim, "Sumbt: Slot-utterance matching for universal and scalable belief tracking," *arXiv preprint arXiv:1907.07421*, 2019.

[13] D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the ATIS task: The ATIS-3 corpus," in *Workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 43–48.

[14] Y. Liu, K. Han, Z. Tan, and Y. Lei, "Using context information for dialog act classification in dnn framework," in *EMNLP*, 2017.

[15] D. Ortega, C.-Y. Li, G. Vallejo, P. Denisov, and N. T. Vu, "Context-aware neural-based dialog act classification on automatically generated transcriptions," in *ICASSP*, 2019.

[16] L. Devillers, H. Maynard, S. Rosset, P. Paroubek, K. McTait, D. Mostefa, K. Choukri, L. Charnay, C. Bousquet, N. Vigouroux, et al., "The French MEDIA/EVALDA project: the evaluation of the understanding capability of spoken language dialogue systems.," in *LREC*, 2004.

[17] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *NAACL-HLT*, 2016.

[18] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017.

[19] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of NAACL-HLT*, 2018, pp. 2227–2237.

[20] Z. Yin and Y. Shen, "On the dimensionality of word embedding," in *Advances in Neural Information Processing Systems*, 2018, pp. 887–898.

[21] X. Zhang, Y. Li, D. Shen, and L. Carin, "Diffusion maps for textual network embedding," in *NIPS*, 2018.

[22] Y. Yaghoobzadeh, K. Kann, T. J. Hazen, E. Agirre, and H. Schütze, "Probing for semantic classes: Diagnosing the meaning content of word embeddings," in *ACL*, 2019.

[23] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring ASR-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *ASRU*, 2017.

[24] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," *arXiv preprint arXiv:1809.09190*, 2018.

[25] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, "Towards end-to-end spoken language understanding," *arXiv preprint arXiv:1802.08395*, 2018.

[26] S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, and E. Morin, "End-to-end named entity and semantic concept extraction from speech," in *SLT*, 2018, pp. 692–699.

[27] Y.-P. Chen, R. Price, and S. Bangalore, "Spoken language understanding without speech recognition," in *ICASSP*, 2018.

[28] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," in *Interspeech*, 2019.

[29] N. Tomashenko, A. Caubrière, and Y. Estève, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," *Interspeech*, 2019.

[30] N. Tomashenko, A. Caubriere, Y. Esteve, A. Laurent, and E. Morin, "Recent advances in end-to-end spoken language understanding," in *SLSP*, 2019.

[31] A. Caubrière, N. Tomashenko, A. Laurent, E. Morin, N. Camelin, and Y. Estève, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," in *Interspeech 2019*, pp. 1198–1202.

[32] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[33] S. Deena, R. W. Ng, P. Madhyashta, L. Specia, and T. Hain, "Semi-supervised adaptation of rnnlms by fine-tuning with domain-specific auxiliary features," in *Interspeech*. ISCA, 2017, pp. 2715–2719.

[34] F. Béchet and C. Raymond, "Benchmarking benchmarks: introducing new automatic indicators for benchmarking Spoken Language Understanding corpora," in *Interspeech*, Sept. 2019.

[35] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[36] Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news.," in *LREC*, 2010.

[37] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[38] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *LREC*, 2012.

[39] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE corpus: a multimodal corpus for person recognition.," in *LREC*, 2012, pp. 1102–1107.

[40] F. Bechet, B. Maza, N. Bigouroux, T. Bazillon, M. El-Beze, R. De Mori, and E. Arbillot, "DECODA: a call-centre human-human spoken conversation corpus.," in *LREC*, 2012, pp. 1343–1347.

[41] F. Lefèvre, D. Mostefa, L. Besacier, Y. Estève, M. Quignard, N. Camelin, B. Favre, B. Jabaian, and L. Rojas-Barahona, "Robustness and portability of spoken language understanding systems among languages and domains: the PortMedia project [in French]," in *JEP-TALN-RECITAL*, 2012, pp. 779–786.

[42] H. Bonneau-Maynard, C. Ayache, F. Bechet, A. Denis, A. Kuhn, F. Lefèvre, D. Mostefa, M. Quignard, S. Rosset, C. Servan, et al., "Results of the French Evalda-Media evaluation campaign for literal understanding," in *LREC*, 2006.

[43] V. Vukotic, C. Raymond, and G. Gravier, "Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?," in *Interspeech*, 2015.

[44] F. Chollet, "Keras, https://github.com/fchollet/keras," 2015.

[45] Amodei et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *ICML*, 2016.