ERROR ANALYSIS APPLIED TO END-TO-END SPOKEN LANGUAGE UNDERSTANDING

Antoine Caubrière, Sahar Ghannay, Natalia Tomashenko,
Renato De Mori, Antoine Laurent, Emmanuel Morin, Yannick Estève

LIUM - Le Mans Université, Université Paris-Saclay, CNRS, LIMSI, 91400, Orsay, France,
LIA - Avignon Université, LS2N - CNRS/Université de Nantes

ABSTRACT

This paper presents a qualitative study of errors produced by an end-to-end spoken language understanding (SLU) system (speech signal to concepts) that reaches state of the art performance. Different studies are proposed to better understand the weaknesses of such systems: comparison to a classical pipeline SLU system, a study on the cause of concept deletions (the most frequent error), observation of a problem in the capability of the end-to-end SLU system to segment correctly concepts, analysis of the system behavior to process unseen concept/value pairs, analysis of the benefit of the curriculum-based transfer learning approach. Last, we proposed a way to compute embeddings of sub-sequences that seem to contain relevant information for future work.

***Index Terms*—** Spoken language understanding, end-to-end system, error analysis, neural network

1. INTRODUCTION

Despite recent progress, spoken language understanding (SLU) systems make a significant amount of errors in some tasks even with sophisticated end-to-end (E2E) neural architectures. Limited effort has been made so far for making a systematic analysis of these errors, probably because explaining the cause of them appears to be difficult. It would be useful to use prior knowledge to associate errors groups with observations that may explain them. For example, it would be useful to know how deletions are due to the fact that clue words defining concepts in dictionaries have been uttered but not recognized or ambiguities of clue words or word spans annotated with the mention of a concept have been correctly hypothesized, but the SLU system has not been able to focus on distant context relevant for reducing the ambiguity of the hypothesized words. Explaining these errors may suggest modifications on neural architecture components for error analysis. Useful prior knowledge can be the semantic model of an application, the difficulty of representing words uttered with very few phonemes not well distinguished by internal latent representations or the erroneous detection of the boundaries of concept mentions. Methods for sentence, grammatical and semantic error correction applied to text documents can be found in [1–3]. ASR error adaptation for SLU has been proposed in [4]. Error analysis for ASR is discussed in [5]. In [6], a SLU improvements have been presented by managing error reduction based on agreement of different SLU components.

First neural end-to-end SLU systems, that directly extract semantic concept from speech audio signal appeared in 2018 [7–9]. Very recently, such approaches reached state of the art results [10], similar to the results got by pipeline (PIP) approaches that apply sequential processes to extract semantic information from speech signal: automatic speech recognition (ASR), automatic enrichment of ASR outputs (part of speech tagging, chunking, dependency labeling…, and at last natural language understanding (NLU) process applied on enriched ASR outputs.

In this study, we analyze the errors made by an state-of-the art end-to-end system. By understanding its main weaknesses, we expect to discover how to continue improving the performance of such approaches.

2. SYSTEM DESCRIPTION

The end-to-end SLU system used for this work is the same as the one used in [10, 11]. Its architecture is very closed to the DeepSpeech2 architecture [12]. It consists of a stack of two 2D-invariant convolutional layers (CNN), five bidirectional long short term memory layers (bLSTM) with sequence-wise batch normalisation and a final softmax layer.

This system is trained with the Connectionist Temporal Classification (CTC) loss function [13]. This function allows the system to learn an alignment between an audio input and a character sequence to produce. Input features are sequences of log-spectrograms of power normalized audio clips calculated on 20ms windows. Output sequences consist of a sequence of characters composed of word and semantics concepts. Semantics concepts are represented by starting tags and ending tags before and after words supporting these concepts.

Starting tags defines the nature of concept while ending tag will only close an opened tag. We use several starting tags, one for each semantic concept, but only one ending tag. For example, the sentence "I would like two double-bed rooms" is represented with its semantic information as "I would like <**nb_room** two > <**room_type** double-bed rooms >". In this example, <**nb_room** and <**room_type** are two starting tags defining respectively the semantics concepts "number of room" and "room type". The '>' symbol represents the unique closing tags. Notice that starting and ending tags are actually represented by a single character within the character sequence produced by the neural network. Previous example become "I would like **ł** two > **ø** double-bed rooms >", where '**ł**' is "<**nb_room**" and '**ø**' is "<**room_type**".

In addition, we use the same star mode as presented in [10, 14]. This mode allows the CTC loss function to be more sensitive on concepts and their values instead of unlabelled words. It consists of replacing all the characters between two concepts by a single star. Previous example become "* **ł** two > **ø** double-bed rooms >".

Our end-to-end SLU system is trained following the curriculum-based transfer learning approach we proposed in [10]. It consists on training the same model successively with different tasks following a curriculum strategy. To respect this strategy, tasks are ordered from the most generic one to the most specific one: speech recog-

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.

nition (ASR), then named entity recognition (NER) and finally semantic concept extraction (SLU). Named entity recognition task is trained following the same way as the semantic concept extraction task. We add boundaries of named entity concepts inside the character sequences to be produced. We apply transfer learning between each task and keep all the parameters of the produced model of the current training step as initialization of the next training step, except the top layer (softmax). Parameters of this layer are fully reseted because of the change of output labels at each training step. Thanks to this strategy, our end-to-end models reached state-of-the art performance. More details are described in [10].

3. MEDIA CORPUS

The MEDIA corpus is a French dataset of audio recordings with manual annotations, dedicated to semantic extraction from speech in a context of human/machine dialogues. The corpus has manual transcription and semantic annotation of dialogues from 250 speakers. It is split into the following three parts [15]: (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 the value *2001/02/03* can be associated to the concept *date* [15–17]. 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. Some concept value pairs appearing in turns of the test set may not appear in the train set, requiring generalizations that the SLU architecture may not be able to perform. Thus problem has been investigated in [17].

4. ERROR ANALYSIS

To start the analysis on the MEDIA data, we propose to compare the global distribution of errors produced by two state-of-the-art SLU systems. The first one is built following a classical pipeline approach while the second system is the end-to-end approach presented above. As described in [10], the pipeline approach consists of a component chain composed of a speech recognition (ASR) component, diverse natural language processing components in order to enrich the ASR outputs with linguistic information (part-of-speech, chunking, governor words...), and natural language understanding component, based on a Condition Random Field model, that labels the enriched ASR outputs with semantic labels.

Error distribution of the pipeline system on the MEDIA development dataset is provided on figure 1. For greater clarity, we have only kept the 30 concepts with the highest number of errors.

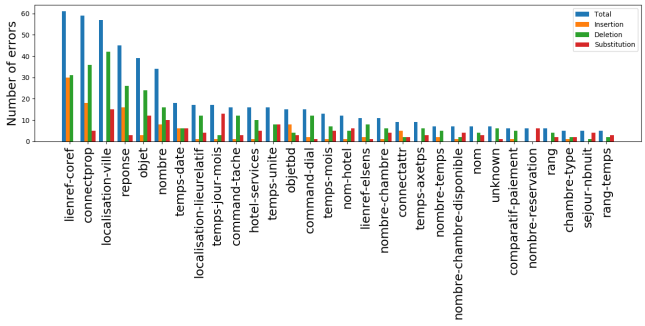

Fig. 1. Error distribution on MEDIA dev dataset for PIP approach

The figure 1 shows us that for most of the concepts, the major error type is deletion. The five concepts with most errors are "lienref-coref" (that represents coreference word that refers to a previous entity), "connectProp" (that is a word that connect two properties of the same concept frame), "localisation-ville" (city location), "reponse" (response) and "objet (object)". Error distribution of the end-to-end system on the same dataset is provided on figure 2. We also have only kept the 30 concepts with the highest number of errors.

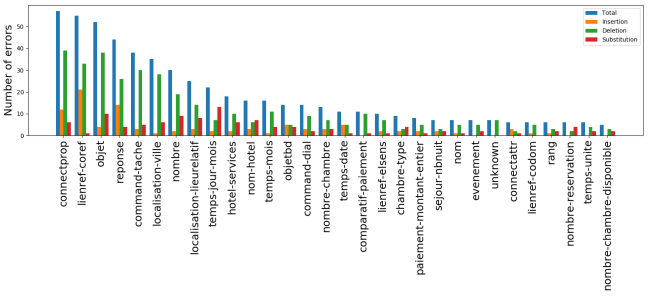

Fig. 2. Error distribution on MEDIA dev dataset for E2E approach

As for the pipeline error distribution, we can observe that deletion is also the most represented error. The five concepts with most errors are "connectProp", "lienref-coref", "objet", "reponse" and "command-tache" (task command). Among the five top erroneous concepts, we notice that four ones are shared by both end-to-end and pipeline approaches, while *city location* seems less problematic for end-to-end system than *task command* that is better handled by the pipeline system. These error distributions appear to be similar and show frequent errors on domain independent concepts corresponding to logic operators such as AND, or reference mentions such as IT or ordinal numbers. Furthermore, these concepts are mentioned by a single word that is frequently used for connecting items that are not relevant for the application semantic domain. A further analysis of the in-domain mentions of these concepts shows that the most relevant context for reducing the detection uncertainty is made of mention of specific domain concepts. This suggest that it is worth considering, in future work, the introduction of an additional component that performs a sort of specific island-driven semantic parsing.

4.1. Transcription problem

In this section, we investigate the potential causes of the high number of concept deletions, especially on the end-to-end system.

We focus this study on the concepts with highest number of errors on the MEDIA dev dataset. They are "connectProp", "lienref-coref" and "objet". The "connectProp" concept is "logical connection" whose most frequent mention is a function word *et* (and in English) uttered only by the vowel phoneme, and is annotated only for connections between application domain concept mentions. The "lienref-coref" concept is the reference to a coreferent that could be in the dialogue history. Its most frequent mentions are function words such as "it" or other short spans such as "the first one", difficult to segment or to detect. The "objet" concept is relevant only if related to the mention of an application domain concept and required a context difficult to characterize for reducing ambiguities.

The analysis reveals that there are three major cases when a deletion occurs:

1. automatic transcription is OK, meaning that the end-to-end system succeeded in recognizing the word supporting the concept, while the concept was not detecting;
2. automatic transcription is wrong, so the concept cannot be detected;
3. automatic transcription is OK, but the word supporting the concept has been nested in another contiguous concept.

Table 1. Most frequent deletions errors on the MEDIA dev dataset

Focused concept	Nb Deletion	Correct ASR	Wrong ASR	Nested
connectProp	39	**28**	6	5
lienref-coref	33	**19**	10	4
objet	38	**31**	4	3

As we can see in table 1, a major part of the time a deletion occurs even if the system produces a good transcription. Deletion of concepts are not mainly caused by the speech transcription capability of our end-to-end system. It is a noticeable result since for "connectProp" and "lienref-coref" support words are very small ones with very few phonemes, like "*et*" (and) or "*il*" (it). These results show that the major part of the errors comes from semantic labeling problem.

As a complement, in the sequences produced by the system for these deletion, we regularly observe the presence of ending tags without any associated starting tags. For instance for the 39 deletion errors of the "connectProp" concepts in the development corpus, we observe 11 cases with an ending tag without starting tag. This represents more than 28 % of the deleted concepts. We propose in the next section an analysis of this segmentation problem.

4.2. Segmentation problem

To tackle the concept segmentation problem, we propose to learn the segmentation as another task added inside the curriculum-based transfer learning approach. We replace each starting tag by a unique '<' inside the sequence to be produced by the system for the MEDIA task. The segmentation task is learned before the training on the final MEDIA task. So, the curriculum-based transfer learning approach consists of a learning chain of 4 tasks following this order: ASR, NER, SLU segmentation, and SLU. To augment our SLU data we use two different corpora for this tasks. MEDIA which is our target task and PORTMEDIA which is a corpus dedicated to theater ticket booking close to the MEDIA task. By the use of these two different corpora, we split the SLU task into two training steps, first step with PORTMEDIA and MEDIA together (PM+M) and second step with only MEDIA dataset (M). The segmentation task is applied only for the final MEDIA task and placed between the two SLU training steps. Final learning chain consists on five different training following this order: "$ASR \bullet NER \bullet PM + M \bullet M_{seg} \bullet M$". Another learning chain consists of applying the star mode for the final MEDIA task.

For these experiments, we used two different metrics Concept Error Rate (CER) and Concept/Value Error Rate (CVER) for evaluation. CER is a metrics similar to the word error rate metrics but applied on concepts. CVER is very close to CER but evaluates concept/value pairs instead of evaluating only concepts. Results with a greedy decoder on the MEDIA test set are reported in table 2. The training chain 1 and 3 are reported from previous work in [10].

Table 2. Comparison between SLU systems with and without a segmentation task in greedy decoding on the MEDIA test

Training chain	CER	CVER
1. $ASR \bullet NER \bullet PM + M \bullet M$	21.6	27.7
2. $ASR \bullet NER \bullet PM + M \bullet M_{seg} \bullet M$	20.7	27.2
3. $ASR \bullet NER \bullet PM + M \bullet M\star$	20.1	27.2
4. $ASR \bullet NER \bullet PM + M \bullet M_{seg} \bullet M\star$	20.8	27.7

These results show improvement by the learns of the segmentation as an intermediate task by reducing the CER from 21.6 to 20.7 in the case of a training chain without star mode.

In our previous work [10], we reach the best result by the use of a star mode which allows the CTC loss function to be more focus on concepts and their values instead of unlabelled words. It consists of replacing all the characters between two concepts by a single star. Example of the section 2 become "* **ł** two > **ø** double-bed rooms >". We applied this mode on the MEDIA task after the segmentation task training. Results show that we cannot reach an improvement by the learns of the segmentation before the use of the star mode. We also notice that the integration of the segmentation task into the curriculum-based transfer learning approach allows us to get the same CVER value (27.2%) with the normal mode (that produces both full transcription and semantic tags) as with the star mode (that produced semantic tags and only transcription of concept value).

4.3. Unseen Concept/Value pairs

In this section, we propose to analyze the behavior of our system to process Unseen Concept/Value pairs (UCV). We define the UCV as the Concept/Value pairs seen in the MEDIA development dataset which do not appear in the training dataset. On the development dataset, there is a total of 533 UCV. Table 3 reports the number of UCV correctly retrieved by the end-to-end system with and without the use of the NER task in the curriculum-based transfer learning approach. For both system, we report the number of correct value.

Table 3. UCV correctness on MEDIA dev dataset. CV means Concepts/Values pairs are corrects and V means Values only are corrects

System	correct CV	correct V
$ASR \bullet NER \bullet PM + M \bullet M$	132	38
$ASR \bullet PM + M \bullet M$	124	36

For both systems, we can observe a small number of correct values. As there are 533 UCV in the MEDIA development dataset and the well recognized UCV represent around a quarter of the total UCV, that means that speech transcription is wrong for a major part of the UCV.

We also notice that the use of the NER task during the training yields an improvement of 6% of relative gain. In order to get a more precise idea of the contribution of the use of the NER task during the training, we measured the evolution of the number of errors for each concept, with or without the use of the NER task during the training. Figure 3 shows this evolution. Globally, this evolution is strongly positive for a lot of concepts, but sometimes with a negative impact.

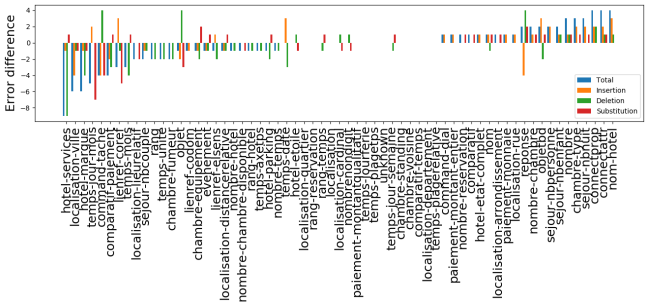

Fig. 3. Error difference by the use of NER task in the curriculum-based transfer learning approach

5. EMBEDDINGS ANALYSIS

To give more insight to the error analysis we propose to perform a visual evaluation of concept embeddings by computing the t-SNE representations.

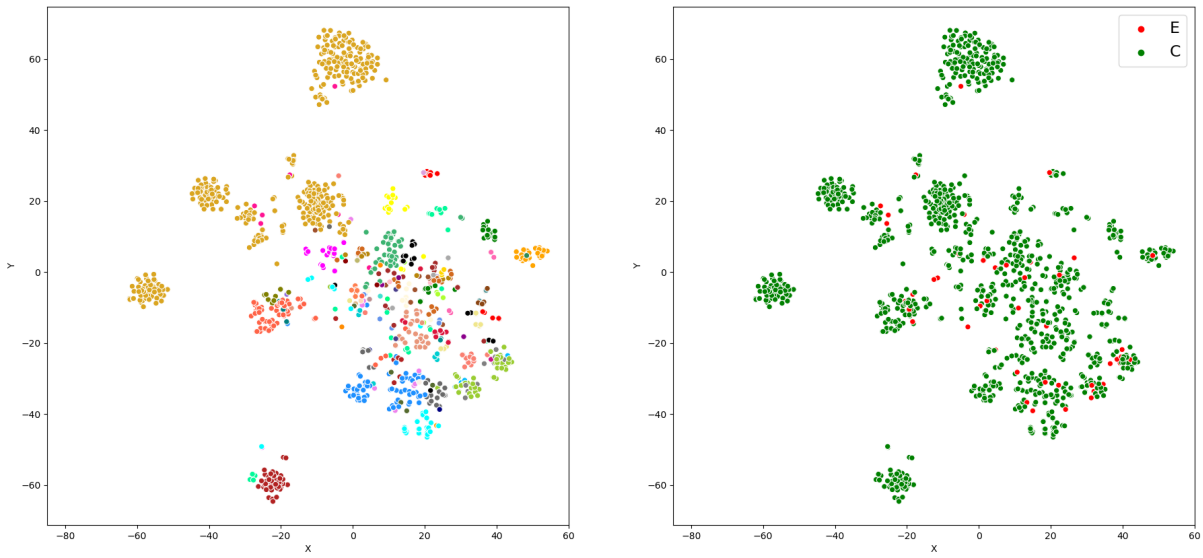

Fig. 4. t-SNE representations of concept projection with colored dots in function of the recognized semantic class (left) versus the same with red/green colors that correspond to Error/Correct concepts (right).

5.1. Frames to words and concepts: embedding extraction

To accomplish the visual evaluation, we have to extract the word and concept embeddings from the end-to-end SLU system. The embeddings are extracted from the last bLSTM layer, that corresponds to sequential representations of the input frames. To obtain a representation for each word and concept we were based on CTC outputs to know the number of frames for each word and concept. Hence, the word or concept embeddings correspond to the sum of its frame embeddings. For example : if the CTC output of the recognized word "bonjour" corresponds to "bbb@@@oon@jjourrr" (17 frames), this word is represented by the sum of its 17 frame embeddings.

5.2. Visualization

Figure 4 illustrates two t-SNE representations of the concept embeddings extracted from Dev dataset. The first representation(left) corresponds to the projection with colored dots of the different recognized semantic classes (concepts). It shows that most of the concepts of the same class are compact and clustered in the continuous space. The second one (right), illustrates the projection of the different recognized semantic classes with red/green colors, that corresponds to Error/Correct concepts. Note that the annotation of the concepts to Error/Correct is computed based on the alignment with the reference transcriptions. From the second representation we observe that the main errors (red dots) made by our end-to-end SLU are located in the areas where the concepts are mixed without any structure in the continuous space.

We consider this observation important for future work. Three axes would be explored:

1. how to take benefit of this semantic representation cartography to improve the performance?
2. how to force the system to represent concepts in a more relevantly structured space, for instance by injecting some *a priori* knowledge in the training process.
3. how to take benefit of this cartography to auto-detect uncertainty and errors? Indeed, it seems that the position in the continuous space of the embedding provides relevant information. Notice that these embedding are computed on variable sub-sequences of frame embedding, while currently the neural network takes decision frame by frame: these embeddings bring another representation of information not used as is by the model.

6. CONCLUSION

This paper presents a qualitative study of errors produced by an end-to-end SLU system (speech signal to concepts) that reaches state of the art performance. We made a comparison to a classical pipeline SLU system about the error distribution among concepts. Also, a study on the cause of deletions (the main nature of errors) showed that the problem does not come from the speech recognition capability of the system. We detected a problem in the capability of the end-to-end SLU system to segment correctly concepts and proposed a way to attenuate this during the training. Then, We analyzed the behavior of the system to process unseen concept/value pairs and confirmed the interest of using transfer learning from the named entity recognition task to address this issue. Last, we proposed a way to compute embeddings of sub-sequences that seem to contain relevant information for future work. The error analysis suggests that concepts requiring specific attention are generic and not domain dependent, while their relevance may be domain dependent. In fact application domains as MEDIA may have connector instances relating domain semantic contents, while other instances of the same connector mention domain irrelevant semantic contents and should not be taken into account.

7. REFERENCES

[1] Georgios Spithourakis, Isabelle Augenstein, and Sebastian Riedel, "Numerically grounded language models for semantic error correction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 987–992.

[2] Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao, "A nested attention neural hybrid model for grammatical error correction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 753–762.

[3] Allen Schmaltz, Yoon Kim, Alexander Rush, and Stuart Shieber, "Adapting sequence models for sentence correction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2807–2813.

[4] Su Zhu, Ouyu Lan, and Kai Yu, "Robust spoken language understanding with unsupervised asr-error adaptation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6179–6183.

[5] Askars Salimbajevs and Jevgenijs Strigins, "Error analysis and improving speech recognition for latvian language," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 2015, pp. 563–569.

[6] Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève, and Renato De Mori, "Asr error management for improving spoken language understanding," in *Interspeech 2017*, 2017.

[7] Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio, "Towards end-to-end spoken language understanding," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5754–5758.

[8] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin, "End-to-end named entity and semantic concept extraction from speech," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 692–699.

[9] Yuan-Ping Chen, Ryan Price, and Srinivas Bangalore, "Spoken language understanding without speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6189–6193.

[10] Antoine Caubrière, Natalia Tomashenko, Antoine Laurent, Emmanuel Morin, Nathalie Camelin, and Yannick Estève, "Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability," in *Interspeech 2019*, Graz, Austria, Sept. 2019.

[11] Natalia Tomashenko, Antoine Caubrière, and Yannick Estève, "Investigating adaptation and transfer learning for end-to-end spoken language understanding from speech," *Interspeech*, 2019.

[12] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[14] Natalia Tomashenko, Antoine Caubriere, Yannick Esteve, Antoine Laurent, and Emmanuel Morin, "Recent advances in end-to-end spoken language understanding," in *SLSP*, 2019.

[15] Vedran Vukotic, Christian Raymond, and Guillaume Gravier, "Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?," in *Interspeech*, 2015.

[16] Laurence Devillers, Hélène Maynard, Sophie Rosset, Patrick Paroubek, Kevin McTait, Djamel Mostefa, Khalid Choukri, Laurent Charnay, Caroline Bousquet, Nadine Vigouroux, et al., "The French MEDIA/EVALDA project: the evaluation of the understanding capability of spoken language dialogue systems.," in *LREC*, 2004.

[17] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.