Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding

Salima Mdhaffar, Valentin Pelloin, Antoine Caubrière, Gaëlle Laperrière, Sahar Ghannay, Bassam Jabaian, Nathalie Camelin, Yannick Estève

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2949–2956
Marseille, 20-25 June 2022

Impact Analysis of the Use of Speech and Language Models Pretrained by Self-Supersivion for Spoken Language Understanding

Salima Mdhaffar[1], Valentin Pelloin[2], Antoine Caubrière[1], Gaëlle Laperrière[1], Sahar Ghannay[3], Bassam Jabaian[1], Nathalie Camelin[2], Yannick Estève[1]
[1]LIA - Avignon Université, France
[2]LIUM - Le Mans Université, France
[3]LISN - CNRS/Université Paris-Saclay, France
[1]{firstname.lastname}@univ-avignon.fr, [2]{firstname.lastname}@univ-lemans.fr, [3]{firstname.lastname}@limsi.fr

Abstract

Pretrained models through self-supervised learning have been recently introduced for both acoustic and language modeling. Applied to spoken language understanding tasks, these models have shown their great potential by improving the state-of-the-art performances on challenging benchmark datasets. In this paper, we present an error analysis reached by the use of such models on the French MEDIA benchmark dataset, known as being one of the most challenging benchmarks for the slot filling task among all the benchmarks accessible to the entire research community. One year ago, the state-of-art system reached a Concept Error Rate (CER) of 13.6% through the use of an end-to-end neural architecture. Some months later, a cascade approach based on the sequential use of a fine-tuned wav2vec 2.0 model and a fine-tuned BERT model reaches a CER of 11.2%. This significant improvement raises questions about the type of errors that remain difficult to treat, but also about those that have been corrected using these models pre-trained through self-supervision learning on a large amount of data. This study brings some answers in order to better understand the limits of such models and open new perspectives to continue improving the performance.

Keywords: Spoken Language Understanding, Slot Filling, Error Analysis, Self-supervised models

1. Introduction

Spoken language understanding (SLU) aims at extracting a semantic representation from a speech signal in human-computer interaction applications (Tur and De Mori, 2011), like named entity recognition from speech, call routing, slot filling task in a context of human-machine dialogue,...

Last years, SLU task received a lot of attention by the research community and many approaches have been proposed. Traditional SLU approaches were processed through a cascade approach that is based on the use of an automatic speech recognition (ASR) system followed by a natural language understanding (NLU) module applied to the automatic transcription (De Mori, 2007) from the ASR sytem. For both ASR and NLU, deep neural networks have made great advances, leading to impressive improvements of qualitative performance for final SLU tasks (Amodei et al., 2016; Collobert et al., 2011; Vaswani et al., 2017).

Nowadays, state-of-the-art SLU systems are populated with end-to-end neural approaches, based on deep neural networks, that are proposed in order to directly extract semantic information from speech signal (Serdyuk et al., 2018; Ghannay et al., 2018). For cascade systems, the intermediate transcription may contain recognition errors, and the NLU module has to deal with these errors. The main advantage of end-to-end approaches is to skip the use of an intermediate speech transcription, and so to avoid ASR errors propagation. In addition, end-to-end approaches permit us to optimize the entire model to the final task, while cascade approaches need to optimize each module on a sub-task separately. Moreover, cascade models may have bottleneck issues because all information from the source features (speech) needs to be reduced into a single flat representation (words) before being transformed into the target representation (semantic annotation).

Very recently, works on self-supervised training with unlabelled data has opened new perspectives in terms of performance both for ASR and natural language processing (Baevski et al., 2020; Devlin et al., 2019). They can be applied to SLU tasks.

The recent state-of-the-art results on the French MEDIA benchmark were obtained by using a cascade approach that takes benefit from acoustic-based and linguistic-based models pre-trained on unlabelled data: wav2vec 2.0 (Baevski et al., 2020) model for the ASR module and BERT-based model (Devlin et al., 2019) for the NLU module.

This study presents the results of three state-of-art SLU systems on the French MEDIA benchmark corpus, that is one of the most challenging benchmarks for SLU task (Béchet and Raymond, 2019). We describe these systems and analyse the errors corrected by these approaches and the residual errors that remain hard to correct.

2. MEDIA Dataset: a Challenging French Benchmark Dedicated to the Slot Filling Task

The French MEDIA corpus (Bonneau-Maynard et al., 2005), is dedicated to semantic extraction from speech in a context of human-machine dialogues in a hotel room reservation task with touristic information.

(a) I would like to book one double room in Paris up to one hundred and thirty euros

(b) <booking I would like to book > <number-room one > <room-type double room > in <location Paris > <comparative-payment up to > <amount-payment one hundred and thirty > <currency-payment euros >

Figure 1: An example of a MEDIA dataset sample. (a) corresponds to the transcribed sentence. (b) the same sample with its additional semantic tags. Here, '<number-room ' is an opening tag starting the support word sequence 'one' and expressing that this word sequence is associated with the *number-room* semantic concept. The character '>' represents the closing tag and it is used to close all concept tags.

Data	Nb. words	Nb. utterances	Nb. concepts	Nb. hours
train	94.2k	13.7k	31.7k	10h 46m
dev	10.7k	1.3k	3.3k	01h 13m
test	26.6k	3.7k	8.8 k	02h 59m

Table 1: The official MEDIA dataset distribution

This dataset was created as a part of the Technolangue project of the French government in 2002. It aims, among others, to set up an infrastructure for the production and dissemination of language resources, and the evaluation of written and oral language technologies.

The MEDIA dataset is made of telephone dialogue recordings with their manual transcriptions and semantic annotations. It is composed of 1257 dialogues from 250 different speakers, collected with a Wizard-of-Oz setting between two humans: one plays a computer, the other plays the user. The dataset is split into three parts (train, dev and test) as described in Table 1. In this work, we used the user part of MEDIA, since it has both speech and semantic annotations.

The semantic domain of this corpus is represented by 76 semantic concept tags such as *room number, hotel name, location, etc.* Some more complex linguistic tags, like co-references, are also used in this corpus.

The sentence (translated from French) in Figure 1 is an example of a user utterance in MEDIA, with its corresponding semantic annotation.

Béchet and Raymond (2019) showed why the MEDIA task can be considered as the most challenging SLU benchmark available, in comparison to other well-known benchmarks such as ATIS (Dahl et al., 1994), SNIPS (Coucke et al., 2018), and M2M (Shah et al., 2018).

3. State-of-the-Art SLU Systems Description

We describe in this section the three systems used in our error analysis.

3.1. End-to-End Encoder-Decoder Approach with Attention Mechanism

The first system used in our analysis is an attention-based encoder-decoder neural network proposed by Pelloin et al. (2021).

This system achieved state-of-the-art performance on the MEDIA task in 2021. The encoder part is composed of four two-dimensional convolution layers (CNN) with batch normalisation. CNN layers are followed by four bidirectional Long Short-Term Memory (bLSTM) layers. The decoder part is a stack of four LSTM, two fully connected, and a softmax layer. The input features of the network are 40-dimensional MelF-Banks with a Hamming window of 25ms and 10ms strides. Figure 2 illustrates the architecture of the system.

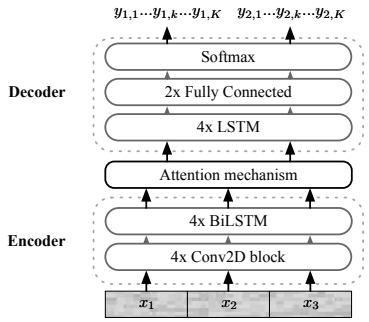

Figure 2: Architecture of Encoder-Decoder SLU (Pelloin et al., 2021)

This encoder-decoder system was trained following transfer learning approach, with external data for ASR supervised pretraining that is not easily accessible (French Broadcast news).

3.2. End-to-End Fine-Tuning of wav2vec 2.0 Models for SLU

The second system used in this paper is an end-to-end system fine-tuned for SLU. More precisely, a wav2vec 2.0 model (Evain et al., 2021) was first fine-tuned on the CommonVoice data for ASR, then on MEDIA for ASR, and last on MEDIA for the SLU task (to produce words and their associated semantic concepts). wav2vec 2.0 is introduced in (Baevski et al., 2020) and it is one of the current state-of-the-art Self-Supervision Language (SSL) model for ASR. It takes raw audio as input and computes contextual representations that can be used as input for speech recognition systems. We fine-tune it into an SLU system by adding concept boundaries inside the sequences to be produced as suggested in (Ghannay et al., 2018).

In Figure 3 an overview of the end-to-end system fine-tuned for SLU is presented.

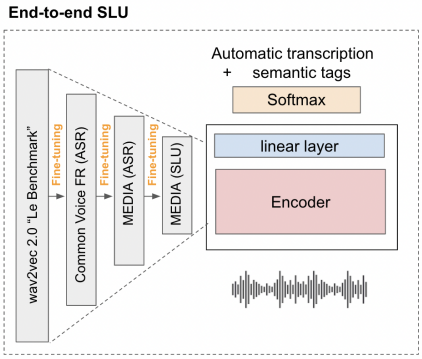

Figure 3: Overview of fine-tuning wav2vec 2.0 models for SLU: (1) Use of a French self-supervised pre-trained wav2vec 2.0 model (2) Finetune the model with the French CommonVoice (3) Finetune the model with MEDIA ASR (4) Finetune the model with MEDIA SLU.

3.3. Cascade Approach with Pre-Trained Models

The third system we propose to analyse is based on a cascade approach with pre-trained models for each component (Ghannay et al., 2021). The ASR part in this cascade approach is composed of the large pre-trained French wav2vec 2.0 model[1] (Evain et al., 2021), a linear layer of 1024 units, and the softmax output layer. This ASR system is trained by optimizing, firstly, the ASR system on the French CommonVoice dataset. Then, it is finetuned for speech recognition on the French MEDIA corpus, the wav2vec 2.0 weights being updated at each training stage. The loss function used at each fine-tuning step is the Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006).

The NLU system is applied to the automatic transcriptions provided from the ASR system, to obtain semantic annotations. This system is based on the fine-tuning of the French CamemBERT (Martin et al., 2020) model, on the manual transcriptions of MEDIA corpus. It achieved state-of-the-art results on manual transcriptions of MEDIA corpus (Ghannay et al., 2020), yielding to 7.56% of Concept Error Rate (CER).

In Figure 4, an overview of the cascade approach proposed in Ghannay et al. (2021) with pre-trained models is presented.

4. Systems Performance

SLU systems can be evaluated with different metrics. Conventional metrics jointly used on the MEDIA corpus are the Concept Error Rate (CER) and the Concept Value Error Rate (CVER). The CER is computed similarly to the Word Error Rate, by only taking into account the occurrences of concepts in both the reference and the hypothesis files. The CVER metric is an extension of the CER that considers the correctness of the complete concept/value pair.

Table 2 presents the results obtained on the official MEDIA benchmark for both dev and test dataset using the approaches presented in the previous sections. We provide confidence intervals with a confidence degree of 95%.

The best results, considering an automatic transcription, are obtained by the cascade model (Ghannay et al., 2021) composed of the wav2vec 2.0 model and the BERT-like model, reaching a CER of 11.2% and a CVER of 17.2% on the MEDIA test. We also reported in our analysis results of the CamemBERT approach applied on manual transcriptions (manual transcription + CamemBERT) to highlight the impact of speech recognition errors in the cascade approach. With manual transcriptions, the CamemBERT model reaches a CER of 7.5% and a CVER of 12.2% on the MEDIA test dataset.

5. Error Analysis

In this section, we analyse what type of errors the cascade model is able to correct by using CamemBERT and what type of errors are still present compared to the other systems: Enc-Dec/AM (Pelloin et al., 2021), wav2vec 2.0 fine-tuned for SLU and manual transcription + CamemBERT.

5.1. Error Distribution

First, Table 3 summarises the performance of the different systems on Dev and Test in terms of insertions, substitutions and deletions[2] for concept evaluation. We observe that the major error type is deletion for all systems with automatic transcription. This can be explained by transcription errors present in those systems preventing them from capturing any concept.

The comparison of the deletions between the wav2vec 2.0 + CamemBERT and the manual transcription + CamemBERT systems confirms that there is less concept deletions when the transcription is correct. As an example on the dev, we observe only 3.2% of deletions with manual transcriptions compared to 5.1% considering the automatic transcriptions. We can also notice that the wav2vec 2.0 + CamemBERT system has the lowest substitution rate in comparison to the two end-to-end systems.

Error distributions among the different semantic concepts of the four systems on the MEDIA Dev dataset are provided in Figure 5. For greater clarity, we only kept the 40 concepts with the highest number of errors,

[1] https://huggingface.co/LeBenchmark

[2] We use SCLITE from the SCTK toolkit to generate this error distribution (https://github.com/usnistgov/SCTK)

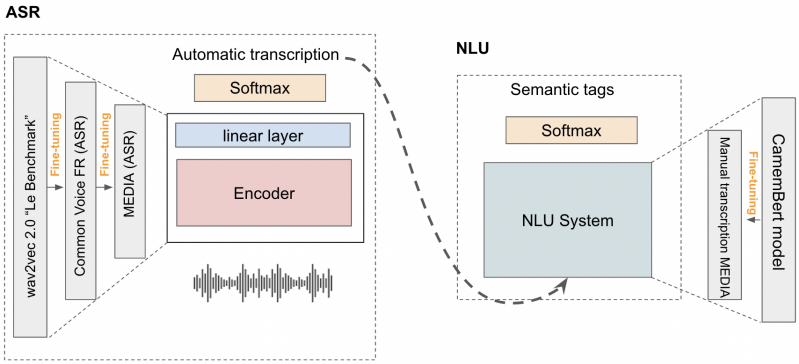

Figure 4: Overview of cascade approach with pre-trained models (1) ASR part (a) Use of a French self-supervised pre-trained wav2vec 2.0 model (b) Finetune the model with the French Common Voice (c) Finetune the model with MEDIA ASR (2) NLU part (a) Use of a French CamemBERT (b) Finetune the model with manual transcription of MEDIA (c) Extraction of semantic concepts for automatic transcription

Model	Dev		Test	
	CER	CVER	CER	CVER
Enc-Dec/AM (Pelloin et al., 2021)	16.1 (±1.2)	20.4 (±1.3)	13.6 (±0.7)	18.5 (±0.8)
wav2vec 2.0 fine-tuned for SLU	15.2 (±1.2)	19.6 (±1.3)	14.5 (±0.7)	18.8 (±0.8)
wav2vec 2.0 + CamemBERT	**12.2** (±1.1)	**16.7** (±1.2)	**11.2** (±0.7)	**17.2** (±0.8)
manual transcription + CamemBERT	9.2 (±1.0)	13.2 (±1.1)	7.5 (±0.6)	12.2 (±0.7)

Table 2: Performance on MEDIA dev and test in terms of CER and CVER scores (with a 95% confidence interval).

Model	Dev			Test		
	Ins	Sub	Del	Ins	Sub	Del
Enc-Dec/AM (Pelloin et al., 2021)	5.3	4.9	5.9	4.3	4.3	4.9
wav2vec 2.0 fine-tuned for SLU	4.1	4.1	7.1	3.8	3.8	6.9
wav2vec 2.0 + CamemBERT	4.1	2.9	5.1	3.4	2.8	4.9
manual transcription + CamemBERT	3.5	2.5	3.2	2.8	2.1	2.6

Table 3: Detailed performances on considering only concept for MEDIA dev and test in terms of insertions (Ins), substitution (Sub) and deletions (Del).

based on the Enc/Dec system, the previous end-to-end state-of-the-art system.

When the cascade system (wav2vec 2.0 + CamemBERT) is used, the five concepts that generate the most important amount of error are, in descending order, "*linkref-coref*" (that represents a coreference word that refers to a previous entity), "*proposition-connector*" (that is a word that connects two propositions), "*response*", "*object*" and "*location-city*". The first two are known as the most challenging to retrieve. We observe that the cascade system reduces the number of errors for a majority of concepts, mainly by reducing deletion and substitution errors. However, some concepts still remain hard to be recognised by this system, for instance, the *location-city* concept. Indeed, this concept is conveying by a lots of different values (*e.g.* all French city names, with some never seen in the train). Among the five top erroneous concepts of Enc-Dec/AM, we notice that four of them are also in the top five erroneous concepts of the systems which do not use manual transcriptions.

The cascade system (wav2vec 2.0 + CamemBERT) is particularly more effective to extract concepts related to date: "*time-day-month*", "*time-month*", "*time-date*", but this is not the case for other concepts like "*location-city*". Figure 5(d) shows that using manual transcriptions, errors related to *location-city* are mostly corrected. A first assumption could be that we have a higher WER for the words supporting the concept "*location-city*". In order to verify this assumption, we propose to analyse the number of errors concerning only the words contained in a concept support in the next subsection. A second assumption is that errors may come from prepositions in the neighbourhood of the concept, since the "*location-city*" concept is usually close to prepositions like '*à*' ("to" in English). Notice that prepositions are often part of a "*proposition-connector*" or "*linkref-coref*" concept. Last, it is very interesting to notice that the "*proposition-connector*" concept is not really better processed by the use of the CamemBERT model, while this is the most frequent error. For the "*linkref-coref*" concept, the trend seems

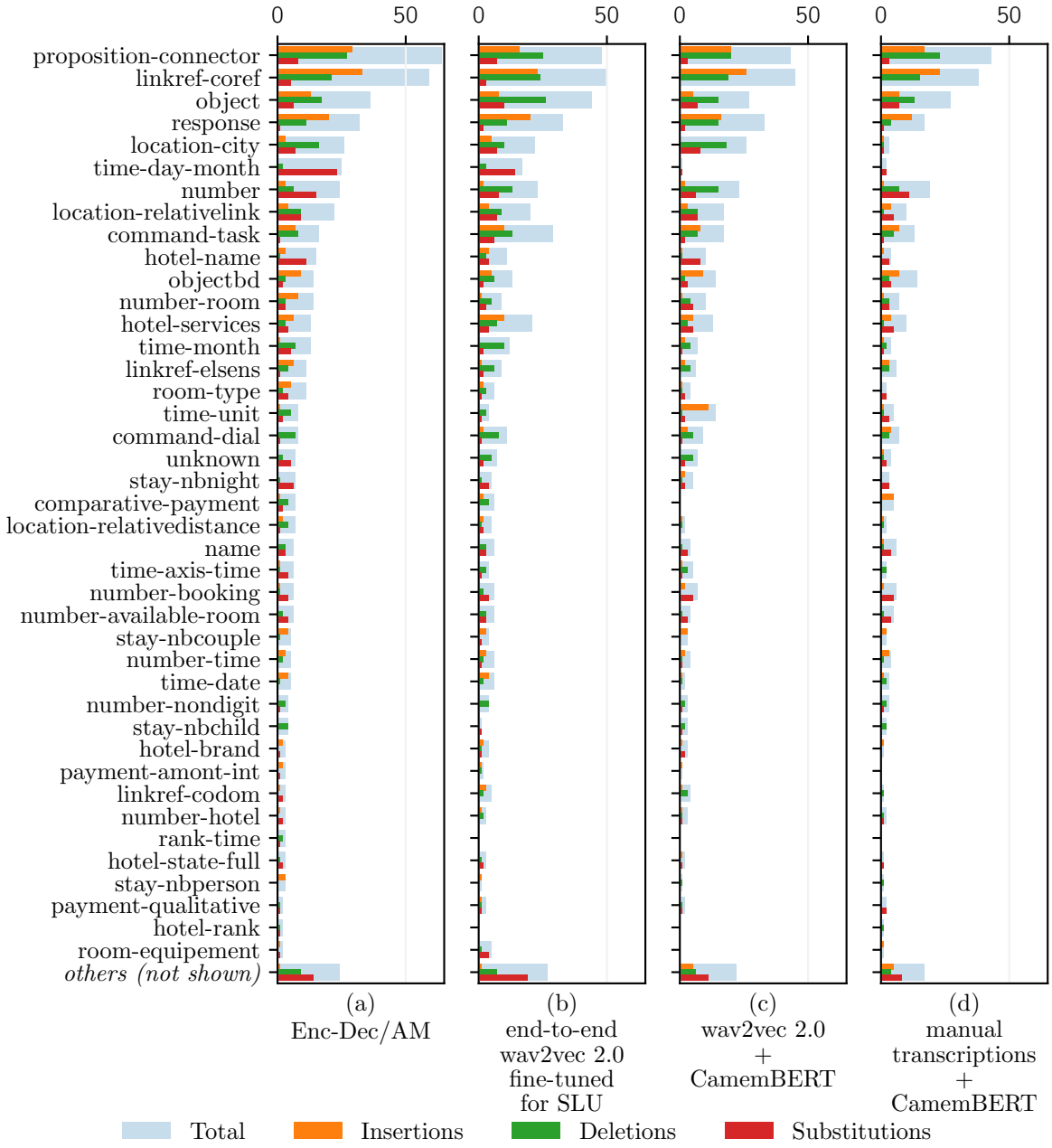

Figure 5: Error distribution for (a) *Enc-Dec/AM* (Pelloin et al., 2021) (b) end-to-end (wav2vec 2.0 fine-tuned for SLU) and (c) cascade (wav2vec 2.0 + CamemBERT) on Dev. dataset

the same.

5.2. Recognizing Unseen Concept/Value Pairs

It is interesting to measure the generalization capability of the different models. We define the Unseen Concept/Value (UCV) pairs as the concept/value pairs seen in the MEDIA development dataset which do not appear in the training dataset. There are a total of 543 UCV pairs on the MEDIA development dataset. The number of well-recognized UCV pairs (*C+V ok*) for the different approaches is reported in Table 4. We also report the number of UCV pairs for which the value has been correctly retrieved, while the concept has been misrecognized (*V ok only*).

As we can see, the wav2vec 2.0 + CamemBert system recognizes 44.5% (242 out of 543) of the concept/value pairs unseen in the MEDIA training corpus, while the end-to-end wav2vec 2.0 fine-tuned for SLU recognizes only 29.1% UCV pairs. This shows that the cascade (wav2vec 2.0 + CamemBert) system is really better for generalization. Besides, we can see that it recognizes correctly less values of UCV pairs (16) when the concept has been misrecognized, but this is probably due to the higher number of UCV pairs entirely well recognized for the cascade approach (the (V) OK only

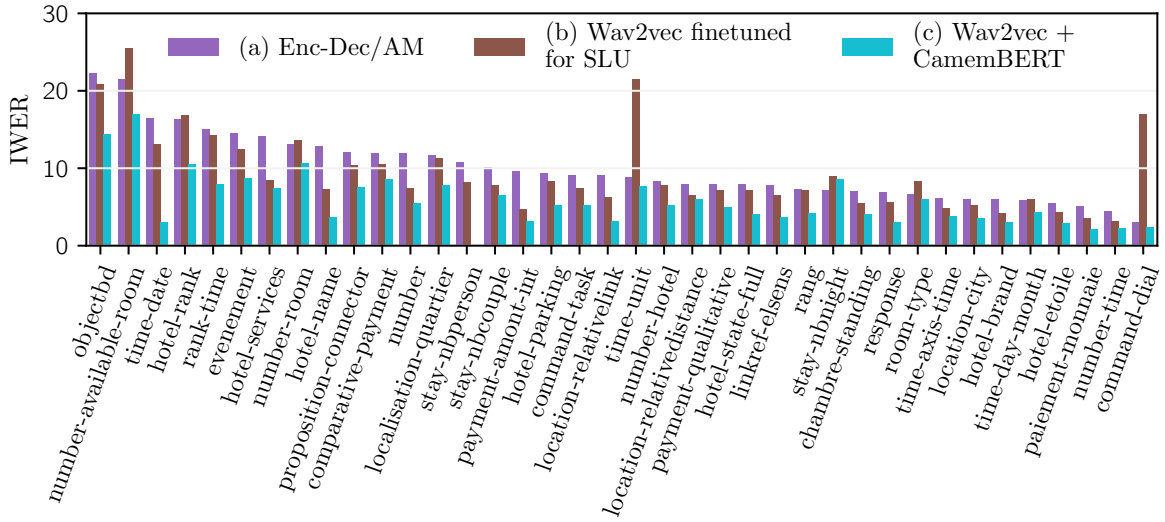

Figure 6: IWER for support words by concept (concepts are sorted by the IWER of the Enc-Dec/AM) on Dev dataset. We kept only concepts with support words occurring fewer than 20 times are not shown.

Model	**(C+V) OK**	**(V) OK only**
Enc-Dec/AM (Pelloin et al., 2021)	168	32
wav2vec 2.0 fine-tuned for SLU	158	47
wav2vec 2.0 + CamemBERT	242	16
manual transcription + CamemBERT	375	29

Table 4: UCV correctness on MEDIA dev dataset. (C+V) means Concept/Value pairs are correct and V means only the Value is correct while the concept is misrecognized

values missing may be already in the (C+V) OK). We can notice that the cascade system recognizes 47.5% of values (242+16 of 543) unseen in the training data, while the end-to-end wav2vec2.0 fine-tuned for SLU system is able to recognize only 37.8% of them. The last line in Table 4 shows that with manual transcription + CamemBert, the system is able to recognize 74.4% of values unseen in the training data.

5.3. IWER Analysis

The Word Error Rate (WER) metric assigns a global error score to transcriptions, which implies that each transcription error has the same impact on the reported ASR performance. We propose to exploit IWER (Individual Word Error Rate) metric (Goldwater et al., 2010) in order to evaluate the errors for only a specific set of words, composed of support words. Support words are words that are associated to a concept: support words are among the words involved in the values of a concept. For example, the words "*double*" and "*room*" are two support words for the concept "*room-type*" in Figure 1.

The IWER metric is computed as follows: for deletion and substitution errors, the principle is the same as for the WER. We attribute 1 or 0 by comparing the hypothesis to the reference. However, for insertion errors, there may be two adjacent words which cause the error. As we have no way to know which word is, we simply assign equal partial responsibility for any insertion errors to both of the adjacent words. So, for the i^{th} reference word w_i, the IWER is calculated as:

$$IWER(w_i) = del_i + sub_i + \alpha.ins_i \qquad (1)$$

where $del_i = 1$ if w_i is deleted, $sub_i = 1$ if w_i is substituted and ins_i = number of insertions adjacent to the word w_i. The parameter α is computed as follows: $\alpha = \frac{I}{\sum_{w_i} ins_i}$ where I is the number of insertions in all the corpus (the total penalty for insertion errors is the same as when computing WER). The IWER for a set of words is the average IWER for individual words:

$$IWER(w_1...w_n) = \frac{1}{n}\sum_{i=1}^{n} IWER(w_i) \qquad (2)$$

Table 5 shows that the global WER (IWER computed for all words is equal to WER) of the cascade approach (7.7%) is much lower than the WER of wav2vec 2.0 fine-tuned for SLU (12%). However, the wav2vec 2.0 models used on the two systems are very close: for the end-to-end version, the wav2vec 2.0 used for ASR in the cascade model has been fine-tuned on the SLU task to emit word and semantic concepts. Note that, to compute the WER of the end-to-end version, we removed the semantic concepts generated by the model.

Model	Global	Support words
Enc-Dec/AM (Pelloin et al., 2021)	12.37	13.66
wav2vec 2.0 fine-tuned for SLU	12	13.5
wav2vec 2.0 + CamemBERT	7.7	9.27

Table 5: IWER (%) results for all words (=global WER) and for support words that are involved in values associated to concepts on dev dataset MEDIA.

It seems that during its fine-tuning on the SLU task, the wav2vec 2.0 model forgot some knowledge about automatic speech recognition. This is maybe due to the increase of the number of token output (the same number of characters + 76 symbols to handle the semantic concepts).

The IWER of support words for each concept is illustrated in Figure 6. Stop-words are removed from all support words except for some concepts like *proposition-connector* that is mainly supported by coordinating conjunctions and prepositions usually included in stop-word lists. We observe that end-to-end systems (Enc/dec or wav2vec 2.0 fine-tuned for SLU) tend to degrade ASR results compared to the wav2vec 2.0 used for ASR used in the cascade model. This figure confirms the significant better performance of the wav2vec 2.0 used for ASR in the cascade model, for all the lists of support words related to all the concepts (*i.e.* blue bars are always under purple and brown ones).

We also can observe that wav2vec 2.0 fine-tuned for SLU makes generally fewer errors than Enc/Dec (Pelloin et al., 2021) system (*i.e.* brown bars often under purple bars), except for some concepts like *time-unit*, *command-dial* and *number-available-room*, on which this model produces some peaks of errors.

Finally, we can see that our first assumption related to the concept *location-city* in section 5.1 is not validated. Figure 6 shows that support words for this concept does not observe a particularly great IWER. Moreover, support words seem to be even well recognized by the wav2vec 2.0 used for ASR in the cascade model.

6. Conclusion

In this paper, we present an error analysis reached by the use of self-supervised models on the French MEDIA dataset. Three systems outputs have been compared in this study: (1) an encoder-decoder approach that reached the state-of-the-art on MEDIA dataset in the beginning of 2021; (2) an end-to-end system fine-tuned for SLU that takes benefit from the use of a pre-trained wav2vec 2.0 model; (3) a cascade approach that also takes benefit from the use of a pre-trained wav2vec 2.0 model dedicated to French language and of a pre-trained BERT-like model in French, CamemBERT. We also compare results of CamemBERT applied to a manual transcription. The error analysis shows different points. First, the use of a BERT-like model has a great impact: this makes possible a better generalization that allows a better detection of words and concepts unseen in the training data. The CamemBERT model is relevant in order to improve the recognition of several concepts, especially the semantic tags related to the expression of date. But for some concepts, like *location-city* for instance, CamemBERT with automatic transcription seems in difficulty. One of our future work will tend to reinforce such weaknesses during the CamemBERT fine-tuning.

Surprisingly, while the wav2vec 2.0 models used in the cascade and end-to-end approaches in our experiments are very close, the WER got by the cascade model is much lower than the one got in the end-to-end approach. It seems that during the fine-tuning of the wav2vec 2.0 model on the SLU data, wav2vec 2.0 forgot some of its automatic speech recognition abilities. This is maybe due to the increase of the number of token output (same number of characters + 76 symbols to handle the semantics concepts), which increases the difficulty for the model. As a result, the end-to-end model reaches very bad results in terms of WER in comparison to the wav2vec 2.0 fine-tuned for the ASR task only.

Combining wav2vec 2.0 and BERT model in an end-to-end approach is also a promising perspective, left for future work. The knowledge extracted from the error analysis presented in this paper could be useful in order to optimize this combination.

7. Acknowledgements

This paper was partially funded by the European Commission through the SELMA project under grant number 957017 and by the AISSPER project supported by the French National Research Agency (ANR) under contract ANR-19-CE23-0004-01. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012551 made by GENCI. It has been also made possible thanks to the Saclay-IA computing platform.

8. Bibliographical References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Béchet, F. and Raymond, C. (2019). Benchmarking benchmarks: introducing new automatic indicators for benchmarking spoken language understanding corpora. In *Interspeech*, Graz, Austria.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005). Semantic annotation of the french media dialog corpus. In *INTERSPEECH*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al. (2018). Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Dahl, D. A., Bates, M., Brown, M. K., Fisher, W. M., Hunicke-Smith, K., Pallett, D. S., Pao, C., Rudnicky, A., and Shriberg, E. (1994). Expanding the scope of the atis task: The atis-3 corpus. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

De Mori, R. (2007). Spoken language understanding: a survey. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 365–376. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021). LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Proc. Interspeech 2021*, pages 1439–1443.

Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., and Morin, E. (2018). End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 692–699. IEEE.

Ghannay, S., Servan, C., and Rosset, S. (2020). Neural networks approaches focused on French spoken language understanding: application to the MEDIA evaluation task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2722–2727, Barcelona, Spain (Online), december. International Committee on Computational Linguistics.

Ghannay, S., Caubrière, A., Mdhaffar, S., Laperrière, G., Jabaian, B., and Estève, Y. (2021). Where are we in semantic concept extraction for spoken language understanding? In *International Conference on Speech and Computer*, pages 202–213. Springer.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Pelloin, V., Camelin, N., Laurent, A., De Mori, R., Caubrière, A., Estève, Y., and Meignier, S. (2021). End2end acoustic to semantic transduction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7448–7452.

Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., and Bengio, Y. (2018). Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE.

Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., and Heck, L. (2018). Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.