HAL
open science

# Where are we in semantic concept extraction for Spoken Language Understanding? ⋆

Sahar Ghannay, Antoine Caubrière, Salima Mdhaffar, Gaëlle Laperrière, Bassam Jabaian, Yannick Estève

## To cite this version:

Sahar Ghannay, Antoine Caubrière, Salima Mdhaffar, Gaëlle Laperrière, Bassam Jabaian, et al.. Where are we in semantic concept extraction for Spoken Language Understanding? ⋆. SPECOM 2021 23rd International Conference on Speech and Computer, Sep 2021, Saint Petersburg, Russia. hal-03372494

**HAL Id: hal-03372494**

**https://hal.science/hal-03372494v1**

Submitted on 10 Oct 2021

**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:2106.13045v1 [cs.CL] 24 Jun 2021

# Where are we in semantic concept extraction for Spoken Language Understanding? *

Sahar Ghannay[1], Antoine Caubrière[2], Salima Mdhaffar[2], Gaëlle Laperrière[2], Bassam Jabaian[2], Yannick Estève[2]

[1] Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
firstname.lastname@limsi.fr
[2] LIA - Avignon Université, France
firstname.lastname@univ-avignon.fr

**Abstract.** Spoken language understanding (SLU) topic has seen a lot of progress these last three years, with the emergence of end-to-end neural approaches. Spoken language understanding refers to natural language processing tasks related to semantic extraction from speech signal, like named entity recognition from speech or slot filling task in a context of human-machine dialogue. Classically, SLU tasks were processed through a cascade approach that consists in applying, firstly, an automatic speech recognition process, followed by a natural language processing module applied to the automatic transcriptions. These three last years, end-to-end neural approaches, based on deep neural networks, have been proposed in order to directly extract the semantics from speech signal, by using a single neural model. More recent works on self-supervised training with unlabeled data open new perspectives in term of performance for automatic speech recognition and natural language processing. In this paper, we present a brief overview of the recent advances on the French MEDIA benchmark dataset for SLU, with or without the use of additional data. We also present our last results that significantly outperform the current state-of-the-art with a Concept Error Rate (CER) of 11.2%, instead of 13.6% for the last state-of-the-art system presented this year.

**Keywords:** Spoken language understanding · End-to-end approach · Cascade approach · Self supervised training

## 1 Introduction

Spoken language understanding (SLU) refers to natural language processing tasks related to semantic extraction from the speech signal [34], like named entity recognition from speech, call routing, slot filling task in a context of human-machine dialogue...

Usually, SLU tasks were processed through a cascade approach that consists in applying first an automatic speech recognition (ASR) process, followed by a

* This work was granted access to the HPC resources of IDRIS under the allocation 2020-AD011011838 made by GENCI.

natural language processing module applied to the automatic transcription [14]. For both automatic speech recognition and natural language processing, deep neural networks (DNN) have made great advances possible, leading to impressive improvements of qualitative performance for final SLU tasks [1,11,35].

These three last years, end-to-end neural approaches, based on deep neural networks, have been proposed in order to directly extract the semantics from speech signal, by using a single neural model [29,18]. A first advantage of such approaches consists on a joint optimization of the ASR and NLP part, since the unique neural model is optimized only for the final SLU task. Another advantage is the limitation of the error propagation: when using a cascade approach, an error in the first treatment implies errors in the following ones. In a neural end-to-end approach, the model decision is delayed to the output layer: all the information uncertainty is handled until the final decision.

Very recently, works on self-supervised training with unlabeled data open new perspectives in term of performance for automatic speech recognition and natural language processing [2,16]. They can be applied to SLU task.

This study presents experimental results on the French MEDIA benchmark dataset. This benchmark dataset is one of the most challenging benchmarks for SLU task. In this paper, we present a brief overview of the performance evolution of state-of-the-art systems on this benchmark dataset. We also present an approach that takes benefit from acoustic-based and linguistic-based models pre-trained on unlabelled data: this approach represents the next milestone to be surpassed.

## 2 MEDIA dataset

The French MEDIA corpus [5], is dedicated to semantic extraction from speech in a context of human-machine dialogues for a hotel booking task. This dataset was created as a part of the Technolangue project of the French government in 2002. Its main objective is to set up an infrastructure for the production and dissemination of language resources, the evaluation of written and oral language technologies, the participation in national and international standardisation bodies and an information monitoring in the field.

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, 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 following sentence (translated from French) is an example of the MEDIA content: "I would like to book one double room in Paris up to one hundred and thirty euros". It will be annotated as (I would like to book, *reservation*), (one,

**Table 1.** The official MEDIA dataset distribution

| Data | Nb Words | Nb Utterances | Nb Concepts | Nb Hours |
|---|---|---|---|---|
| train | 94.2k | 13.7k | 31.7k | 10h46m |
| dev | 10.7k | 1.3k | 3.3k | 01h13m |
| test | 26.6k | 3.7k | 8.8 k | 02h59m |

*number-room*), (double room, *room-type*), (up to, *comparative-payment*), (one hundred and thirty, *amount-payment*), (euros, *currency-payment*).

In [4], Béchet and Raymond 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 [13], SNIPS [12], and M2M [31].

## 3 Overview of approaches proposed for the MEDIA benchmark

### 3.1 Cascade approach

Conventional SLU systems are designed as a cascade of components. Each of them solves separately a specific problem. First, an ASR module, trained on a large amount of data, maps speech signals to automatic transcriptions. This is then passed on to a natural language understanding (NLU) module that predicts semantic information from the automatic transcriptions. In this approach, error propagation is unavoidable, despite the performance of current ASR and NLU systems. In addition, those modules are optimized separately under different criteria. The ASR system is trained to minimize the word error rate (WER), while the NLU module is trained to minimize the concept error rate (CER) in case of slot filling task. This separate optimization suggests that a cascade SLU system is suboptimal.

Working on automatic transcriptions, for an SLU task on MEDIA corpus, is highly challenging. Many approaches have been proposed. Early NLU approaches were based on generative models such as Stochastic finite state transducers (FST), on discriminative or conditional models such as conditional random fields (CRFs) and support vector machines(SVMs)[21]. In the light of the success of neural approaches in different fields, some studies developed neural architectures for SLU task. In [27], the author presents the first recurrent neural architecture dedicated to SLU for the ATIS benchmark corpus. This neural model was applied to transcriptions despite speech signals directly. In [33,32], for the first time, an encoder-decoder neural network structure with attention mechanism [3] was proposed for this task. This time, it was on manual and automatic transcriptions from the MEDIA corpus. In order to reduce the unavoidable SLU performance decline due to ASR errors, the authors in [33] have proposed ASR confidence measures to localize ASR errors. These confidence measures have been used as additional SLU features to be combined with lexical and syntactic features, useful for characterizing concept mentions. In [32], the authors proposed

an approach to simulate ASR errors from manual transcriptions, to improve the performance of SLU systems. The use of the resulting corpus prepares the SLU system to ASR errors during their training and makes it more robust to ASR errors.

### 3.2 End-to-end approach

As seen in the previous section, one problem with cascaded approaches is the propagation of errors through the components. The intermediate transcription is noisy due to speech recognition errors, and the NLU component has to deal with these errors. The other problem comes from the separate optimzations of the different modules.

To tackle these issues, end-to-end approaches were proposed in order not to use an intermediate speech transcriptions. This kind of approach aims to develop a single system directly optimized to extract semantic concepts from speech.

SLU end-to-end systems are usually trained to generate both recognized words and semantic tags [18,15].

Until now, mainly two kinds of neural architectures have been proposed on the MEDIA benchmark. The first one is based on the use of the Connectionist Temporal Classification (CTC) loss function [20], while the other one is based on the use of an encoder-decoder architecture with attention mechanism [3].

#### 3.2.1 CTC approach

In this work, we call CTC approach the neural architecture trained by using the CTC loss function. This loss function allows the system to learn an alignment between the input speech and the word and concept sequences to produce.

To our knowledge, the best-published results with a CTC approach on MEDIA were obtained by [6]. In this study, the authors proposed a neural architecture largely inspired by the DeepSpeech 2 speech recognition system. The neural architecture is a stack of two 2D-invariant convolutional layers (CNN), followed by five bidirectional long short term memory (bLSTM) layers with sequence-wise batch normalization, a classical fully connected layer, and the softmax output layer. As input features, we used spectrograms of power to normalize audio clips, calculated on 20ms windows. This system was trained following the curriculum-based transfer learning approach, which consists in training the same model through successive stages, with different tasks ranked from the most generic one to the most specific one. The authors used speech recognition tasks, then named entity extraction and finally semantic concept extraction tasks.

#### 3.2.2 Encoder-decoder approach with attention mechanism

The encoder-decoder architecture was initially implemented in the machine translation context. This approach quickly showed its benefits for the speech recognition task [9,7,8], and more recently for SLU tasks [29,28].

The encoder-decoder architecture is divided into two main parts. First, an encoder receives the speech features as input, and provides its hidden states to build a high-level representation of the features. This high-level representation is then passed on to an attention module. It identifies the parts of these representations that are relevant for each step of the decoding process. Next, the attention module computes a context vector from these representations to feed the decoder. Finally, the decoder processes the input context vectors to predict the transcription of speech, enriched with semantic concepts. At each decoding time step, a new context vector is computed from the encoded speech representations. Unlike CTC approaches, the output sequence size of an encoder-decoder approach does not depend on the input sequence size.

A recent study [28] used a similar architecture and obtained the state-of-the-art performance for the MEDIA task. The encoder part is composed of four 2-dimensional convolution layers followed by four bLSTM layers. Each convolution layer is followed by a batch normalization. The decoder part is a stack of four bLSTM layers, two fully connected layers, and a softmax layer. The input features of the network are 40-dimensional MelFBanks with a Hamming window of 25ms and 10ms strides.

This encoder-decoder system is trained following the curriculum-based transfer learning, with the same data used for the CTC approach presented in section 3.2.1, except for the named entity extraction task which was not used.

### 3.3 System performance

SLU systems can be evaluated with different metrics. Historically, on the MEDIA corpus, two metrics are jointly used: 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 concepts occurrences in both the reference and the hypothesis files. The CVER metrics is an extension of the CER. It considers the correctness of the complete concept/value pair. In the example in section 2, both "one hundred and thirty" and *amount-payment* have to be correct to consider the concept/value pair (one hundred and thirty,*amount-payment*) as correct. Errors on the value component can come from a bad segmentation (missing or additional words in the value) or from ASR errors.

Table 2 presents the best results obtained on the official MEDIA benchmark dataset, by the main families of approaches presented in the two previous sections. By computing the 95% confidence interval, we observe a 0.7 confidence margin for CER and 0.8 for CVER, when the CER is 13.6% and the CVER 18.5%. Until now, the best result was reached by an end-to-end encoder-decoder architecture with attention mechanism, trained by following a curriculum transfer-learning approach [28].

## 4 Improving the state of the art

In the previous section we present state-of-the-art performances. Recently, unsupervised learning on huge amount of data have been successfully proposed

| Architecture | Model | CER | CVER |
|---|---|---|---|
| Cascade (2018) | HMM/DNN ASR + neural NLU [32] | 20.2 | 26.0 |
| Cascade (2018) | HMM/DNN ASR + CRF [32] | 20.2 | 25.3 |
| Cascade (2019) | HMM/TDNN ASR + CRF [6] | 16.1 | 20.4 |
| End-to-end (2019) | E2E CTC [6] | 16.4 | 20.9 |
| End-to-end (2021) | E2E encoder-decoder with attention [28] | **13.6** | **18.5** |

**Table 2.** Best results obtained on the official MEDIA benchmark dataset, by the main families of approaches presented in this paper. Results are given in both Concept Error Rate and Concept/Value Error Rate

to pre-train Transformers-based models [16,2]. Thanks to these models, ASR state-of-the-art performance [2] and NLP state-of-the-art performance [16] have been outperformed, with respectively wav2vec and BERT models. In this section, we present a cascade system using both BERT and wav2vec optimized on the MEDIA task.

### 4.1 BERT and CamemBERT models

For the NLU module, we propose to use the one that achieved the state-of-the-art result on manual transcriptions of MEDIA corpus [19]. This system is based on a fine-tuning of BERT [16] on MEDIA SLU task using the French CamemBERT [26] model.

BERT [16] is a deeply bidirectional, unsupervised language representation model, which stands for Bidirectional Encoder Representations from Transformers. It is designed to pre-train deep bidirectional representations from unlabeled text, taking into account both left and right context in all layers. The resulting pre-trained BERT model can be fine-tuned with just one additional output layer, to create state-of-the-art models for a wide range of NLP tasks. BERT is pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus which include the Toronto Book Corpus and Wikipedia.

The French CamemBERT model is based on RoBERTa (Robustly Optimized BERT Pre-training Approach) [25] which is based on BERT. CamemBERT is similar to RoBERTa, which dynamically change the masking pattern applied to the training data, and remove the next sentence prediction task. In addition, it uses the whole word masking and the SentencePiece tokenization [24]. The CamemBERT model is trained on the French CCNet corpus composed of 135GB of raw text.

### 4.2 Wav2vec models

Wav2vec 2.0 [2] is a model pre-trained through self-supervision. It takes raw audio as input and computes contextual representations that can be used as input for speech recognition systems. It contains three main components: a convolutional feature encoder, a context network and a quantization block. The

convolutional feature encoder converts the audio signal into a latent representation. This representation is given to the context network which takes care of the context. The context network architecture consists of a succession of several transformer encoder blocks. The quantization network is used to map the latent representation to quantized representation.

In [17], the authors released French pre-trained wav2vec 2.0 models. Two models have been released for public use[3], a large one and a base one. In this study, we use the large configuration which encodes raw audio into frames of 1024-dimensional vectors. The models are pre-trained in a unsupervised way with 3K hours of unlabeled speech. Details about data used to train the wav2vec models can be found in [17]. The trained model is composed of about 300M parameters.

To get better ASR results than the ones we could reach by fine-tuning the French wav2vec 2.0 model, on the MEDIA training data only, we suggest to ,first, fine-tune on external audio data, as proposed in [6] or [28]. To make the experiments reproducible, instead of using the Broadcast News data used in these works, we used the CommonVoice French dataset[4] (version 6.1), collected by the Mozilla Foundation, and much easily accessible. The train set consists of 425.5 hours of speech, while the validation and test sets contain around 24 hours of speech.

### 4.3 Cascade approach with pre-trained models

As written before, we propose in this work to use a cascade approach, with pre-trained models for each component. The ASR system is composed of the large pre-trained French wav2vec model, a linear layer of 1024 units, and the softmax output layer. First, we optimize the ASR system on the French CommonVoice dataset. Then, we fine-tune it for speech recognition on the French MEDIA corpus, the wav2vec weights being updated at each training stage. The loss function used at each fine-tuning step is the CTC loss function. We call the final ASR model W2V • Common Voice • $M_{ASR}$.

The NLU system is applied on the automatic transcriptions provided from the ASR system, to obtain semantic annotations. This system is based on the fine-tuning of the French CamemBERT [26] model, on the manual transcriptions of MEDIA corpus. It achieved state-of-the-art result on manual transcriptions of MEDIA corpus [19], yielding to 7.56 of CER when there is no error in the transcription.

### 4.4 Results and discussion

The experimental result obtained with the proposed cascade approach is presented in table 3. We compare the performance of this cascade system, named *W2V • Common Voice •* $M_{ASR}$ *+ CamemBERT*, to the E2E encoder-decoder

[3] https://huggingface.co/LeBenchmark
[4] https://commonvoice.mozilla.org/fr/datasets

model proposed in [28], that reached the best result on this task until now, and other wav2vec-based models. All the wav2vec-based models presented in table 3 were implemented thanks to the SpeechBrain toolkit[5], including the fine-tuning of the wav2vec models.

Like in section 3.3, the results are evaluated in terms of CER and CVER. Our new system yields to 17.64% of relative CER improvement and 7.02% of relative CVER improvement, by reaching respectively 11.2% of CER and 17.2% of WER. The result shows the effectiveness of unsupervised pre-trained models like wav2vec and BERT in such a scenario. Notice that the *W2V • Common Voice* • $M_{ASR}$ model allows us to have an effective ASR system that achieved 8.5% of WER.

In system (1), the wav2vec model is fine-tuned directly on MEDIA SLU ($M_{SLU}$) task. In system (2), the wav2vec model is first fine-tuned on the Common Voice data then on $M_{SLU}$ task, and a beam search decoding is applied. In systems (3) and (4) the wav2vec model is first fine-tuned on the Common Voice data, then on MEDIA ASR ($M_{ASR}$), and last on $M_{SLU}$ task, using the greedy or the beam search decoding using a 5-gram language model to rescore. This language model is trained on the manual transcriptions of $M_{SLU}$ training data only.

It is worth to mention that even before the generalisation of the use of neural networks for sequential tagging tasks, such as the slot filling task investigated in this paper, several efforts have been made to better take into account the the ASR system errors during the semantic labeling. Many approaches have been proposed for a joint decoding between speech recognition and understanding, considering the n-best recognition hypotheses during the semantic annotation [22,30,23]. When neural networks have become state-of-the-art systems for SLU, end-to-end approaches have gradually replaced cascade approaches and have shown very good performance, allowing the semantic labeling of a speech signal and minimising the impact of transcription errors on the SLU performance. However, these architectures need a large amount of data and often use pre-trained external module that have been trained separately in out-of-context data. The results presented in table 3 show that if such pre-trained models are used in a cascade architecture, the resulting system reaches or even exceeds the performance of the end-to-end based one. In addition, the result of the cascade system reinforces the idea of the use of pre-trained models at the encoder (wav2vec) and decoder (BERT) levels within end-to-end architecture, as proposed in [10]. This leads us to conclude that the two architectures remain valid and competitive and that the choice should be made according to the availability of additional data and the pre-training models.

## 5 Conclusions

In this paper, we present a brief overview of the recent advances on the French MEDIA benchmark dataset for SLU. We propose a system based on a cascade

[5] https://speechbrain.github.io

| Architecture | Model | CER | CVER |
|---|---|---|---|
| End-to-end | encoder-decoder [28] | 13.6 | 18.5 |
| | (1) W2V • $M_{SLU}$ (Beam 5g) | 18.8 | 23.6 |
| | (2) W2V • Common Voice • $M_{SLU}$ (Beam 5g) | 15.8 | 20.4 |
| | (3) W2V • Common Voice • $M_{ASR}$ • $M_{SLU}$ (greedy) | 15.4 | 20.5 |
| | (4) W2V • Common Voice • $M_{ASR}$ • $M_{SLU}$ (Beam 5g) | 14.5 | 18.8 |
| Cascade | W2V • Common Voice • $M_{ASR}$ + CamemBERT | **11.2** | **17.2** |

**Table 3.** Performance on Test MEDIA in terms of CER and CVER scores of the proposed cascade and end-to-end systems using pre-trained models. "•" formalizes a transfer learning step during the training of the E2E system.

approach, that takes benefit from acoustic-based and linguistic-based models pre-trained on unlabelled data : wav2vec models for the ASR system, and BERT-like model for the NLU system. Experimental results show that our system outperforms significantly the current state of the art with a Concept Error Rate (CER) of 11.2% instead of 13.6% for the last state-of-the-art system presented this year.

This new advance reinforces the idea of the use of pre-trained models at the encoder (wav2vec) and decoder (BERT) levels within an end-to-end architecture. This will explored in our future work.

This study leads us to conclude that the two architectures (cascade *vs.* end-to-end) remain valid and competitive and that the choice should be made according to the availability of additional data and relevant pre-trained models.

## References


1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182. PMLR (2016)
2. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477 (2020)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Béchet, F., Raymond, C.: Benchmarking benchmarks: introducing new automatic indicators for benchmarking spoken language understanding corpora. In: Interspeech. Graz, Austria (2019)
5. Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., Mostefa, D.: Semantic annotation of the french media dialog corpus. In: INTERSPEECH (2005)
6. Caubrière, A., Tomashenko, N., Laurent, A., Morin, E., Camelin, N., Estève, Y.: Curriculum-Based Transfer Learning for an Effective End-to-End Spoken Language Understanding and Domain Portability. In: Proc. Interspeech 2019. pp. 1198–1202 (2019). https://doi.org/10.21437/Interspeech.2019-1832, http://dx.doi.org/10.21437/Interspeech.2019-1832

7. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4960–4964. IEEE (2016)
8. Chiu, C.C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., et al.: State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4774–4778. IEEE (2018)
9. Chorowski, J., Bahdanau, D., Cho, K., Bengio, Y.: End-to-end continuous speech recognition using attention-based recurrent nn: First results. arXiv preprint arXiv:1412.1602 (2014)
10. Chung, Y.A., Zhu, C., Zeng, M.: Splat: Speech-language joint pre-training for spoken language understanding. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1897–1907 (2021)
11. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)
12. Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al.: Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. arXiv preprint arXiv:1805.10190 (2018)
13. Dahl, D.A., Bates, M., Brown, M.K., Fisher, W.M., Hunicke-Smith, K., Pallett, D.S., Pao, C., Rudnicky, A., Shriberg, E.: 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 (1994)
14. De Mori, R.: Spoken language understanding: a survey. In: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). pp. 365–376. IEEE (2007)
15. Desot, T., Portet, F., Vacher, M.: Towards end-to-end spoken intent recognition in smart home. In: 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). pp. 1–8. IEEE (2019)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, `https://www.aclweb.org/anthology/N19-1423`
17. Evain, S., Nguyen, H., Le, H., Zanon Boito, M., 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., Besacier, L.: Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In: Interspeech. Brno, Czechia (2021)
18. Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., Morin, E.: End-to-end named entity and semantic concept extraction from speech. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 692–699. IEEE (2018)
19. Ghannay, S., Servan, C., Rosset, S.: 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. pp. 2722–2727. International Committee on Computational Linguistics, Barcelona, Spain (Online) (december 2020). https://doi.org/10.18653/v1/2020.coling-main.245, https://www.aclweb.org/anthology/2020.coling-main.245

20. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006)
21. Hahn, S., Dinarelli, M., Raymond, C., Lefevre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., Riccardi, G.: Comparing stochastic approaches to spoken language understanding in multiple languages. IEEE Transactions on Audio, Speech, and Language Processing **19**(6), 1569–1583 (2010)
22. Hakkani-Tü, D., Béchet, F., Riccardi, G., Tur, G.: Beyond ASR 1-best: Using word confusion networks in spoken language understanding. Computer Speech and Language (Oct 2005). https://doi.org/10.1016/j.csl.2005.07.005, https://hal.archives-ouvertes.fr/hal-01314993
23. Jabaian, B., Lefèvre, F.: Error-corrective discriminative joint decoding of automatic spoken language transcription and understanding. In: Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P. (eds.) INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013. pp. 2718–2722. ISCA (2013), http://www.isca-speech.org/archive/interspeech_2013/i13_2718.html
24. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://www.aclweb.org/anthology/D18-2012
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
26. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
27. Mesnil, G., He, X., Deng, L., Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Interspeech. pp. 3771–3775 (2013)
28. Pelloin, V., Camelin, N., Laurent, A., De Mori, R., Caubrière, A., Estève, Y., Meignier, S.: End2end acoustic to semantic transduction. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7448–7452 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413581
29. Serdyuk, D., Wang, Y., Fuegen, C., Kumar, A., Liu, B., Bengio, Y.: Towards end-to-end spoken language understanding. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5754–5758. IEEE (2018)
30. Servan, C., Raymond, C., Béchet, F., Nocera, P.: Conceptual decoding from word lattices: application to the spoken dialogue corpus MEDIA. In: The Ninth International Conference on Spoken Language Processing (Interspeech 2006 - ICSLP). Pittsburgh, United States (Sep 2006), https://hal.archives-ouvertes.fr/hal-01160181

31. Shah, P., Hakkani-Tür, D., Tür, G., Rastogi, A., Bapna, A., Nayak, N., Heck, L.: Building a conversational agent overnight with dialogue self-play. arXiv preprint arXiv:1801.04871 (2018)
32. Simonnet, E., Ghannay, S., Camelin, N., Estève, Y.: Simulating ASR errors for training SLU systems. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://www.aclweb.org/anthology/L18-1499
33. Simonnet, E., Ghannay, S., Camelin, N., Estève, Y., De Mori, R.: ASR error management for improving spoken language understanding. In: Interspeech 2017. Stockholm, Sweden (Aug 2017)
34. Tur, G., De Mori, R.: Spoken language understanding: Systems for extracting semantic information from speech. John Wiley & Sons (2011)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)