The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools

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

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1595–1602
Marseille, 20-25 June 2022

The Spoken Language Understanding MEDIA Benchmark Dataset in the Era of Deep Learning: data updates, training and evaluation tools

Gaëlle Laperrière[1], Valentin Pelloin[2], Antoine Caubrière[1], Salima Mdhaffar[1], Nathalie Camelin[2], Sahar Ghannay[3], Bassam Jabaian[1], Yannick Estève[1]
[1]LIA - Avignon Université, France
[2]LIUM - Le Mans Université, France
[3]Université Paris-Saclay, CNRS, LISN, 91405 Orsay, France
[1]{firstname.lastname}@univ-avignon.fr, [2]{firstname.lastname}@univ-lemans.fr, [3]{firstname.lastname}@limsi.fr

Abstract

With the emergence of neural end-to-end approaches for spoken language understanding (SLU), a growing number of studies have been presented during these last three years on this topic. The major part of these works addresses the spoken language understanding domain through a simple task like speech intent detection. In this context, new benchmark datasets have also been produced and shared with the community related to this task. In this paper, we focus on the French MEDIA SLU dataset, distributed since 2005 and used as a benchmark dataset for a large number of research works. This dataset has been shown as being the most challenging one among those accessible to the research community. Distributed by ELRA, this corpus is free for academic research since 2020. Unfortunately, the MEDIA dataset is not really used beyond the French research community. To facilitate its use, a complete recipe, including data preparation, training and evaluation scripts, has been built and integrated to SpeechBrain, an already popular open-source and all-in-one conversational AI toolkit based on PyTorch. This recipe is presented in this paper. In addition, based on the feedback of some researchers who have worked on this dataset for several years, some corrections have been brought to the initial manual annotation: the new version of the data will also be integrated into the ELRA catalogue, as the original one. More, a significant amount of data collected during the construction of the MEDIA corpus in the 2000s was never used until now: we present the first results reached on this subset — also included in the MEDIA SpeechBrain recipe — , that will be used for now as the MEDIA test2. Last, we discuss evaluation issues.

Keywords: Spoken Language Understanding, Benchmark dataset, Deep learning tools

1. Introduction

Spoken Language Understanding (SLU) refers to natural language processing tasks related to semantic extraction from the speech signal. It is "*a field in the intersection of speech processing, natural language processing by leveraging technologies from machine learning and artificial intelligence*" (Tur and De Mori, 2011). For instance, in a human-machine spoken interaction system, such a task aims to convert a user input into a semantic representation according to the user's intention and the target application domain. While treating an SLU task, three questions need to be answered: 1) how is represented the semantic of the domain? 2) Which method is the most adapted to automatically extract the semantic from speech and to project it into the targeted semantic representation? 3) How to evaluate the resulting system?

The semantic representation is dependent on the targeted task of the software application. In most available corpora, this semantic representation can be constructed from semantic concepts supported by words or sequences of words. For example, in ATIS (Air Travel Information System), a very popular corpus (Hemphill et al., 1990), the task is dedicated to air travel planning scenarios, and semantic labels to retrieve are dedicated to this task. The semantic representation is also represented by 17 intents (aircraft, airport, distance, flight, meal...) that correspond to semantic frames, and by semantic slots associated with each frame (depart_date.relative, arrive_date.relative). Other corpora exist for SLU (Coucke et al., 2018; Shah et al., 2018; Lugosch et al., 2021), and each one proposes its own semantic representation, usually based on a list of specific labels like intents or concepts.

Different approaches have been proposed to process SLU tasks. During the last two decades, the main approaches were based on the use of machine learning algorithms. In the 2000s, such approaches were based on different kinds of generative and discriminative algorithms (Raymond and Riccardi, 2007). Until the emergence of deep learning, conditional random fields (CRF) was the most popular tool – since it reached the best results – used to process SLU tasks redefined as word labelling tasks (Hahn et al., 2010). The most recent approaches are now based on neural network architectures: recurrent neural architectures (Mesnil et al., 2013; Kurata et al., 2016; Dupont et al., 2017), encoder-decoder neural networks with attention mechanisms (Simonnet et al., 2017; Li et al., 2018). The latest approaches are based on language representation models pre-trained through self-supervised learning, such as BERT (Devlin et al., 2019). Such models have been shown to achieve state-of-the-art results in different SLU tasks (Korpusik et al., 2019; Ghannay et al., 2020). All the SLU models cited above are based on a cascade approach: first an ASR (Automatic speech recognition) system is used to automatically transcribe the user utterance, and then a NLU (Natural Language Understanding) model takes the automatic transcription as input to extract semantic tags. The intermediate transcription may contain recognition errors, and the NLU module has to deal with these errors. End-to-end approaches were proposed in order to skip the use of an intermediate speech transcription, and to avoid ASR errors propagation. In addition, end-to-end approaches permit to optimize the entire model to the

final task, while cascade approaches need to optimize each module on a sub-task. An end-to-end approach is based on the use of a single system directly optimized to extract semantic concepts from the speech. Such SLU end-to-end systems can be trained to generate both recognized words and semantic tags (Ghannay et al., 2018; Desot et al., 2019; Dinarelli et al., 2020; Evain et al., 2021).

Once cascade or end-to-end systems are built, the crucial issue of the evaluation appeared. This evaluation depends on the complexity of the semantic representation. In an intent detection task, accuracy, precision and recall are sufficient to evaluate the performances. In a more complex semantic task, like in slot filling, at least two aspects have to be considered : the semantic label and its value. The Concept Error Rate (CER) and the Concept Value Error Rate (CVER) are introduced in that case.

In this paper, we focus on the French MEDIA SLU dataset, known as being the most challenging one among the ones accessible to the research community (Béchet and Raymond, 2019). Unfortunately, the MEDIA dataset is not really used beyond the French research community. To facilitate its use, we present a complete recipe of an end-to-end neural architecture, including data preparation, training and evaluation scripts, that has been built and integrated to SpeechBrain, an already popular open-source and all-in-one conversational AI toolkit based on PyTorch. By integrating this recipe to SpeechBrain, we expect to make the MEDIA benchmark more accessible to researchers, and to make the source code persistent through a community maintenance. In addition, based on the feedback of some researchers who have worked on this dataset for several years, we also brought manual corrections to the initial manual annotations. Last, a significant amount of data collected during the construction of the MEDIA corpus in the 2000s was never used until now: we present the first results reached on this subset — also included in the MEDIA SpeechBrain recipe — that will be used for now as the MEDIA test2.

2. The Original MEDIA Benchmark

The French MEDIA benchmark (Bonneau-Maynard et al., 2005) was created as a part of the Technolangue project of the French government in 2002. It is dedicated to semantic extraction from speech in a context of human-machine dialogues for a hotel room reservation task with touristic information. 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 data is distributed by ELRA as the *MEDIA Evaluation Package*[1] [2].

2.1. Data

The MEDIA corpus is composed of telephone dialogue recordings with their manual transcriptions and their semantic annotations. It was recorded using a Wizard-of-Oz (WoZ) method (Green and Wei-Haas, 1985; Dahlbäck et al., 1993): a human (the "Wizard"), pretends to be a computer, while the user is made to believe that he is interacting with an intelligent machine. This results in 1258 official recorded dialogues, from 250 different speakers. The semantic annotations are only available for users turns. The original dataset is split into three official parts (train, dev and test) as described in tables 1, 2 and 3. As shown in these tables, a consequent part of the data included in the MEDIA package has not been used during the official campaign in 2005. The reason is this data was finalized after the end of the campaign. As a consequence, even if this data is present in the archive distributed by ELRA, it is not listed in the official data files and is hidden among the subdirectories that structure the MEDIA archive file system. On our knowledge, this data was never used in research work until now.

Data	Nb. Utterances	Nb. Turns	Nb. Dialogues
train	13.7 k	13.0 k	727
dev	1.4 k	1.3 k	79
test	3.8 k	3.5 k	208
unused	4.0 k	3.8 k	244

Table 1: Original MEDIA dataset distribution considering only the user's utterances.

Data	Nb. Hours	Mean Duration	Median Duration
train	16h56m	4.69s	3.12s
dev	01h40m	4.77s	2.79s
test	04h47m	4.89s	3.34s
unused	05h35m	5.30s	3.86s

Table 2: Original MEDIA time statistics of the user with mean and median duration of their utterances.

Data	Nb. Hours	Mean Duration	Median Duration
train	42h10m	209s	194s
dev	03h37m	165s	158s
test	11h34m	200s	190s
unused	14h30m	214s	196s

Table 3: Original MEDIA time statistics on global recordings (user, WoZ and blanks) with mean and median of the recordings.

The semantic dictionary defined for the MEDIA project includes 83 basic attributes – including 73 database attributes, 4 modifiers, and 6 general attributes – and 19 specifiers (Bonneau-Maynard et al., 2006): *room-number, hotel-name, location* are examples of database attributes,

[1] http://catalog.elra.info/en-us/repository/browse/ELRA-E0024/

[2] International Standard Language Resource Number: 699-856-029-354-6

comparative, relative-distance are examples of modifiers, *proposition-connector* and *attribute-connector* are examples of general attributes, and *address, travel* are examples of specifiers, that specializes the attribute role in a dialogue. Some complex linguistic phenomena, like co-references, are also managed thanks to this mechanism. By combining attributes and specifiers, the total number of possible attribute/specifier pairs is 1121.

These attributes are supported by words or sequence of words, now called **word support**. For each occurrence of an attribute in the semantic annotation, two other pieces of information are provided in addition to the attribute name: the mode and the normalized value of this occurrence. Four modes are possible: affirmative '+', negative '-', interrogative '?' or optional '~' .

The following sentence (translated from French) is an utterance extracted from the MEDIA dataset: "I would like to book one double room in Paris up to one hundred and thirty euros". It will be annotated as a sequence of quadruplets (word support, mode, attribute name, normalized value) as (I would like to book, +, *reservation*, `reservation`), (one, +, *room-number*, `1`), (double room, +, *room-type*, `double room`), (up to, +, *comparative-payment*, `less than`), (one hundred and thirty, +, *amount-payment*, `130`), (euros, +, *currency-payment*, `euro`).

The study proposed by Béchet and Raymond (2019), revealed 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).

2.2. Evaluation

Many metrics can be used to evaluate SLU systems. The evaluation metric adopted during the official MEDIA evaluation campaign in 2005 (Bonneau-Maynard et al., 2006) was called the *understanding error rate*. It consists on aligning – thanks to the Levenshtein distance – the hypothesis semantic representation to the reference one, and to compare them in terms of deletion, insertion, and substitution. This scoring considers as units the triplets (mode, attribute name, normalized value) presented above. The computation of the *understanding error rate* is the same as the one used to compute the *word error rate* (WER) for automatic speech recognition, by considering each triplet (mode, attribute name, normalized value) as a word.

Based on these semantic elements, different scoring processes have been used in the official MEDIA evaluation campaign: the *full scoring* takes into account the whole set of attributes (1121 possibilities), while the *relax scoring* does not consider the specifiers (83 possibilities). In addition, the 'mode' can also be reduced to the binary choice 'negative/affirmative' instead of retaining the four initial possibilities.

2.3. Issues

Evolution in the use of metrics As far as we know, after the original MEDIA evaluation campaign in 2005, only one study continued using the *understanding error rate* in relax or full scoring scenarios and 2 or 4 modes (Lehuen and Lemeunier, 2010). In other works related to the MEDIA benchmark, a simplification of the evaluation task has been done. Raymond and Riccardi (2007) introduced the *concept error rate* (CER) as a scoring metric for MEDIA. The CER is similar to the *understanding error rate*, by limiting the reference/hypothesis alignment to the attribute names only, now called as **concept**. This metric then became the *de facto* metric in MEDIA and has been used in several research works, as in (Hahn et al., 2010; Dinarelli et al., 2020; Ghannay et al., 2018).

Hahn et al. (2010) jointly evaluate the recognized concepts and the normalized value of each concept occurrence. In (Simonnet et al., 2017; Simonnet et al., 2018), the authors named this metric the *concept-value error rate* (CVER), also used in addition to the CER. The CVER is an extension of CER, which considers the correctness of the concept/value pair. Following Simonnet et al. (2017) and Simonnet et al. (2018) works, the joint CER and CVER metrics have been adopted in several recent studies for MEDIA benchmark (Caubrière et al., 2019; Ghannay et al., 2021; Pelloin et al., 2021)

Concept value normalization To normalize word supports into values, Hahn et al. (2010) proposed three possible ways: 1) hand-crafted rules obtained with the training data, needing a human expert effort; 2) stochastic approaches based on Deep Belief Network or Conditional Random Field; 3) a combination of stochastic approaches and human rules. They concluded that the use of human rules outperforms the results obtained with just stochastic approaches. The main reason was the numerous possible values for some concepts with open values like *date*, *payment-amount* or *name-client*, in conjunction with the small size of the training data. As a result, most of latest works considering the values during the evaluation used the script based on human rules.

By using these rules, the CVER should be equal to 0% when evaluating the reference. However, we obtain a CVER of 4.7% on the dev and 5.7% on the test corpus. All these errors are substitutions, i.e. support words that are not or badly normalized.

Pelloin et al. (2021) has introduced a way of obtaining automatically the normalized value directly in the output sequence. They used an end-to-end encoder-decoder system with Attention Mechanism to output the sequence of concept/value pairs from the audio signal. They obtained very good results but similarly to Hahn et al. (2010), the same rule-based system designed to normalize concept values obtained better results, even if this normalization is not robust to speech recognition errors and could be perfectible. However, we think that the automatic value normalization process is still a very important issue that could be addressed by researchers in order to propose new approaches that do not use human expertise. Such data-driven solutions could speed up and reduce the cost of the deployment of other human-machine applications in new domains.

In order to get an additional information about the values brought by word sequences supporting the concepts, but which is independent to the value normalization process, we propose in this paper a new evaluation measure: the *u-CVER*, which takes into account the unnormalized values,

instead of the normalized ones.

Error correction in manual annotation During the MEDIA project, a complete semantic annotation scheme have been proposed to the human annotators. As natural language is subject to interpretation and humans are error-prone, some semantic annotation errors remain, leading to take them into account during training and also leading to false errors during the evaluation. Furthermore, we detected some problems in the audio segmentation. For example, there are very long segments at the end of the dialogues when the user has already hung up the phone. We propose to correct some of these errors detected during several analyses of the corpus with the proposed new updated version of MEDIA.

Data preparation The manual transcription in the MEDIA corpus is very precise and takes into account some details of pronunciation. For instance, disfluences like false starts or truncated words are annotated by using parentheses or asterisks, *e.g.*: "an (ho)tel", that means that the "ho" was not pronounced. Such annotations are not processed in the same way by the different authors, and this can make the comparisons between published experimental results hard to interpret. For the sake of transparency, we share our script for data preparation in the SpeechBrain recipe we have implemented.

Integrating the unused data As seen in subsection 2.1., an important part of collected and manual annotated data was never used before now. We suggest to integrate this data in our recipe – including data preparation – and share in this paper the first experimental results got on it.

3. Data Updates

Corrections have been brought to the initial manual annotations: the new version of the data will be integrated into the ELRA catalogue, as the original one. We started from the already distributed files of the MEDIA dataset to generate the new ones.

3.1. Correction of Manual Annotation

First, a simple normalization has been done on the transcription itself. We removed multiple spaces, corrected apostrophe and hyphen connections to their words, and added uppercase to nouns when forgotten. We also corrected some spelling of words and some erroneous semantic labelling. The audio channel – left or right – in which the user's voice has been recorded was not well indicated. It will now be indicated next to the recording ID. The ID of some users did not always respect the expected format: we corrected this also.

3.2. The MEDIA test2 Dataset

As presented in section 2.1., we have discovered an available but unused manually annotated data in the MEDIA corpus distributed by ELRA. We have decided to use it to create a new test corpus, named *test2* detailed in Tables 1, 2 and 3, on the lines entitled "unused data". The *test2* is even bigger than the original test, making it greatly interesting and useful. This corpus is far similar to the original *test* one. Only a "Full labelling" (attributes+specifiers) manual annotation of the semantic concepts has been realized in *test2*, we automatically generated the "Relax labelling" (attributes only) by removing the specifier parts of the concepts (following the annotation guide). Only one concept was never seen in the other sets: "personne-prénom", standing for "person-firstname". This semantic concept, appearing only once, brought the discussed CVER evaluation problem to the surface.

3.3. Data Statistics

In this section, we present the new MEDIA statistics for all datasets, including *test2*. While Table 4 summarizes statistics about words and truncated words (*cf. Section 2.3. - data preparation*) considering only the user's utterances, Table 5 presents statistics on concept occurrences and lexicon. Notice that the number of concept occurrences is the same in Full and Relax scoring. Indeed, the number of word supports remains unchanged, only the lexicon of the considered concept labels is different.

Data	Occurrences		Lexicon	
	Nb. Words	Nb. Trunc.	Nb. Words	Nb. Trunc.
train	92.6 k	820	2.3 k	372
dev	10.5 k	134	0.8 k	89
test	26.0 k	227	1.4 k	146
test2	28.0 k	159	1.3 k	107

Table 4: New MEDIA number of words and truncated words, considering occurrences and lexicon in user's turns.

Concept	Occurrences	Lexicon	
	Full and Relax	Full	Relax
train	31.7 k	144	73
dev	3.3 k	104	63
test	8.8 k	125	71
test2	9.4 k	129	71

Table 5: New MEDIA number of occurrences of concepts and size of the concept lexicon, considering Full and Relax scorings.

Taking a closer look at the data, we observe that the most common word is "oui", standing for "yes". It is used 3.8% of the time in total. The most common semantic concepts are "response" and "command-task" (when the user asks to book a room). They appear respectively 18.4% and 6.6% of the time in total.

4. The MEDIA SpeechBrain Recipe

A complete recipe for the MEDIA corpus has been built and integrated to SpeechBrain [3], an open-source conversational AI toolkit based on PyTorch. SpeechBrain is a user-friendly toolkit proposing multiple recipes, ready to train. By integrating our recipe to SpeechBrain, we

[3] https://github.com/speechbrain/speechbrain/tree/develop/recipes/MEDIA

expect to make the MEDIA benchmark more accessible to researchers, and to make the source code persistent through a community maintenance.

The MEDIA recipe is available for running either an ASR task or a SLU task (the latest being the former without considering semantic concepts in the output of the system). It is based on end-to-end architectures, and can be trained by fine-tuning a wav2vec 2.0 model (Baevski et al., 2020). wav2vec 2.0 learns speech representations through self-supervision learning from large amounts of speech data, using Convolutional Neural Networks (CNN) and a masked Transformer. After pre-training, this model can then be fine-tuned through supervised learning on labelled data for ASR or SLU task, depending on the final task.
The recipe includes data preparation, training and evaluation scripts. They are detailed in the next subsections.

4.1. Data Preparation

To use the MEDIA SpeechBrain recipe, remind that it is necessary to get the original MEDIA data, distributed by ELRA, beforehand. By running an experiment, the data preparation scripts (if asked in the launching command) will create SpeechBrain compatible *csv files* to train the model. These csv files contain information like the utterance id, the manual annotation, the audio file pathname. Note that some options are available in the recipe in order to sort or remove utterances according to their duration. Indeed, experiments done by many users of the SpeechBrain toolkit have shown that the ascending sorting can be more efficient for ASR or SLU.
All special characters, except chevrons, hyphens and apostrophes have been removed from the csv files. The chevrons are used to mark the beginning of word support, indicating the semantic concept, and the ending of the word support is marked with an ending chevron. For example, "<task-command> I want to book > hum <room-type> a double room >". For the apostrophe, it has been decided to bond it to the preceding word, but not to the following one. For instance, "d'Avignon" would now be written "d' Avignon". The only exception is for the word "c'est" (*i.e. it is*), which is far too common in French to divide it and since the word "c'" does not exist in another form.
Further processing have been made to the MEDIA dataset to be optimized for experiments. Among the removed symbols, asterisks were used to specify very close words in the speech. For example, a user who would say "how are you" very quickly could be transcribed "how* are* you" in the original MEDIA because of a strong assimilation during the co-articulation of these words. It will be transcribed "how are you" after the data preparation process of our recipe.
Concerning the truncated words, round brackets and their content have been replaced by asterisks in order to still have an indication about the *not pronounced* part of the word but also that this word does not exists in the French vocabulary. It also prevents the annotator from misinterpreting any possible word that the user wanted to pronounce. As a result with an example, "*exam(ple)*" written in the original MEDIA corpus is written "*exam**" in the new csv files.
The last modification brought to the transcriptions during the data preparation process was removing hyphens between numbers written in letters. In French, "soixante-dix" (meaning 70) is as much understandable as "soixante dix". Those characters only increase the final vocabulary, but not really adding sense to the transcription.
In the original MEDIA xml files, some synchronization tags are present, in addition to time code related to speaker turns. These synchronization tags are helpful to split a speaker turn into utterances. They have been marked by the human annotators of the original MEDIA corpus. A processing script enables to take some into account in order to reduce the length of utterances.
All the synchronisation tags narrowing the time limits, or cutting the transcription without splitting a semantic annotation labelling a word support, were used.
Table 6 makes an update on actual hours of recordings and speech (for users only, not WoZ), thanks to the segmentation we applied from the manual annotation.

Data	Nb. Hours	Mean Duration	Median Duration
train	10h52m	2.85s	1.69s
dev	01h13m	3.23s	1.91s
test	03h01m	2.88s	1.70s
test2	03h16m	2.94s	1.93s
total	18h22m	2.90s	1.75s

Table 6: Statistics of the user's utterances with mean and median duration after processing the new segmentation script.

4.2. Neural Architecture

The recipe is based on the use of wav2vec 2.0 models. We used LeBenchmark models (Evain et al., 2021) such as the Wav2Vec2-FR-3k large. This model was pretrained through self-supervised learning on 3k hours of speech (mainly read speech and broadcast news) in French language. The LeBenchmark models are freely shared with the community. On the top of the Wav2Vec2-FR-3k large, we added 3 dense layers of 512 neurons, with the LeakyReLU as activation function. These layers are themselves followed by one fully-connected layer and a final softmax layer. The weights of these four additional layers are randomly initialized, while the other weights are initialized by using the pretrained weights in the wav2vec 2.0 part of the neural architecture. As input, the neural network receives a wav audio file sampled at 16 kHz, and the output are characters that cover the alphabet needed to spell all the words of the MEDIA training data, and additional characters that manage the opening and closing tags used to recognize the concepts. After processing through the softmax layer, the outputs are generated thanks to a simple greedy decoder.

4.3. Training Process

The training is done in supervised manner from the semantically labelled MEDIA training data. This can be considered mainly as a fine-tuning of the wav2vec 2.0 model on

the final downstream task.
We also proposed in the recipe an alternative that consists of first fine-tuning the wav2vec 2.0 model on an external audio data, in order to achieve an ASR task. To make the experiments reproducible and accessible, we used the CommonVoice French dataset[4] (version 6.1), collected by the Mozilla Foundation. The train set consists of 425.5 hours of speech, with around 24 hours for the validation and test sets. After removing and re-initializing the weight matrix of the last layer, we finish by fine-tuning the model in the same way as the first solution, on the MEDIA semantically labelled training data.
The optimizer for the wav2vec 2.0 model is Adam, with a learning-rate of 0.0001, and AdaDelta for other layers, with a learning-rate of 1 and momentum of 0.95. Utterances were sorted in ascending order. We only did 30 epochs for the results presented in section 4.5..
Two kinds of models are considered in the following section:

- The models fine-tuned directly on the MEDIA training data to process the targeted SLU task. These models are named **media-base**.
- The models that are first fine-tuned on the CommonVoice data for an ASR task, then fine-tuned on the MEDIA training data to process the targeted SLU task. These models are named **media-comvoice**.

4.4. Evaluation

We integrated in our recipe several evaluation metrics. As proposed in the original MEDIA campaign, we align the reference and the hypothesis to compute the number of errors in terms of deletions, insertions and substitutions. Several components are then considered in the reference/hypothesis. The recipe proposes three types of metrics to evaluate the performance of the neural models:

- the character error rate **ChER** was already integrated in SpeechBrain. It considers all characters in the output, considering also the semantic concept but as only one single character.
- the concept error rate **CER**, as proposed by Hahn et al. (2010), compares the sequence of concepts only. Each token is considered as one unit, not each character like in ChER. Lets take again the example "<task-command> I want to book > hum <room-type> a double room >". The evaluated hypothesis will be "task-command room-type".
- we introduce a new measure: the unnormalized concept value error rate **u-CVER**. The words appearing in a word support for a concept are concatenated in one single word and also concatenated to the concept. The previous example will be presented as "task-command_I_want_to_book room-type_a_double_room". As a consequence, if any of the characters in the word support prediction or the concept tag prediction differs from the reference, then the entire prediction of that concept is counted as an error.

For the sake of comparison, in this paper we also present results with the CVER, when the CVER is computed by the human-rules system as proposed in (Hahn et al., 2010). In order to prevent ambiguity, we rename this metric **r-CVER** (r- for 'rule-based'). Here, the previous example will be evaluated considering the sequence "task-command_booking room-type_double".
Last, while the "Full scoring" was still not used in the last published papers related to the MEDIA corpus, we reintroduce this scoring since we consider it makes the challenging MEDIA benchmark even more challenging. Thus, our recipe includes both "Relax" and "Full" scorings.

4.5. First Experimental Results

An experiment with the recipe has been launched to obtained the models described in Section 4.2.. Notice that for each kind of scoring (Full or Relax), a specific model was trained (as the semantic lexicon considered for the output is different). Thus, for "Full" or "Relax" scorings, two different **media-base** models were trained, the same for **media-comvoice**.
The results on the new Media, considering both Full and Relax scorings, are presented in Table 7 for the two kinds of neural models present in our MEDIA SpeechBrain recipe. As the rules based system (Hahn et al., 2010) we used is only available for the Relax scoring, we do not compute the *r-CVER* for the FULL scoring.
We can see that the recipe reaches a little under end-to-end system's state-of-the-art results with 16.3% of CER and 23.7% of r-CVER on the test. This proves the recipe is operational and simply needs tuning to enhance the results. As expected, the proposed metrics *u-CVER* is stricter than the *r-CVER*.
In Table 8, the results obtained by our best systems on the new test2 dataset are presented. These systems are based on the **media-comvoice** models. We can also observe that the already analysed test corpora are quite similar, and that the test2 is well suited to the task. Furthermore, while the original test dataset has been used by researchers for over fifteen years, the results of the test2 appear to show that the models applied to the original test corpus have not been implicitly "overfitted" during this time.
In a future use of a model trained on the MEDIA dataset, we need to be able to evaluate the results obtained by using a better normalization process for concept value than the current rule-based one.

5. Conclusion

In this paper, we present some updates to the MEDIA benchmark dataset for spoken language understanding. This update will also be integrated into the ELRA catalogue, as the original dataset. We expect to facilitate the use of this very challenging dataset, which became free for academic research in 2020. We also present a complete recipe, including data preparation, training and evaluation scripts, built and integrated to SpeechBrain, an already popular open-source and all-in-one conversational AI toolkit based on PyTorch. Last, a significant amount of data collected during the construction of the MEDIA corpus in the

[4]https://commonvoice.mozilla.org/fr/datasets

Scoring	Model	dev				test			
		ChER	CER	u-CVER	r-CVER	ChER	CER	u-CVER	r-CVER
Full	**media-base**	8.4	28.9	41.2	-	8.2	26.1	37.5	-
	media-comvoice	7.2	24.0	34.4	-	6.9	20.3	30.8	-
Relax	**media-base**	8.1	23.3	37.1	28.9	7.9	21.8	34.1	29.4
	media-comvoice	6.8	18.1	30.4	23.3	6.7	16.3	27.7	23.7

Table 7: Results on the new MEDIA data of the **media-base** and **media-comvoice** models on both "Full" and "Relax" scoring mode.

Scoring	test2			
	ChER	CER	u-CVER	r-CVER
Full	6.7	21.1	30.9	-
Relax	6.4	16.4	27.1	21.0

Table 8: Results on test2 corpus with the **media-comvoice** models.

2000s was never used until now: we present the first results reached on this subset — also included in the MEDIA SpeechBrain recipe — that can be used for now as the MEDIA test2.

We expect a growing community will use our recipe to start working on the MEDIA corpus. While the research community is more and more interested by SLU problems and benchmarks, MEDIA stays one of the most challenging corpus, even in the era of deep learning. For this reason, this corpus constitutes a relevant dataset to investigate new solutions that can have a real impact on such human/machine application.

6. 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.

7. Bibliographical References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

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*.

Bonneau-Maynard, H., Ayache, C., Bechet, F., Denis, A., Kuhn, A., Lefevre, F., Mostefa, D., Quignard, M., Rosset, S., Servan, C., and Villaneau, J. (2006). Results of the French evalda-media evaluation campaign for literal understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Caubrière, A., Tomashenko, N., Laurent, A., Morin, E., Camelin, N., and Estève, Y. (2019). Curriculum-based transfer learning for an effective end-to-end spoken language understanding and domain portability. In *20th Annual Conference of the International Speech Communication Association (InterSpeech)*, pages 1198–1202.

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*.

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies—why and how. *Knowledge-based systems*, 6(4):258–266.

Desot, T., Portet, F., and Vacher, M. (2019). Towards end-to-end spoken intent recognition in smart home. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. 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.

Dinarelli, M., Kapoor, N., Jabaian, B., and Besacier, L. (2020). A data efficient end-to-end spoken language understanding architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8519–8523. IEEE.

Dupont, Y., Dinarelli, M., and Tellier, I. (2017). Label-dependencies aware recurrent neural networks. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 44–66. Springer.

Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., et al. (2021). Task agnostic and task spe-

cific self-supervised learning from speech with lebenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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.

Green, P. and Wei-Haas, L. (1985). The rapid development of user interfaces: Experience with the wizard of oz method. *Proceedings of the Human Factors Society Annual Meeting*, 29(5):470–474.

Hahn, S., Dinarelli, M., Raymond, C., Lefevre, F., Lehnen, P., De Mori, R., Moschitti, A., Ney, H., and Riccardi, G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1569–1583.

Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990). The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Korpusik, M., Liu, Z., and Glass, J. (2019). A comparison of deep learning methods for language understanding. In *Interspeech, September 15–19, 2019, Graz, Austria*, Graz, Austria, September 15–19.

Kurata, G., Xiang, B., Zhou, B., and Yu, M. (2016). Leveraging sentence-level information with encoder lstm for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2077–2083.

Lehuen, J. and Lemeunier, T. (2010). A robust semantic parser designed for spoken dialog sytems. *2010 IEEE Fourth International Conference on Semantic Computing*, pages 52–55.

Li, C., Li, L., and Qi, J. (2018). A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833.

Lugosch, L., Papreja, P., Ravanelli, M., HEBA, A., and Parcollet, T. (2021). Timers and such: A practical benchmark for spoken language understanding with numbers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Mesnil, G., He, X., Deng, L., and Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

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.

Raymond, C. and Riccardi, G. (2007). Generative and Discriminative Algorithms for Spoken Language Understanding. In *Interspeech 2007 - 8th Annual Conference of the International Speech Communication Association*, Anvers, Belgium, August.

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*.

Simonnet, E., Ghannay, S., Camelin, N., Estève, Y., and De Mori, R. (2017). ASR error management for improving spoken language understanding. In *Interspeech 2017*, Stockholm, Sweden, August.

Simonnet, E., Ghannay, S., Camelin, N., and Estève, Y. (2018). Simulating ASR errors for training SLU systems. In *LREC 2018*, Miyazaki, Japan, May.

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