Albayzin Evaluation: IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge

Eduardo Lleida[1], Alfonso Ortega[1], Antonio Miguel[1], Virginia Bazán[2], Carmen Pérez[2], Manuel Gómez[2], and Alberto de Prada[2]

[1] Vivolab, Aragon Institute for Engineering Resarch (I3A)
University of Zaragoza, Spain
{ortega,ivinalsb,amiguel,lleida}@unizar.es
http://www.vivolab.es
[2] Corporación Radiotelevisión Española, Spain http:www.rtve.es

Abstract. The IberSPEECH-RTVE Speech to Text Transcription is a new challenge in the ALBAYZIN evaluation series. The evaluation is supported by the *Spanish Thematic Network on Speech Technology* (RTTH) and *Cátedra RTVE Universidad de Zaragoza* and is organized by ViVoLab Universidad de Zaragoza.The evaluation will be conducted as part of the Iberspeech 2018[3] conference to be held in Barcelona, Spain from 21 to 23 November 2018.

1 Introduction

The IberSPEECH-RTVE 2018 Speech to Text Transcription Challenge aims to evaluate Automatic Speech Recognition (ASR) systems in realistic TV shows. The task will evaluate state of the art ASR technology to be used for applications as subtitling and automatic metadata generation for audiovisual content. Subtitling is the process by which we get a transcription of the audio portion of a program. Automatic metadata generation for audiovisual content is the process by which we analyze the content of the audiovisual document to archive, retrieve and filter audio-visual segments (for example, a special interview), objects (a special person) and events (a special goal in a football match)[1].

Tremendous progress has been observed during the last years in the performance of ASR systems. However they still entail errors, mainly due to challenging acoustic conditions, speaking rate, spontaneous speech, out-of-vocabulary words or language ambiguities. The resulting errors are of varying importance depending on the application in which the ASR system is being used. The most common measure of the ASR performance is the word error rate (WER). The WER is the edit distance between a reference word sequence and its automatic transcription. However, WER does not consider whether some words may be more important to the meaning of the message. In fact, humans perceive different ASR errors as having different degrees of impact on a text. The ASR errors

[3] http://iberspeech2018.talp.cat

have different impact on both application, subtitling and automatic metadata generation. Usually, subtitling needs a closer verbatim transcription than automatic metadata generation as in the later the goal is to retrieve the relevant information present in the audiovisual document. These differences lead to different ways of measuring the performance of ASR systems. In this challenge, we will use word error rate (WER) as primary scoring measure but we will explore the use of other measures as Word Information Loss [2], which is more suitable than WER for the evaluation of any application in which the proportion of word information communicated is more meaningful than edit cost or Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[3] widely used for text summarization and machine translation. We intent to use ROUGE measures to compare ASR transcription against reference subtitles.

2 Challenge Description and Databases

The Speech to Text transcription evaluation consists of automatically transcribe different types of TV shows. For this evaluation, RTVE has licensed around 569 hours of own TV production jointly with the corresponding subtitles. The shows cover a great variety of scenarios from scripted content to live broadcast, from read speech to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents. Some of the contents have been labeled thanks to the Spanish Thematic Network on Speech Technology (RTTH) and Cátedra RTVE en la Universidad de Zaragoza.

RTVE2018[4] database has a total of 569 hours and 22 minutes of audio. About 460 hours are provided with the subtitles and about 109 hours have been human-revised transcribed. Be aware that in most of the cases, subtitles could not contain an verbatim word transcription as most of them have been generated by a re-speaking procedure.

The database has been divided into 4 partitions, a *train* one, two development partitions *dev1*, *dev2* and finally a *test* partition. Additionally, the database includes a set of text files extracted from all the subtitles broadcasted by the RTVE 24H Channel during 2017.

Detailed information about the RTVE2018 database content can be found in the RTVE2018 database description report `http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf`. Here we give a simple description of the database partitions.

2.1 Training and Development data

The train partition consists of all the audio files without human-revised transcriptions, which means that only subtitles are available.

For development, two partitions have been defined. Partition *dev1* contains about 53 hours of audios and their corresponding human-revised transcriptions

[4] `http://catedrartve.unizar.es/reto2018.html`

and partition *dev2* with about 15 hours of audios and their corresponding human-revised transcriptions and speaking-turns timestamps. For this challenge, both partitions can be used for either development or training.

Training conditions The ASR systems can be evaluated over a *closed* or *open* training condition.

- **Closed condition -** The *closed* condition limits the system training to the use of the training and development dataset of the RTVE2018 database. The use of pretrained models on data other than RTVE2018 is not allowed in this condition. Participants can use any external phonetic transcription dictionary.
- **Open condition -** The *open* training condition removes the limitations of the *closed* condition. Participants are free to use RTVE2018 training and development set or any other data to train their systems provided that these data are fully documented in the systems description paper.

Reference result. As reference of the performance of state-of-the-art commercial ASR, we provide the margin of WER values obtained for a TV show with different commercial ASR systems. The TV show is the one corresponding with file LM-20171107.acc in *dev1* partition. Using the primary metric defined in 3.1 with the LM-20171107.stm reference file, the WER is in the range of 22% to 27%.

2.2 Evaluation data

The evaluation data will contain a set of TV shows covering a variety of scenarios. RTVE2018 database includes a *test* partition with all the files needed to evaluate ASR systems. The detailed information about the *test* partition will be released along the evaluation data by September 24.

3 Performance Measurement

ASR system output will be evaluated with different metrics but a primary metric will be used for ranking ASR systems. All the participants will provide as ASR output for evaluation a free-form text with no page, paragraphs, sentence or speaker breaks with *.txt* extension using the utf-8 charset per test file. The text may include punctuation marks to be evaluated with an alternative metric. An example can be found in the *doc* folder of the RTVE2018 database.

3.1 Primary metric

Word Error Rate (WER) will be the primary metric for the Speech to Text Transcription task. The text will be normalized removing all the punctuation marks, numbers will be written with letters and text will be lowercased. The WER is defined as

$$WER = \frac{S + D + I}{N_r} \tag{1}$$

where $\mathrm{N_r}$ is the total words in the reference transcription, S is the number of substituted words in the automatic transcription, D is the number of words from the reference deleted in the automatic transcription and I is the number of words inserted in the automatic transcription not appearing in the reference. WER will be computed using the sclite tool included in the NIST Speech Recognition Scoring Toolkit (SCTK[5]). To use sclite tool it is necessary to translate the reference transcription files to any sclite reference format. Sclite accepts as reference files a variety of formats[6]. In this evaluation, we will use the *stm* format as reference. The stm format describes the segment time marked files consisting of a concatenation of text segment records from a waveform file. Each record is separated by a newline and contains: the waveform's filename and channel identifier [A|B], the talkers ID, begin and end times (in seconds), optional subset label and the text for the segment. The stm files are built from the transcription files (trn) using dummy segment time marks. Hypothesis files will be simply free-form text with no page, paragraphs, sentence or speaker breaks with *.txt* extension.
Here is an example of stm file:

20H 1 Presentador1 2079.102 2086.618 <,,> El premio se les concedió por sus descubrimientos sobre los mecanismos moleculares que controlan los ritmos cardiacos
20H 1 Presentador2 2086.642 2092.578 <,,> En la información que van a ver a continuación van a intentar explicar qué es exactamente eso .
20H 1 Voz_off8 2093.900 2101.040 <,,> Los ritmos circadianos podrían traducirse popularmente como los mecanismos de nuestro reloj biológico interno

3.2 Alternative metrics

In addition to the primary metric, other alternative metrics may be computed, but not taking into account for the challenge.

Punctuation marks evaluation (PWER) - The WER is computed with the punctuation marks given by the ASR system.

Text Normalized Word Error Rate (TNWER) - text normalization techniques as stopword removal and lemmatization are applied to the ASR output. In this sense, common errors as verbal conjugations, gender or number substitutions, articles, determiners, and quantifiers deletion/insertions will not have impact on the ASR performance. The same text normalization will be applied to

[5] https://www.nist.gov/itl/iad/mig/tools

[6] `http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm`

both the reference and automatic transcriptions before proceeding to calculate WER. The freeling[7] lemmatizer will be used.

Word Information Loss - Word Information Loss (WIL)[2], was introduced as replacements for WER in settings where high error rates are common. WIL is a probabilistic approach that approximates the proportion of the word information lost due to the presence of errors. WIL is more suitable metric than WER for the evaluation of any application in which the proportion of word information communicated is more meaningful than edit cost. WIL metric is computed as

$$WIL = 1 - \frac{H^2}{(H+S+D)(H+S+I)} \tag{2}$$

where H as the number of correctly recognized words.

ROUGE - the Recall-Oriented Understudy for Gisting Evaluation (ROUGE)[3] is widely used for text summarization and machine translation evaluation. ROUGE is a metric used for evaluating text summarization and machine translation systems. ROUGE metrics compare an automatically produced summary or translation against a reference (human-produced) summary or translation. ROUGE is a metric based on N-gram co-occurrence statistics, it measures how much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries. We intent to use ROUGE measures to compare ASR transcription against reference subtitles.

4 Evaluation Protocol

This challenge is conducted as an open evaluation where the test data is sent to the participants who process the data locally and submit the output of their systems to the organizers for scoring.

4.1 Registration rules

The organizers encourage the participation of all researchers interested in speech to text transcription. All teams willing to participate in this evaluation must send an e-mail to

- lleida@unizar.es
- ortega@unizar.es

Indicating the following Information:

- RESEARCH GROUP:

[7] `http://nlp.lsi.upc.edu/freeling/`

- INSTITUTION:
- CONTACT PERSON:
- E-MAIL:

with CC to Iberspeech 2018 Evaluation organizers at:

- albayzinevaluations@gmail.com

before September 24th, 2018.

4.2 Data License Agreement

The RTVE data is available to the evaluation participants only and subject to the terms of a licence agreement with the RTVE. The license agreement can be downloaded from Cátedra RTVE-UZ web page:
http://catedrartve.unizar.es/reto2018.html

Participants must sign the agreement and send a scanned copy attached to the email. A copy signed by RTVE representative will be returned. Please read carefully the information provided on the Cátedra RTVE-UZ web page related with the use of the RTVE data after the evaluation campaign.

4.3 Evaluation Rules

Submission procedure. Each participant team must submit at least a primary system in one condition, open-set or closed-set, but they can also submit up to two contrastive systems. Each and every submitted system must be applied to the whole test database. The ranking of the evaluation will be done according to results of the primary systems but the analysis of the results of the contrastive systems will be also processed and presented during the evaluation session at Iberspeech. All participant sites must agree to make their submissions (system output, system description, ...) available for experimental use by the rest of the participants and the organizing team.

The participant teams will notify and provide the total time required to run the set of tests for each submitted system (specifying the computational resources used). No manual intervention is allowed for each developed system to generate its output, thus, all developed systems must be fully automatic. Listening to the evaluation data, or any other human interaction with the evaluation data, is not allowed before all results have been submitted. The evaluated systems must use only audio signals.

4.4 Results Submission Guidelines

The evaluation results must be presented in just one ZIP file per submitted system. The ZIP file must contain one TXT file per test audio file using utf-8 charset.
Each TXT file must be identified by the following code:
<FILENAME>_<SITE>_<SYSID>_<SET>.txt
where,

- <**FILENAME**>: Refers to the filename of the test audio file without the extension (LM-20171215)
- <**SITE**>: Refers to the acronym identifying the participant team (UPM, UPC, UVI, ...)
- <**SYSID**>: Is an alphanumeric string identifying the submitted system. For the primary system the SYSID string must begin with p-, c1- for contrastive system 1 and c2- for contrastive system 2.
- <**SET**>: Refers to the training condition *open* for open condition training or *closed* for closed condition training.

The zip output file must be identified by the following code:
<SITE>_<SYSID>_<SET>.zip

Each participant team must send an e-mail with the corresponding ZIP result files to

- lleida@unizar.es
- ortega@unizar.es

4.5 System Descriptions

Participants must send, along with the result files, a PDF file with the description of each submitted system. The format of the submitted documents must fulfil the requirements given in the IberSpeech 2018 call for papers. You can use the templates provided for the Iberspeech conference (WORD or LaTeX). Please, include in your descriptions all the essential information to allow readers to understand the key aspects of your systems.

5 Schedule

- June 18, 2018: Registration opens and release of the training data.
- July 15, 2018: Registration deadline.
- September 24, 2018: Release of the evaluation data.
- October 21, 2018: Deadline for submission of results and system descriptions.
- October 31, 2018: Results distributed to the participants.
- Iberspeech 2018 workshop: Official publication of the results.

6 Acknowledgments

The organizing team would like to thank Corporación Radiotelevisión Española and Cátedra RTVE de la Universidad de Zaragoza for their efford in providing the data for the 2018 evaluation. Thanks also to the organizing committee of Iberspeech 2018 for their help and support.

References

[1] Kohler, J., Biatov, K., Larson, M., Eckes, C., Eickeler, S., "AGMA: Automatic Generation of Metadata for Audio-Visual Content in the Context of MPEG-7", Proceedings of Cast01, 2001.

[2] Morris, A. C., Maier, V. and Green, P. D., "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.." Paper presented at the meeting of the INTERSPEECH, 2004.

[3] Lin, Chin-Yew, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp 74-81, Barcelona 2004.