*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# BERT-based Semantic Model for Rescoring N-best Speech Recognition List

*Dominique Fohr, Irina Illina*

Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
{dominique.fohr,irina.illina}@loria.fr

## Abstract

This work aims to improve automatic speech recognition (ASR) by modeling long-term semantic relations. We propose to perform this through rescoring the ASR N-best hypotheses list. To achieve this, we propose two deep neural network (DNN) models and combine semantic, acoustic, and linguistic information. Our DNN rescoring models are aimed at selecting hypotheses that have better semantic consistency and therefore lower WER. We investigate a powerful representation as part of input features to our DNN model: dynamic contextual embeddings from Transformer-based BERT. Acoustic and linguistic features are also included. We perform experiments on the publicly available dataset TED-LIUM. We evaluate in clean and in noisy conditions, with n-gram and Recurrent Neural Network Language Model (RNNLM), more precisely Long Short-Term Memory (LSTM) model. The proposed rescoring approaches give significant WER improvements over the ASR system without rescoring models. Furthermore, the combination of rescoring methods based on BERT and GPT-2 scores achieves the best results.

**Index Terms**: automatic speech recognition, semantic context, embeddings, BERT

## 1. Introduction

ASR systems have made significant progress in recent years. Classical ASR systems only take into account acoustic, lexical, and syntactic information (local n-gram language models (LM)). When conditions between training and testing differ, like noisy environments, the audio signal is distorted, and the acoustic model may not be able to compensate for this variability. Even if noise compensation methods work well [10], it is of interest to incorporate *semantic knowledge* into the decoding process to help the ASR better account for the long-term semantic context and furthermore to combat adverse conditions. This improvement should also be useful when training and testing conditions match, since semantic information is important for ASR systems.

Some studies have tried to include this information into ASR. The authors of [18] use a semantic context for recovering proper names missed in the ASR process. [1] integrate semantic frames and target words into recurrent neural network LM. In [2] the re-ranking of the ASR hypotheses using an in-domain LM and a semantic parser significantly improves the accuracy of the transcription and semantic understanding. Furthermore, [6] introduce semantic grammars applicable for ASR and understanding using ambiguous context information.

Studies have shown that *rescoring the ASR N-best hypotheses list* can be an efficient solution to incorporate long-range semantic information. [20] formalize the N-best list rescoring as a learning problem and use a wide range of features with automatically optimized weights. [13][14] introduce N-best rescoring through an LSTM-based encoder network followed by a fully-connected feed-forward NN-based binary-class classifier. [19] propose a bi-directional LM for rescoring, and utilize the word prediction capability of the *BERT* [3][24].

In this work, we aim to *add long-range semantic information* to ASR through rescoring the N-best hypotheses list. We believe that some ASR errors can be corrected by taking into account distant contextual dependencies, which is important for noisy conditions. We theorize that in noisy parts of speech, the semantic model (SM) will help remove acoustic ambiguities. The core ideas of the proposed rescoring approaches are as follows. First, we use a continuous SM to represent each hypothesis: *BERT* model. Different semantic properties and efficiencies of *BERT* motivated us to explore it for our task. Second, we compare ASR hypotheses two per two and propose two BERT-based models. Finally, we propose efficient DNN architecture to train together semantic, acoustic, and linguistic information. The obtained score is combined with the ASR scores attached to each hypothesis (acoustic and linguistic) and used to rescore the ASR N-best list. Regarding [8], where semantic relations between entities are extracted from *DBpedia* and used as features for rescoring, we use transformer *BERT* model that are pre-trained on large and diverse text corpora. Compared to [13][14], we use a more powerful model with several transformer layers. Compared to [19], where masked word prediction is performed for *BERT*, we use the *sentence prediction* capability of the *BERT* model. Compared to our previous work [9], we use the *BERT* SM to represent the hypotheses at the sentence level, and train hypotheses representations by a DNN. In experiments using a publicly available speech corpus, we systematically explore the effectiveness of the proposed features and their combinations. The proposed approaches steadily outperform the baseline ASR system in clean and all noisy conditions. The proposed approaches are competitive compared to GPT-2 rescoring. This research work was carried out as part of an industrial project.

## 2. Proposed methodology

### 2.1. Introduction

A classical speech recognition system provides an acoustic score $P_{ac}(w)$ and a linguistic score $P_{lm}(w)$ for each of the hypothesized words $w$ of the utterance to recognize. The best sentence hypothesis is the one that maximizes the likelihood of the word sequence:

$$\widehat{W} = argmax_{h_i \epsilon H} \prod_{w \epsilon h_i} P_{ac}(w)^{\alpha} * P_{lm}(w)^{\beta} \qquad (1)$$

$\widehat{W}$ is the recognized sentence (the end result); $H$ is the set of $N$-best hypotheses; $h_i$ is the $i$-th sentence hypothesis; $w$ is a



http://dx.doi.org/10.21437/Interspeech.2021-313

hypothesized word. $\alpha$ and $\beta$ represent the weights of the acoustic and language models.

An efficient way to take into account semantic information is to re-evaluate (rescore) the best hypotheses of the ASR system. We propose to introduce for each hypothesis $h_i$ the semantic probability $P_{sem}(h_i)$ to take into account the semantic context of the sentence. In our rescoring approach, $P_{ac}(h_i)$, $P_{lm}(h_i)$, and the semantic score $P_{sem}(h_i)$ are computed and *combined* using specific weights $\alpha$, $\beta$ and $\gamma$ (for $P_{sem}(h_i)$) for each hypothesis:

$$\widehat{W} = argmax_{h_i \epsilon H} \; P_{ac}(h_i)^{\alpha} * P_{lm}(h_i)^{\beta} * P_{sem}(h_i)^{\gamma} \quad (2)$$

We propose to rescore using a pair of ASR hypotheses, one at a time. We use hypothesis pairs to get a tractable size of the rescoring DNN input vectors. Each hypothesis of each pair is represented by *semantic* information, produced by proposed DNN rescoring models based on *BERT* representation. We also explored the *word2vec* model [12] but given its poor performance, we have not included its results in this article.

Furthermore, we propose to go beyond a simple score combination, like in eq. (2). We propose two DNN-based rescoring models producing $P_{sem}(h_i)$: (a) the first model, called $BERT_{sem}$, is purely semantic and only uses textual information as input; (b) the second model, called $BERT_{alsem}$, takes *acoustic, linguistic,* and *textual information* as the input. We believe that the acoustic and linguistic information should be trained together with the semantic information to give an accurate rescoring model.

### 2.2. DNN-based rescoring models

Our proposed DNN models use a pair of hypotheses. For each hypothesis pair ($h_i$, $h_j$), the expected DNN *output v* is: (a) *1,* if the WER of $h_i$ is lower than the WER of $h_j$ ; (b) otherwise, 0.

The overall algorithm of the N-best list rescoring is as follows. For a given sentence, for each hypothesis $h_i$ we want to compute the cumulated score $score_{sem}(h_i)$. To perform this, for each hypothesis pair $(h_i, h_j)$ of the N-best list of this sentence:

- we apply the DNN model and obtain the output value $v_{ij}$ (between 0 and 1). A value $v_{ijs}$ close to 1 means that $h_i$ is better than $h_j$. We use this value to compute the scores for these hypotheses.
- we update the scores of both hypotheses as:
  $score_{sem}(h_i)$ += $v_{ij}$; $score_{sem}(h_j)$ += 1-$v_{ij}$

The obtained cumulated score $score_{sem}(h_i)$ is used as a *pseudo* probability $P_{sem}(h_i)$ and combined with the acoustic and linguistic likelihoods with a proper weighting factor (to be optimized) according to eq. (2). In the end, the hypothesis that obtains the best score is chosen as the recognized sentence.

Our two proposed DNN-based rescoring models producing $P_{sem}(h_i)$ are based on *BERT*, which is a multi-layer bidirectional transformer encoder that achieves state-of-the-art performances for various natural language tasks (NLP). The pre-trained *BERT* model can be fine-tuned using task-specific data [22]. Since the cosine distance is not meaningful for *BERT* SMs [25][26], we compute the semantic information at the sentence level, as described below.

In our approach, we employ a pre-trained *BERT* model. Two methods can be used to fine-tune *BERT* using application-specific data: *masked LM* and *next sentence prediction*. We base our *BERT* fine-tuning on a task similar to the latter.

The first proposed model, $BERT_{sem,}$, consists of performing the fine-tuning of *BERT* (with a fully connected layer at the top of BERT) using only embeddings of CLS tokens (the first token of a sentence, used for sentence classification). During the training, we input a hypothesis pair ($h_i$, $h_j$), that we want to compare, and the output is set to 1 (or 0) if the first (or the second) hypothesis achieves the lowest WER.

The second proposed model, $BERT_{alsem}$, takes input as feature vectors which include *acoustic, linguistic,* and *textual information*. Figure 1 shows the architecture of this model: the text of the hypothesis pair is given to the *BERT* model. Then, the embedding token of *BERT*, representing this pair, is given to the tra, followed by max pooling and average pooling, and then by a fully connected layer (FC) with a ReLU (*Rectified Linear Unit*) activation function. Finally, the output of this FC is concatenated with the acoustic and linguistic information of the hypothesis pair and passed through the second FC layer followed by a sigmoid activation function (to obtain a value between 0 and 1). In the end, the output $v_{ij}$ is obtained. In this setting, we use a one-layer Bi-LSTM and two layers of FC. More complex DNN architectures can also be considered. Advantage of this model is that the weights of acoustic, linguistic, and semantic information are learned together to provide a more powerful model.

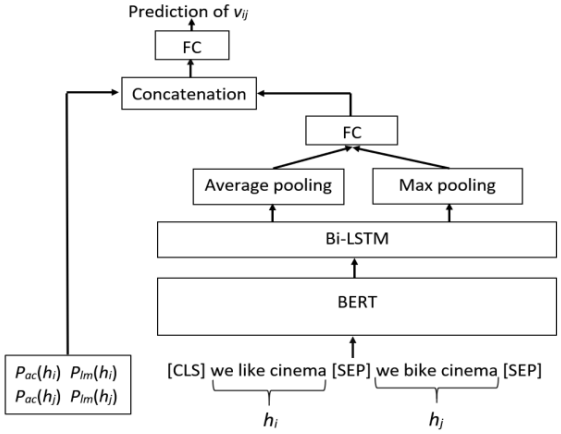


Figure 1: *Proposed $BERT_{alsem}$ rescoring model.*

## 3. Experimental conditions

### 3.1. Corpus description

For this study, we use the publicly available TED-LIUM corpus [4], which contains recordings from TED conferences. Each conference, within the corpus, is focused on a particular subject, so the data is well suited to our study. We use the train, development, and test partitions provided within the TED-LIUM corpus: 452 hours for training (268k segments), 8 conferences (507 segments) for development, and 11 conferences (1155 segments) for the test set (see Table 1). As usual, we apply the development set to choose the best parameter configuration, and the test set to evaluate the proposed methods with the best configuration. We compute the WER to measure the performance. Since our model only compares two hypotheses and cannot estimate the word probabilities, it is not possible to calculate the perplexity of our model. Therefore, in this article, we will not be providing any results related to perplexity.

This research work was carried out as part of an industrial project, studying the recognition of speech in noisy conditions, more precisely in fighter aircrafts. Then, we add noise to the development and test sets to get closer to the actual conditions

of an aircraft: noise added at 10 dB and 5dB SNR (noise of an F16 from the NOISEX-92 corpus [23]). The noise is *not added* to the training set. Furthermore, we evaluate the proposed approaches in clean conditions (training and testing).

Table 1: *The statistics of the TED-LIUM dataset*.

| ***Data*** | ***Nbr. of talks*** | ***Nbr. of words*** | ***Duration*** |
|---|---|---|---|
| Train | 2,351 | 4.8M | 452h |
| Development | 8 | 17,783 | 1h36 |
| Test | 11 | 27,500 | 2h37 |

### 3.2. Recognition system description

We use a recognition system based on the Kaldi voice recognition toolbox [15]. TDNN triphone acoustic models are trained on the training part (without noise) of TED-LIUM using sMBR training (*State-level Minimum Bayes Risk*). The lexicon and LM were provided in the TED-LIUM distribution. The lexicon contains 150k words. The LM has 2 million 4-grams and was estimated from a textual corpus of 250 million words. We also perform N-best list generation using the RNNLM model (LSTM) [11] [21]. Our objective for this is to verify if, by using a more powerful LM, the proposed rescoring models can improve the ASR. In all experiments, during N-best rescoring, the LM (4-gram or RNNLM) is not modified.

### 3.3. Rescoring models

According to our previous work on semantic models [9], the use of 5 or 10 hypotheses of the N-best list is not enough for efficient rescoring. Using more than 25 hypotheses shows no further improvement. In this study, we chose to use an N-best list of 20 hypotheses in all our experiments. Moreover, this size of N-best lists seems to be reasonable to generate the pairs of hypotheses and to have a tractable computational load during the training. In the case that a larger sized N-best list was to be required, a different pair comparison strategy would have been used [13]. During the training, the hypothesis pairs that get the same WER are not used. During evaluation (with development and test sets), all hypothesis pairs are considered, because we don't have the word error rate for these hypotheses.

For all experiments, combination weights are: $\alpha=1$, $\beta$ is between 8 and 10, and $\gamma$ is between 80 and 100. For each model, the weight values performing the best N-best rescoring performance for the development data were selected as the optimal value for the test data.

For $BERT_x$ models, we downloaded the pre-trained *BERT* model provided by Google with 110M parameters, 12 layers, and the size of the hidden layers at 768 [22]. For the $BERT_{sem}$ fine-tuning three epochs are performed. For $BERT_{alsem}$ we use *Adam* optimizer and binary cross-entropy loss function. We iterate the training for four epochs: during the first two epochs the weights of $BERT_{sem}$ are frozen, during the last two epochs all weights are updated. The dropout is 30 %. We train the proposed models using all training set.

As the *Generative Pre-Training Transformer 2* model (GPT-2) showed good performance in several NLP tasks [16], we applied this model in our experiments. The pre-trained GPT-2 LM was downloaded from the *Hugging Face* site. This model contains 117M parameters and was trained by *OpenAI* on 40GB of Internet text. This model is used in our experiments as the LM during the N-best rescoring (instead of n-gram). We also performed similar experiments using Masked Language Model (MLM) [17]. The results are less good than those of GPT-2 and are not presented here for lack of space.

## 4. Experimental results

We investigated the different hyperparameters of the proposed models. We can say that for the *BERT*-based rescoring models, it is important to use a large corpus of training (millions of pairs of hypotheses) and to choose a model with many hidden layers (we tested 4, 8, and 12 layers). For lack of space, we do not give these results in this article.

We report the WER for the development and the test sets of TED-LIUM with clean speech and in noise conditions of 10 and 5 dB. In Tables 2 and 3, the first line of results (method *Random*), corresponds to the random selection of the recognition result from the N-best hypotheses without the use of the proposed rescoring models. The second line of the Tables (method *Baseline*), corresponds to WER performance without using the rescoring models (standard ASR). The last line of the Tables (method *Oracle*) represents the maximum performance that can be obtained by searching in the N-best hypotheses: we select the hypothesis, which minimizes the WER for each sentence. The other lines of the Tables give the performance of the proposed approaches.

To fairly compare the proposed transformer-based models to other state-of-the-art transformer-based models introducing long-range context dependencies, we experiment with a rescoring based on the GPT-2 model. It corresponds to the rescoring of N-best hypotheses according to two configurations: (a) using eq. (1) with $P_{lm}(h)$ given by the GPT-2, instead of n-gram, while the SM is not used (*GPT2 comb. with ac. scores* in Tables); (b) using eq. (2) with $P_{sem}(h)$ given by the $BERT_{alsem}$ model and $P_{lm}(h)$ given by the GPT-2, instead of the n-gram ($BERT_{alsem}$ *comb. with ac./GPT-2 scores* in Tables).

For our rescoring models, we study three configurations:

- Rescoring using only the scores $score_{sem}(h)$ computed with *BERT*-based rescoring methods (denoted $BERT_x$ in Tables). In this case, in eq. (2) $\alpha=0$, $\beta=0$, *and* $\gamma=1$.
- Rescoring using a combination of the *BERT*-based $score_{sem}(h)$, and the acoustic score $P_{ac}(h)$ ($BERT_x$ *comb. with ac. scores* in Tables). In this case, $score_{sem}(h)$ is used as a *pseudo probability* and multiplied to the acoustic likelihood with a proper weighting factor $\gamma$. $P_{lm}(h_i)$ is not used in this combination, namely in eq. (2) $\beta=0$.
- Rescoring using a combination of the *BERT*-based score, the acoustic score $P_{ac}(h)$ and the linguistic score $P_{lm}(h)$ ($BERT_x$ *comb. with ac./x scores,* in Tables). For the most efficient $BERT_{alsem}$ model, we also use the GPT-2 score as a linguistic score to combine as described above.

From Table 2 we can observe that for all conditions and all evaluated rescoring models, the proposed rescoring models outperform the baseline system. This shows that the proposed Transformer-based rescoring models are efficient at capturing a significant proportion of the semantic information. Combining the acoustic score with the $BERT_{sem}$ model ($BERT_{sem}$ *comb. with ac. scores* in Tables) improves the performance. Indeed, the acoustic score is an important feature and should be taken into account. On the other hand, combining the linguistic score alone with the *BERT* rescoring gives no improvement compared to the *BERT* model. We do not present this result in the Tables. Google's BERT model, trained on billions of sentences, probably captures the linguistic structure of the language better than a simple n-gram LM trained on a much smaller corpus. Using the linguistic and acoustic scores with the *BERT* rescoring model ($BERT_{sem}$ *comb. with ac./4-gram scores*) brings small additional improvement. From these

Table 2: *ASR WER (%) on the TED-LIUM development and test sets, SNR of 10 and 5 dB, 20-best hypotheses,* ***4-gram LM****. "*" denotes significantly different result compared to "GPT-2 comb. with ac. scores" configuration.*

| ***Methods/systems*** | ***SNR 5 dB*** | | ***SNR 10 dB*** | | ***no added noise*** | |
|---|---|---|---|---|---|---|
| | ***Dev.*** | ***Test*** | ***Dev.*** | ***Test*** | ***Dev.*** | ***Test*** |
| Random system | 33.5 | 41.3 | 16.9 | 22.9 | 10.6 | 12.1 |
| Baseline system | 32.7 | 40.3 | 15.7 | 21.1 | 8.7 | 8.9 |
| GPT-2 comb. with ac. scores | 30.0 | 37.1 | 13.1 | 17.9 | 6.8 | 7.3 |
| *$BERT_{sem}$* | 31.1 | 38.7 | 14.4 | 19.8 | 8.0 | 8.7 |
| *$BERT_{sem}$ comb. with ac. scores* | 30.6 | 37.9 | 14.2 | 19.4 | 7.9 | 8.6 |
| *$BERT_{sem}$ comb. with ac./4-gram scores* | 30.6 | 37.9 | 14.1 | 19.4 | 7.8 | 8.5 |
| *$BERT_{alsem}$* | 30.4 | 37.5 | 13.5 | 18.6 | 6.9 | 7.3 |
| *$BERT_{alsem}$ comb. with ac./4-gram scores* | 30.2 | 36.9 | 13.4 | 18.3 | 6.8 | 7.0 |
| *$BERT_{alsem}$ comb. with ac./GPT-2 scores* | **29.7*** | **36.6*** | **12.8*** | **17.5*** | **6.4*** | **6.6*** |
| Oracle | 27.5 | 33.2 | 11.2 | 15.0 | 5.2 | 4.7 |

Table 3: *ASR WER (%) on the TED-LIUM development and test sets, SNR of 10 and 5 dB, 20-best hypotheses,* ***RNNLM (LSTM)****. "*" denotes significantly different result compared to "GPT-2 comb. with ac. scores" configuration.*

| ***Methods/systems*** | ***SNR 5 dB*** | | ***SNR 10 dB*** | | ***no added noise*** | |
|---|---|---|---|---|---|---|
| | ***Dev*** | ***Test*** | ***Dev*** | ***Test*** | ***Dev*** | ***Test*** |
| Random system | 29.2 | 38.4 | 13.9 | 20.2 | 8.9 | 10.8 |
| Baseline system | 28.2 | 37.1 | 12.3 | 17.7 | 6.6 | 7.2 |
| GPT-2 comb. with ac. scores | 26.2 | 34.9 | 11.0 | 15.9 | 6.1 | 6.7 |
| *$BERT_{sem}$* | 27.0 | 35.9 | 12.0 | 17.4 | 7.1 | 8.1 |
| *$BERT_{sem}$ comb. with ac. scores* | 26.6 | 35.3 | 11.6 | 17.1 | 6.9 | 7.1 |
| *$BERT_{sem}$ comb. with ac./RNNLM scores* | 26.5 | 35.4 | 11.5 | 16.9 | 6.0 | 6.6 |
| *$BERT_{alsem}$* | 26.1 | 35.2 | 11.0 | 16.4 | 5.9 | 6.6 |
| *$BERT_{alsem}$ comb. with ac./RNNLM scores* | 25.9* | 34.5* | 10.8* | 15.9 | 5.6* | 6.1* |
| *$BERT_{alsem}$ comb. with ac./GPT-2 scores* | **25.5*** | **34.4*** | **10.4*** | **15.4*** | **5.4*** | **5.7*** |
| Oracle | 23.1 | 30.2 | 8.3 | 12.1 | 3.8 | 3.5 |

results, for *$BERT_{alsem}$* we decided to combine the acoustic and linguistic scores. We observe that in all cases, *$BERT_{alsem}$*, trained with acoustic and linguistic scores, provides further large WER reductions compared to *$BERT_{sem.}$*. *$BERT_{alsem}$* combined with the GPT-2 model, a widely-used robust model, brings a *significant improvement* compared to the rescoring using *GPT-2 comb. with ac. scores* (denoted by *"*"* in Tables). This indicates that *$BERT_{alsem}$* can efficiently incorporate semantic and acoustic information, while remaining competitive, and bring complementary information compared to the GPT-2 model.

For *BERT*-based results, all improvements are *significant* compared to the *baseline* system (confidence interval at 5% significance level is computed according to the matched-pairs test [7]). On the test set, *$BERT_{alsem}$ comb. with ac./4-gram scores* achieves 8 % (for 5 dB), 13 % (for 10 dB), and 21 % (for clean) relative WER compared to the baseline system. Compared to the *GPT-2 comb. with ac. scores, $BERT_{alsem}$ comb. with ac./GPT-2 scores* allow us to obtain additional significant improvements.

As n-gram LM is limited in its ability to model long-range dependencies, we performed the ASR experiments using the more powerful RNNLM (LSTM). Table 3 reports the results for the same set of experiments, but, instead of n-gram, the RNNLM (LSTM) is used to generate the N-best hypotheses. The proposed rescoring *BERT*-based methods give consistent improvements compared to the n-gram LM results. All previous observations are valid for RNNLM-based experiments. The best system (*$BERT_{alsem}$ comb. with ac./RNNLM scores)* gives between 7 % and 14 % of relative improvement on the test set compared to the baseline system. These improvements are also significant. Compared to the 4-gram results, the improvements of RNNLM are smaller, but this is due to the fact that RNNLM can take into account more distant dependencies than n-gram.

## 5. Conclusions

In this article, we focus on the task of improving automatic speech recognition in clean and noisy conditions. Our methodology is based on taking into account semantics through powerful representations that capture the long-term relations of words and their contexts. The semantic information of the utterance is taken into account through a rescoring module on ASR N-best hypotheses. We proposed two effective DNN approaches based on the *BERT* model: one approach uses *BERT*-fine-tuning and represents a purely SM. The second approach uses a DNN and *BERT* models trained using semantic, acoustic, and linguistic information. On the corpus of TED-LIUM conferences, the system *$BERT_{alsem}$ with ac./LM scores* achieves between 7% and 21% of relative improvement compared to the baseline system. These improvements are statistically significant for all evaluated (clean and noisy) conditions and the two LMs: n-gram and RNNLM (LSTM). *$BERT_{alsem}$* remains competitive compared to GPT-2 rescoring, and the best performance is obtained by *$BERT_{alsem}$ with ac./GPT-2 scores.* Future work will include the introduction of an attention mechanism and context information beyond the utterance level [5].

## 6. Acknowledgements

The authors thank the DGA (*Direction Générale de l'Armement*, part of the French Ministry of Defence), Thales AVS and Dassault Aviation who are supporting the funding of this study and the "*Man-Machine Teaming*" scientific program in which this project is taking place.

# 7. References


[1] A. Bayer, G. Riccardi, "Semantic Language Models for Automatic Speech Recognition", *Proceedings of the IEEE Spoken Language Technology Workshop (SLT),* 2014.

[2] R. Corona, J. Thomason, R. Mooney, "Improving Black-box Speech Recognition using Semantic Parsing", *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pp.122–127, 2017.

[3] J. Devlin, M.-W. Chang and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of NAACL-HLT*, 2019.

[4] H. Fernandez, H. Nguyen, S. Ghannay, N. Tomashenko and Y. Esteve, "TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation", *Proceedings of SPECOM*, pp. 18–22, 2018.

[5] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, T. Kawahara, "Distilling the Knowledge of BERT for Sequence-to-Sequence ASR," *Proceedings of Interspeech*, 2020.

[6] J. Gaspers, P. Cimiano, B., "Semantic Parsing of Speech using Grammars Learned with Weak Supervision", *Proceedings of the HLT-NAACL*, pp. 872–881, 2015.

[7] L. Gillick and S. Cox S, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, v. 1, pp. 532–535, 1989.

[8] A. Kumar, C. Morales, M.-E. Vidal, C. Schmidt, S Auer," Use of Knowledge Graph in Rescoring the N-best List in Automatic Speech Recognition", *arXiv:1705.08018v1*, 2017.

[9] S. Level, I. Illina and D. Fohr, "Introduction of Semantic Model to Help Speech Recognition", *International Conference on Text, Speech and Dialogue,* 2020.

[10] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-robust Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[11] T. Mikolov, S. Kombrink, L. Burget, J.-H. Cernocky, S. Khudanpur, "Extensions of Recurrent Neural Network Language Model", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 5528–5531, 2011.

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", *Advances in Neural Information Processing Systems,* 26, pp. 3111-3119, 2013.

[13] A. Ogawa, M. Delcroix, S. Karita and T. Nakatani, "Rescoring N-best Speech Recognition List Based on One-on-One Hypothesis Comparison Using Encoder-Classifier Model," *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2018.

[14] A. Ogawa, M. Delcroix, S. Karita and T. Nakatani, "Improved Deep Duel Model for Rescoring N-best Speech Recognition List Using Backward LSTMLM and Ensemble Encoders", *Proceedings of Interspeech,* 2019.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, "The Kaldi Speech Recognition Toolkit", *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* 2011.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language Models are Unsupervised Multitask Learners", *Technical Report OpenAI* 2019.

[17] J. Salazar, D. Liang, T. Q. Nguyen, K. Kirchhoff, "Masked Language Model Scoring," *Proceedings of ACL,* 2020.

[18] I. Sheikh, D. Fohr, I. Illina, G. Linares, "Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition", *IEEE/ACM Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers*, 25 (3), pp.598 – 610, 2017.

[19] J. Shin, Y. Lee, K. Yung, "Effective Sentence Scoring Method Using BERT for Speech Recognition", *Proceedings of ACML,* 2019.

[20] Y. Song, D. Jiang, X. Zhao, Q. Xu, R. Wong, L. Fan and Q. Yang. "L2RS: a Learning-to-rescore Mechanism for Automatic Speech Recognition", *arXiv:1910.11496*, 2019.

[21] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM Neural Networks for Language Modeling," *Proceedings of Interspeech*, 2012.

[22] I. Turc, M.-W. Chang, K. Lee and K. Toutanova, "Well-Read Students Learn Better: On the Importance of Pre-training Compact Models*", arXiv:1908.08962v2*, 2019.

[23] A.Varga and H. Steeneken, "Assessment for automatic speech recognition II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems", *Speech Communication*, Volume 12, Issue 3, pp. 247-251, 1993.

[24] A. Wang and K. Cho, "BERT has a Mouth, and it Must Speak: BERT as a Markov Random Field Language Model", *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation,* pages 30–36, 2019.

[25] https://github.com/hanxiao/bert-as-service/, "Bert as service", visited on June 2021.

[26] https://github.com/hanxiao/bert-as-service#q-thecosine-similarity-of-two-sentence-vectors-is-unreasonably-high-eg-always--08-whats-wrong, "Bert as service", visited on June 2021.