Semantic Information Investigation for Transformer-based Rescoring of N-best Speech Recognition

Irina Illina, Dominique Fohr

Semantic Information Investigation for Transformer-based Rescoring of N-best Speech Recognition

Irina Illina, Dominique Fohr

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

Abstract

This article proposes to improve an automatic speech recognition system by rescoring N-best recognition lists with models that could enhance the semantic consistency of the hypotheses. We believe that in noisy parts of speech, the semantic model can help remove acoustic ambiguities. The estimate of a pairwise score for each pair of hypotheses is performed by *BERT* representations. The acoustic likelihood and LM scores are used as features in order to incorporate acoustic, language, and textual information together. In this research work, two new ideas are investigated: to use a fine-grained semantic representation at the word token level and to rely on the previously recognized sentences. On the TED-LIUM 3 dataset, in clean and noisy conditions, the best performance is obtained by leveraging context beyond the current utterance, which significantly outperforms the rescoring using the state-of-the-art GPT-2 model and the work of Fohr and Illina (2021).

Keywords: automatic speech recognition, semantic context, Transformer-based language models.

1. Introduction

Nowadays, automatic speech recognition systems (ASR) are widely used in everyday life. However, in the presence of noise, the degradation in performance can be detrimental to real applications (Deng *et al.*, 2014). In noisy conditions, the speech signal is less reliable and other knowledge is required to guide the recognition process. One possibility is to take into account the long-term context through a *semantic model*.

Semantic information is increasingly explored in recent works. Zhao *et al.* (2021) explore the denoising autoencoder for pretraining sequence-to-sequence semantic correction method and use transfer learning. Level *et al.* (2020) introduce the notions of a context part and possibility zones. Kumar *et al.* (2017) extract the semantic relations from the *DBpedia* (Auer *et al*., 2007) and uses them as features for rescoring.

An efficient solution to incorporate long-range semantic information can be through the *rescoring of the ASR N-best hypotheses list*. Ogawa *et al.* (2018, 2019) introduce N-best rescoring through a Long Short-Term Memory (LSTM) based encoder network. Liu *et al.* (2021) present a domain-aware rescoring framework for achieving domain adaptation during second-pass rescoring. A large range of textual information from different NLP models and a procedure to automatically estimate their weights are used by Song *et al.* (2021). A domain-aware rescoring framework to achieve domain adaptation during second-pass rescoring is proposed by Liu *et al.* (2021). In Xu *et al*. (2022), for second-pass rescoring the authors propose to train a *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin *et al*., 2019; Wang and Cho, 2019) on a discriminative objective such as minimum word error rate.

Some studies have attempted to include semantic information in ASR using a *context larger than the current sentence* to be recognized. Irie et *al.* (2019) train language models based on LSTM and transformers using long training sequences obtained by concatenation of sentences and study their robustness. Parthasarathy *et al.* (2019) focus on the ability of LSTM and transformer language models to learn context across sentence boundaries. Futami *et al.* (2020) exploit both left and right contexts of an utterance by applying *BERT* as an external language model through knowledge distillation. All these works show that it is relevant to rely on a broad context beyond sentence boundaries.

In previous works, Fohr and Illina, (2021) and Illina and Fohr, (2021) incorporated *sentence-level semantic information* (SI) into ASR. For this, the rescoring of the list of N-best hypotheses is carried out using distant contextual dependencies, which are important, especially for noisy conditions. In noisy parts of speech, the semantic model can help remove acoustic ambiguities. An efficient DNN architecture, based on *BERT*, and using semantic, acoustic and language model scores has been proposed. This model deals with pairs of N-best hypotheses to provide a pseudo-probability of the former being semantically more likely than the other. For example, in the following hypotheses for one sentence to recognize, taken from the TED-LIUM 3 corpus: "*hyp1: in antarctica we observe now a negative eyes balance"; "hyp2: in antarctica we observe now a negative ice balance",* the second hypothesis is more coherent semantically.

Compared to the work of Fohr and Illina (2021), the aim of the current paper is to extend this model. Two N-best rescoring approaches are proposed: the first one uses fine-grained information at the word token level; the second one relies on the previously recognized sentences. The combination of these two ideas is also studied. Compared with Ogawa *et al.* (2019), we use the *BERT* model that benefits from the pre-training on large corpora and not just on speech training corpus. Moreover, we exploit previous sentence information. In comparison to (Shin *et al.* (2019), where the *BERT* model computes word-level pseudo probabilities, we use the sentence prediction capability of the *BERT* model and the Generative Pre-Training Transformer (GPT-2) model (Radford *et al.*, 2019). Regarding Irie *et al.* (2019), where previous sentence information is used to improve the language model for the lattice rescoring, we integrate the information from the previous sentence into the *BERT*-based pairwise model for N-best re-ranking.

Our proposed approach using the previous sentence significantly outperforms the state-of-the-art GPT-2 rescoring and the rescoring model of Fohr and Illina (2021). This research work was carried out as part of an industrial project.

2. Proposed methodology

2.1 DNN based rescoring model

Methods, proposed in this article, are based on the methodology presented in (Fohr and Illina, 2021) where it is proposed to take into account the SI by rescoring the best hypotheses list of the ASR system. In this section, we give a brief overview of this methodology.

In this approach, to improve the ASR system, semantic model is introduced and combined with the acoustic probability $P_{ac}(h_i)$, the language model probability $P_{lm}(h_i)$, and the semantic score $P_{sem}(h_i)$, using specific weights α, β and γ:

$$\widehat{W} = argmax_{h_i \epsilon H} \; P_{ac}(h_i)^{\alpha} * P_{lm}(h_i)^{\beta} * P_{sem}(h_i)^{\gamma} \quad (1)$$

where h_i is a hypothesis from the N-best list H. The goal is to estimate the semantic score $P_{sem}(h_i)$ using a DNN-based model.

The rescoring is based on a comparison of ASR hypotheses, two per two to obtain a tractable size of the rescoring DNN input vectors. DNN rescoring model (denoted $BERT_{alsem}$) computes SI, associated with each pair of hypotheses.

For each hypothesis pair (h_i, h_j), during the training 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 computation of $P_{sem}(h_i)$ is done as follows. For each hypothesis h_i of a given sentence, the cumulated score $score_{sem}(h_i)$ is evaluated. For this, for each pair of hypotheses (h_i, h_j) of the N-best list of this sentence: (a) the output value v_{ij} (between 0 and 1) is obtained by DNN model, which relies on the *BERT* model. A value v_{ij} close to *1* means that h_i is better than h_j. This value is used to compute the scores for these hypotheses; (b) the scores of both hypotheses are updated:

$$score_{sem}(h_i) \mathrel{+}= v_{ij}; \qquad score_{sem}(h_j) \mathrel{+}= 1\text{-}v_{ij}$$

We normalise the cumulated score $score_{sem}(h_i)$ by dividing by *N-1* and use it as *pseudo probability* $P_{sem}(h_i)$. The obtained value is combined with the acoustic and language model likelihoods (see eq. (1)). Finally, the hypothesis with the best score is chosen as the recognized sentence.

2.2 $BERT_{alsem}$ rescoring model

In this section, we recall the architecture of the $BERT_{alsem}$ model from (Fohr and Illina, 2021), used as the starting point for the current work. *alsem* denotes "Acoustic, Linguistic and SEMantic" information, because we use *acoustic* and *textual information*. The advantage of this model is that the relative importance of acoustic, language model, and SI is learned together to provide a powerful model.

In Figure 1 (without the dotted block), the text of the pair of hypotheses (h_i and h_j) is given to the *BERT* model. The outputs of *BERT* are given to a bi-LSTM layer, max pooling, average pooling, and then to a fully connected (FC) layer with a ReLU (*Rectified Linear Unit*) activation function (Nair and Hinton, 2010). The output of this FC layer, the acoustic probabilities, and language model probabilities are concatenated. The final FC layer (followed by a sigmoid activation function) computes output v_{ij}.

2.3 Fine grained rescoring model $BERT_{alsem\text{-}fg}$

This section presents the first rescoring method proposed in this paper. The objective is to provide $BERT_{alsem}$ model with fine-grained information (at the word token level and not just at the sentence level as in $BERT_{alsem}$). We would like to integrate the probability of each word token of a given hypothesis. This value represents the probability of a token given *all previous tokens* of the hypothesis. For a pair of hypotheses, two vectors of token probabilities are generated, one for each hypothesis (see Figure 1, dotted part). Each vector is assigned as input to a neural network layer. Since such a vector is a temporal sequence, bi-LSTM or CNN are the best suited to process this type of information and to obtain a fixed length vector. The outputs of these two layers (one for each hypothesis) are concatenated with the acoustic and language model scores, and SI of the hypothesis pair is calculated by *BERT*. This concatenation is passed through an FC layer followed by a sigmoid activation function. Finally, the output v_{ij} of this network is obtained. We call this model *fine-grained* $BERT_{alsem\text{-}fg}$.

To estimate the probability of each word token of a given hypothesis, GPT-2 is used. The first advantage of using GPT-2 is its attention mechanisms allowing the model to selectively focus on the most relevant word tokens. The second potential advantage is to provide complementary information compared to the *BERT* model included in $BERT_{alsem}$.

2.4 Rescoring using previous sentences $P\text{-}BERT_{alsem}$

This part focuses on the second proposed method taking into account the ASR output of the previously recognized sentences for improving the recognition of the current sentence. We would like to combine the SI of one or more previous sentences with the SI of the current sentence. Indeed, the SI contained in the previous sentences and in the current sentence of a discourse are related (Irie *et al.*, 2019). This relationship can link some words from the previous sentences with words from the current sentence. Our objective is to take into account these semantic relations to select the best hypothesis.

The proposed rescoring model using the previously recognized sentences is denoted $P\text{-}BERT_{alsem}$. Compared to the $BERT_{alsem}$, we added the words from the previously recognized sentences to *each hypothesis of a hypothesis pair*. This information is given to the *BERT* model. Concerning the acoustic and language model information part of $BERT_{alsem}$, we modify the language model probabilities by replacing them with the conditional probabilities $P_{lm}(h_i \mid prev\_sent)$ and $P_{lm}(h_j \mid prev\_sent)$. The acoustic probabilities are unchanged.

2.5 Combined rescoring model $P\text{-}BERT_{alsem\text{-}fg}$

The two proposed approaches perhaps contain complementary information and can be combined into a single model, denoted $P\text{-}BERT_{alsem\text{-}fg}$. In this model, for a given pair of hypotheses, the model input is composed of $P_{ac}(h_i)$, $P_{ac}(h_j)$, $P_{lm}(h_i|prev\_sent)$, $P_{lm}(h_j|prev\_sent)$, text of each hypothesis preceded by the text of the previously recognized sentences. The rest of the methodology is unchanged.

3. Experimental conditions

3.1 Corpus description

We use the publicly available TED-LIUM 3 corpus (Fernandez *et al.*, 2018), containing recordings from TED conferences. Each conference of this corpus focuses on a particular subject; thereby the data are well suited to our study. The train, development, and test partitions are provided within the TED-LIUM 3 corpus: (a) train: 2,351 talks, 4.8M words, 452h; (b) development: 8 talks, 17,783 words, 1h36; (c) test: 11 talks, 27,500 words, 2h37. We use the development set to choose the best parameter configuration, and the test set to evaluate the proposed methods with the best configuration.

In this paper, the study of the ASR in noisy conditions was performed because this work is a part of an industrial project (noisy ASR, more precisely in fighter aircrafts). We add noise to the train, development and test sets to get closer to the actual conditions of an aircraft. For the train part, we add different noises from NOISEX-92 corpus (Varga and Steeneken, 1993)

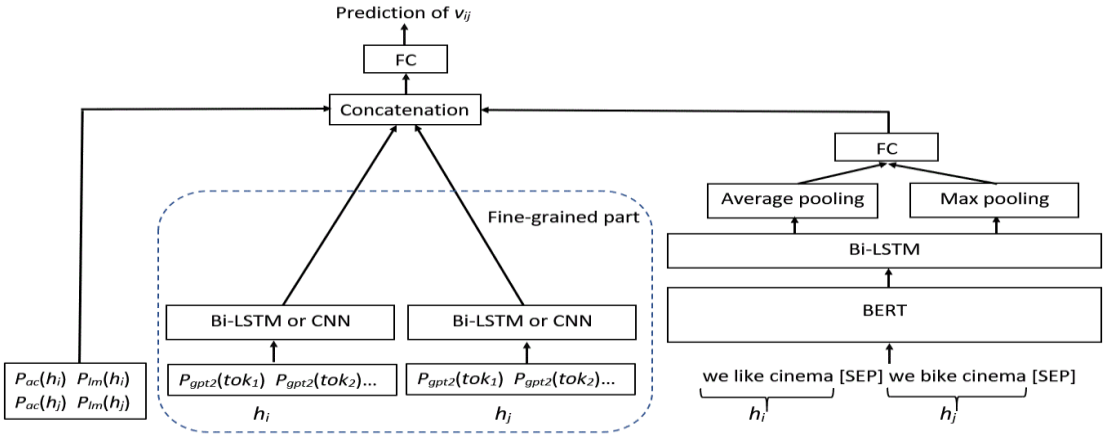

Fig. 1: Architecture of the proposed $BERT_{alsem\text{-}fg}$ rescoring model

(excluding F16 noise, used for development and test data) at SNR from 0 to 20 dB. We keep the size of the training set unchanged (no training data augmentation). Furthermore, we evaluate the proposed approaches in clean conditions.

3.2 Speech recognition system

Our recognition system is based on the *Kaldi speech recognition toolbox* (Povey *et al.*, 2011). Time Delay Neural Network (TDNN) (Waibel *et al*., 1989) triphone acoustic models are trained on the training part of TED-LIUM 3 using sMBR training (*State-level Minimum Bayes Risk*) (Kingsbury, 2009). The lexicon and language models were provided in the TED-LIUM 3 distribution. The lexicon contains 150k words. The LM used for the lattice generation has 2 million 4-grams and was estimated from a textual corpus of 250 million words. We perform N-best list generation with a more powerful LM: the RNNLM model (LSTM) (Sundermeyer *et al.*, 2012). Since this DNN model only compares two hypotheses and cannot *output* the word probabilities, it is not possible to calculate the perplexity of this model. We compute the word error rate (WER) to measure the performance.

It is worth noting that in (Fohr and Illina, 2021) the acoustic model was trained only on clean speech. In the current work, we carry out training on noisy data. The obtained model is more accurate for noisy ASR and the WERs are lower than in (Fohr and Illina, 2021).

3.3 Rescoring models

We have chosen to use an N-best list of 20 hypotheses in all experiments (Illina and Fohr, 2021). During the training of the proposed models, we do not use the hypothesis pairs which obtain the same WER. When evaluating (development and testing), we consider all hypothesis pairs, because the WERs are not available for these hypotheses.

For each model, the combination weight values α, β, and γ achieving the best rescoring performance on the development set are selected as the optimal value for the test data. For all the experiments, optimal values of the combination weights are: α=1, β is between 8 and 10, and γ is between 80 and 100. This large difference between the values is explained by the fact that we use likelihood or pseudo probabilities that are not normalized.

For our semantic models, we downloaded Google's pre-trained *BERT* model (110M parameters, 12 layers, and the size of the hidden layers is 768) (Turc *et al.*, 2019). We use the Adam optimizer (Kingma and Ba, 2015) and binary cross-entropy loss function.

We iterate the training of $BERT_{alsem}$ as follows: during the first epoch, the layer weights of the *BERT* model are frozen, and during the following epochs all *BERT* weights are updated. The dropout is 30 %. Two methods could be employed to use *BERT* with application-specific data: *masked LM* and *next-sentence prediction*. We use next-sentence prediction because we put two hypotheses as input to the *BERT* model (see figure 1, right part).

We downloaded pre-trained GPT-2 LM from the Hugging Face site. The model has 117M parameters and was trained by OpenAI on 40GB of Internet text. In our experiments, this model is used for several purposes: (a) as a language model $P_{lm}(h_i)$ during N-best rescoring (see eq. (1)); (b) inside $BERT_{alsem}$ to represent the language model score $P_{lm}(h_i)$ of each hypothesis (see figure 1); (c) inside $BERT_{alsem\_fg}$ to compute the score of each word token $P_{gpt2}(tok)$ of each hypothesis; (d) inside $P\text{-}BERT_{alsem}$ to compute $P_{lm}(h|prev\_sent)$. In all configurations, GPT-2 is fine-tuned on the transcriptions (references) of the train part of TED-LIUM 3.

In our preliminary experiments, during N-best rescoring, a Masked Language Model (MLM) (Salazar *et al.*, 2020) performed worse than GPT-2 and therefore the results will be not presented here.

4. Experimental results

We report the WER for the development and test sets of TED-LIUM 3 in clean speech and under noisy conditions (noise added at 10 and 5 dB). We recall that the acoustic model is trained on noisy speech. In Table 1, different notations are introduced: (a) ***Random*** represents the random selection of the recognition result from the N-best hypotheses (without using a rescoring model); (b) ***Baseline*** corresponds to the standard speech recognition system (without using a rescoring model); (c) ***Oracle*** gives the *maximum performance* that can be obtained by selecting in the N-best hypotheses: the hypothesis which minimizes the WER for each sentence is chosen; (d) ***GPT-2 resc***. corresponds to a state-of-the-art rescoring based on the fine-tuned GPT-2 model. We perform this rescoring to fairly compare the proposed transformer-based models to a state-of-the-art transformer-based model introducing long-range context dependencies. We rescore N-best hypotheses using eq. (1), where P_{lm} *(h)* is computed by the GPT-2. The semantic model is not used (γ=0); (e) ***$BERT_{alsem}$ with GPT-2 resc*** (Fohr and Illina, 2021) is performed to compare the

	Methods/systems	SNR 5 dB		SNR 10 dB		no added noise	
		Dev	Test	Dev	Test	Dev	Test
1	Random system	15.1	19.4	10.8	13.9	9.2	11.0
2	Baseline system	13.6	17.1	8.6	10.9	6.9	7.4
3	Baseline system with GPT-2 resc.	11.6	14.6	7.3	8.9	5.8	6.0
4	$BERT_{alsem}$ with GPT-2 resc. (Fohr and Illina, 2021)	11.4	14.5	7.1	8.9	5.6	5.9
5	$BERT_{alsem\text{-}fg}$ (CNN) with GPT-2 resc.	11.5	14.5	7.1*	8.8*	5.6*	5.9
6	$BERT_{alsem\text{-}fg}$ (bi-LSTM) with GPT-2 resc.	11.4*	14.5	7.1*	8.8*	5.6*	5.9
7	*P-*$BERT_{alsem}$ with GPT-2 resc, 1sent	11.2*~	14.3*	6.9*~	8.5*~	**5.3***~	**5.7***~
8	*P-*$BERT_{alsem}$ with GPT-2 resc,1sent30w, stop wrds remov.	**11.1 *~**	**14.2*~**	6.9*~	**8.4***~	**5.3***~	**5.7***~
9	*P-*$BERT_{alsem}$ with GPT-2 resc, 2sen30w, stop wrds remov.	11.2 ***~**	**14.2*~**	**6.8*~**	8.5***~**	**5.3*~**	**5.7*~**
10	Oracle	9.5	11.3	5.4	6.1	4.0	3.6

Table 1. ASR WER (%) on the TED-LIUM 3 development and test sets, SNR of 10 and 5 dB, 20-best hypotheses. "*" denotes significantly different result compared to GPT-2 resc. configuration (line 3). "~" denotes significantly different result compared to $BERT_{alsem}$ with GPT-2 resc. configuration (line 4)

transformer-based models, proposed in this paper, with $BERT_{alsem}$ proposed by Fohr and Illina (2021). It corresponds to the rescoring of the N-best hypotheses using eq. (1) with $P_{sem}(h)$ given by $BERT_{alsem}$ and $P_{lm}(h)$ given by the GPT-2. The other lines of Table 1 give the performance of the proposed approaches. The best results are presented in bold.

For the rescoring models proposed in this article, we studied three configurations: (a) ***$BERT_{alsem\text{-}fg}$ with GPT-2*** represents rescoring using $BERT_{alsem\text{-}fg}$. Configurations with CNN and bi-LSTM models are shown; (b) ***P-$BERT_{alsem}$ with GPT-2*** gives the results for the approach taking into account the previous sentence; (c) ***P-$BERT_{alsem\text{-}fg}$ with GPT-2:*** the combined model gives no additional improvement compared to ***P-$BERT_{alsem}$ with GPT-2*** and the results are not presented in this paper.

For *P-*$BERT_{alsem}$ rescoring model, to avoid the overflow of the number of the *BERT* input tokens, we use at most *M* last words from the previous sentence (*M=30*). Nevertheless, to compute $P_{lm}(h|prev\_sent)$ with the GPT-2, the whole previous sentences are used.

To analyse the results, we make the comparisons with: (a) the state-of-the-art rescoring model with GPT-2 (line 3); (b) the best configuration of $BERT_{alsem}$ rescoring model (line 4, (Fohr and Illina, 2021)).

The *significance of the results* is indicated by "*" in Table 1 compared to line 3, and by "~" compared to line 4. The confidence interval at the 5% significance level is calculated using the matched pairs test (Gillick and Cox, 1989), considering the effects of two different treatments (algorithms) on equivalent subjects (speech segments) aligned by a dynamic programming algorithm.

$BERT_{alsem}$ rescoring model. By studying the results of $BERT_{alsem}$ (line 4), we see that the conclusions given by Fohr and Illina (2021) are still valid when the noisy acoustic model is used: the $BERT_{alsem}$ provides consistent WER reduction compared to the baseline model with GPT-2 rescoring (line 3).

Fine-grained rescoring model: ***$BERT_{alsem\text{-}fg}$*** The $BERT_{alsem\text{-}fg}$ shows an improvement over the baseline system with GPT-2 rescoring (line 6 versus line 3, the significance is indicated by "*" in Table 1). The CNN architecture (line 5) shows similar results to the bi-LSTM one (line 6).

The proposed $BERT_{alsem\text{-}fg}$ displays a similar performance as $BERT_{alsem}$ (lines 5, 6 versus line 4). This means that probably adding fine-grained information (GPT-2 probabilities at the word token level) does not bring complementary information compared to the $BERT_{alsem}$ model. It is difficult to predict whether pre-trained GPT-2 and *BERT* models contain complementary information because these two models are learned on different corpora but are based on the same principle (Transformers).

Rescoring model using previous sentences: ***P-$BERT_{alsem}$*** The lines 8 and 9 display the results for our *P-*$BERT_{alsem}$ model. We use one or two previous sentences. Two configurations were studied: using M words of previous sentences or using only non-stop M words of the previous sentences (stop words contain little semantic information and were removed). The results for the second case are slightly better, therefore we present only the results for the second case.

Statistically significant improvements are observed for all noise levels and clean speech for the *P-*$BERT_{alsem}$ compared to the *GPT-2 resc* configuration (lines 8 and 9 versus line 3, the significance is indicated by "*"). Comparing *P-*$BERT_{alsem}$ with $BERT_{alsem}$ model (Fohr and Illina, 2021) (without the previous sentence information, lines 8 and 9 versus line 4), we see that the integration of the previous sentence information helps in the selection of the best hypothesis. This improvement is significant in almost all configurations (the significance is indicated by "~" in Table 1).

Analysing the results of *P-*$BERT_{alsem}$, we observe that the model corrects syntactic and semantic errors, compared to $BERT_{alsem}$ (lines 8 and 9 versus line 4). Here is one example of a semantic error corrected by *P-*$BERT_{alsem}$ model for 5dB noisy condition, test set:

***ref**: I got to **lhasa** that i understood the face behind the statistics you hear about six thousand sacred monuments…*

***hyp1**: I got to **loss** that i understood the face behind these statistics you hear about six thousand sacred monuments…*

***hyp2**: I got to **lhasa** that i understood the face behind these statistics you hear about six thousand sacred monuments…*

The second hypothesis is selected as the sentence recognized by *P-*$BERT_{alsem}$ because the previous sentence contains the word "Tibet".

Using one or two previous sentences (lines 8 and 9) gives similar results. We performed the experiments using three previous sentences and obtained no improvement. The results are not given here.

In conclusion, the best system *P-*$BERT_{alsem}$ gives between 1% and 3% relative WER reduction compared to $BERT_{alsem}$ rescoring model (Fohr and Illina, 2021) (lines 8, 9 versus line 4). These improvements are *statistically significant* according to the matched pairs test (Gillick and Cox, 1989) (see "~" in Table 1).

5. Conclusion

The aim of this article is to improve ASR in clean and noisy environments. In the framework of the pairwise rescoring of ASR N-best hypotheses, we would like to enrich the rescoring

model $BERT_{alsem}$. We have introduced two rescoring approaches, based on semantic representations. The first one is designed to integrate fine-grained information at the word token level. The second one exploits the context beyond the current utterance by considering the previously recognized sentences. The proposed models are based on DNN, *BERT*, and GPT-2 models. Experimental evaluation, carried out on TED-LIUM 3 corpus with clean and noisy speech, showed that the approach using one previous sentence gives a statistically significant improvement, outperforming the state-of-the-art rescoring using the advanced GPT-2 model and previous work of Fohr and Illina (2021) in almost all configurations.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The Semantic Web Lecture Notes in Computer Science*, Volume 4825/2007, pp. 722-735.

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

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

Fohr, D. and Illina, I. (2021). BERT-based Semantic Model for Rescoring N-best Speech Recognition List. *Proceedings of Interspeech*.

Futami, H. Inaguma, H., Ueno, S., Mimura, M., Sakai, S. and Kawahara, T. (2020). Distilling the Knowledge of BERT for Sequence-to-Sequence ASR. *Proceedings of Interspeech*.

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

Illina, I. and Fohr, D. (2021). DNN-based semantic rescoring models for speech recognition. *Proceedings of the Internat. Conference on Text, Speech and Dialogue*, TSD.

Irie, K., Zeyer, A., Schluter, R. and Ney, H. (2019). Training Language Models for Long-span Cross-Sentence Evaluation. *Proceedings of IEEE Automatic Speech Recognition & Understanding*, ASRU.

Kingma, P. D. and Ba, J. (2015). A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural-network acoustic modelling. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 3761–3764.

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

Level, S., Illina, I. and Fohr, D. (2020) Introduction of Semantic Model to Help Speech Recognition, *Proceedings of the International Conference on Text, Speech and Dialogue*, TSD.

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

Liu, L., Gu, Y., Gourav, A., Gandhe, A., Kalmane, S., Filimonov, D., Rastrow, A. and Bulyko, I. (2021). Domain-Aware Neural Language Models for Speech Recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *ICML*.

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

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

Parthasarathy, S., Gale, W., Chen, X., Polovets, G. and Chang, S. (2019). Long-span Language Modeling for Speech Recognition. *CoRR abs/1911.04571*.

Radford, A., Wu, J., Child, R., Luan, Amodei, D. and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *Technical Report OpenAI*.

Salazar, J., Liang, D., Nguyen, T. and Kirchhoff, K. (2020). Masked Language Model Scoring. *Proceedings of ACL*.

Shin, J., Lee, Y. and Yung, K. (2019). Effective Sentence Scoring Method Using BERT for Speech Recognition. *Proceedings of ACML*.

Song, Y., Jiang, D., Zhao, X., Xu, Q., Wong, R., Fan, L. and Yang, Q. (2021). L2RS: a Learning-to-rescore Mechanism for Automatic Speech Recognition. *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1157–1166.

Sundermeyer, M., Schluter, R. and Ney, H. (2012). LSTM Neural Networks for Language Modeling. *Proceedings of Interspeech*.

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

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

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339.

Wang, A. and Cho, K. (2019). 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.

Xu, L., Gu, Y., Kolehmainen, J., Khan, H., Gandhe, A. Rastrow, A. Stolcke, A., Bulyko, I. (2022). RescoreBERT: Discriminative Speech Recognition Rescoring with BERT. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP.

Zhao, Y., Yang, X., Wang, J., Gao, Y., Yan, C. and Zhou, Y. (2021). BART based Semantic Correction for Mandarin Automatic Speech Recognition System. *Proceedings of Interspeech*.