BERT Semantic Context Model for Efficient Speech Recognition

Irina Illina, Dominique Fohr

Lorraine University, CNRS, Inria, LORIA, F-54000 Nancy, France

illina@loria.fr, dominique.fohr@loria.fr

INTRODUCTION

An automatic speech recognition system (ASR) contains three main parts: an acoustic model, a lexicon and a language model. ASR in **noisy environments** is still a challenging goal. Indeed, when training and/or testing conditions are corrupted by noisy environments, the audio signal is distorted, and the acoustic model may not be able to compensate for this variability. A better language model gives limited performance improvement, modelling mainly local syntactic information. In this paper, we propose **a semantic model** to take into account the long-term semantic context information and thus to reduce the acoustic ambiguities of noisy conditions. Proposed semantic model is based on deep neural networks (DNN) and is aimed at selecting hypotheses that have a better semantic consistency and therefore a lower word error rate (WER). This model is applied during the **rescoring of N-best recognition hypotheses**.

As part of input features to our DNN semantic model, we investigate a powerful representation: dynamic contextual embeddings from **Transformer-based BERT** model [Devlin2018]. BERT takes into account the semantic context of words and has been shown to be effective for several natural language processing tasks. The efficiency and the semantic properties of BERT representations motivate us to explore them for our task. Thus, our ASR is supplemented by a semantic context analysis module. This semantic module re-evaluates (rescoring) the N-best transcription hypotheses and can be seen as a form of **dynamic adaptation in the specific context of noisy data.** Compared to [Ogawa, 2019], we use a more powerful model with several transformer layers.

PROPOSED METHODOLOGY

Introduction

An effective way to take into account semantic information is to re-evaluate (rescore) the N-best hypotheses of the ASR. For each sentence to be recognized, the ASR system can give multiple hypotheses (N-best list), ranked according to the cumulated acoustic and linguistic scores. More precisely, the recognition system provides for each hypothesis h of the sentence to be recognized an acoustic score $P_{acc}(h)$ and a linguistic score $P_{ml}(h)$. The best sentence is the one that maximizes the probability:

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

$\widehat{W}$ is the recognized sentence; H is the set of sentence hypotheses; h_i is the i-th sentence hypothesis; α and β represent the weights of the acoustic and language models.

We propose to add semantic information to guide the recognition process. For this, we modify the computation of the best sentence presented in (1) as follows:

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

We supplemented the formula (1) by the **semantic probability** of the hypothesis *h*, $P_{sem}(h)$, weighted by γ. We propose to estimate this semantic probability by a DNN-based model.

We propose to go beyond a simple score combination, like in eq. (2). We design a DNN-based rescoring model estimating $P_{sem}(h)$ as follow*:* the model takes *acoustic, linguistic*, and *textual information* as input. We believe that the acoustic and linguistic information should be trained together with the semantic information to give an accurate rescoring model.

N-best Rescoring Procedure

To keep a tractable size of the input vectors of the rescoring DNN, the rescoring is based on the comparison of ASR hypotheses, two per two. Then, our proposed DNN model uses 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) 0 otherwise.
The 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 $score_{sem}(h_i)$. 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 the 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.

To compute the cumulated score $score_{sem}(h_i)$, for each hypothesis pair (h_i, h_j) of the N-best list of this sentence:

- we apply the DNN semantic model and obtain the output value v_{ij} (between 0 and 1). A value v_{ij} close to 1 means that h_i is better than h_j.
- we update the scores of both hypotheses as:

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

DNN-based Semantic Model

The proposed model takes input as feature vectors which include *acoustic* (likelihood given by the acoustic model), *linguistic* (probability given by the language model), and *textual information* (text of the hypotheses). We hypothesize that training all this information together is better than combining the probabilities obtained by different models.

The text of the hypothesis pair is given to the *BERT* model. Then, the token embeddings of *BERT*, representing this pair, are given to a bi-LSTM layer, 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). Finally, the output v_{ij} is obtained. Figure 1 presents the architecture of the proposed rescoring model.

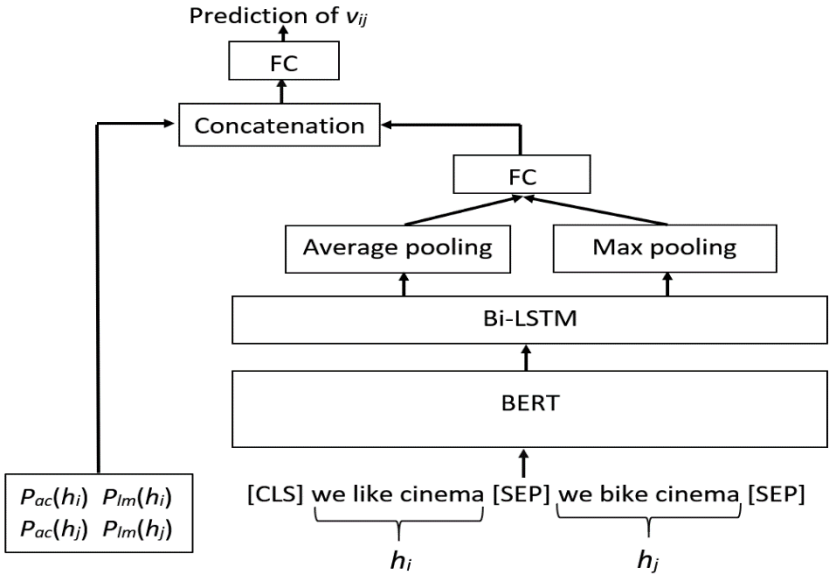

Figure 1: Architecture of the proposed rescoring model.

EXPERIMENTAL CONDITIONS

Corpus Description

For this study, we use the publicly available TED-LIUM corpus [Fernandez, 2018], which contains recordings from TED conferences. Each conference of the corpus is focused on a particular subject. We chose this corpus because it contains enough data to train a reliable acoustic model. We use the train, development, and test partitions provided within the TED-LIUM corpus: 2351 conferences for training (452 hours), 8 conferences (1h36) for development, and 11 conferences (2h27) for the test set. We use 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.

This research work was carried out as part of an industrial project, studying the recognition of speech in noisy conditions, more precisely in fighter aircraft. As it is very difficult to have an access to real data recorded in a fighter aircraft, 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 (excluding F16 noise, used for development and test data) corpus at SNR from 0 to 20 dB. For the development and test sets, the noise is added at 10 dB and 5dB SNR (noise of an F16 from the NOISEX-92 corpus [Varga, 1993]). We evaluate the proposed approaches in clean and noisy conditions. Compared to our previous work [Illina, 2021], we train the acoustic models using noisy data.

Recognition System Description

We use a recognition system based on the Kaldi voice recognition toolbox [Povey, 2011]. 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 perform N-best list generation using a RNNLM model (LSTM) [Sundermeyer, 2012].

For all experiments, combination weights are: $\alpha=1$, β is between 8 and 10, and γ 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.

EXPERIMENTAL RESULTS

We report the WER for the test sets of TED-LIUM with clean speech and in noise conditions at 10 and 5 dB SNR. In Table 1, the first line of results (*Random* method), corresponds to the random selection of the recognition result from the N-best hypotheses without the use of the proposed rescoring model. The second line of the Table (*Baseline* method), corresponds to WER performance without using the rescoring model (standard ASR). The last line of the Table (*Oracle* method) 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 table give the performance of the proposed approach, called $BERT_{alsem}$. For this model, rescoring is performed using a combination of the *BERT*-based score, the acoustic score P_{ac} *(h)* and the linguistic score P_{lm} *(h)* following eq. (2) ($BERT_{alsem}$ *comb. with ac./x scores,* in Table 1).
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. For the $BERT_{alsem}$ model, we also use the Generative Pre-trained Transformer Model (GPT-2) score as a linguistic score to combine as described above.
From Table 1 we can observe that for all conditions and all evaluated rescoring models, the proposed rescoring model outperforms the baseline system. This shows that the proposed Transformer-based rescoring model is efficient at capturing a significant proportion of the semantic information. For *BERT*-based results, all improvements are *significant* compared to the *baseline* system. Compared to the *GPT-2 comb. with ac. scores,* the proposed BERT-based semantic model allows us to obtain additional significant improvements. Confidence interval at 5% significance level is computed according to the matched-pairs test [Gillick, 1998].

Methods/systems	***5 dB***	***10 dB***	***No added noise***
Random system	19.4	13.9	11.0
Baseline system	17.1	10.9	7.4
GPT-2 comb. with ac. scores	15.3	9.6	6.7
$BERT_{alsem}$ *comb. with ac./RNNLM scores*	15.1	9.7	6.5
$BERT_{alsem}$ *comb. with ac./GPT-2 scores*	**14.7***	**9.0***	**6.0***
Oracle	11.3	6.1	3.6

Table 1: ASR WER (%) on the TED-LIUM 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.

CONCLUSION

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. The proposed approach uses BERT-based DNN models trained using semantic, acoustic, and linguistic information. On the corpus of TED-LIUM conferences, the proposed system achieves statistically significant improvement for all evaluated (clean and noisy) conditions compared to the baseline system. The proposed approach is competitive compared to GPT-2 rescoring.

Acknowledgement

The authors would like to thank the DGA (*Direction Générale de l'Armement*), Thales AVS and *Dassault Aviation* which support the funding of this study and the scientific program "Man-Machine Teaming" in which this research project occurs.

REFERENCES

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

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

L. Gillick and S. Cox S (1998). 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.

I. Illina, D. Fohr (2021). DNN-based semantic rescoring models for speech recognition. *TSD 2021 - 24th International Conference on Text, Speech and Dialogue.*

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

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 (2011). The Kaldi Speech Recognition Toolkit, *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).*

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

A.Varga and H. Steeneken (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", *Speech Communication*, Volume 12, Issue 3, pp. 247-251.