MULTILINGUAL SECOND-PASS RESCORING FOR AUTOMATIC SPEECH RECOGNITION SYSTEMS

Neeraj Gaur[*]*, Tongzhou Chen*[*]*, Ehsan Variani, Parisa Haghani, Bhuvana Ramabhadran, Pedro J. Moreno*

Google Inc.

ABSTRACT

Second-pass rescoring is a well known technique to improve the performance of Automatic Speech Recognition (ASR) systems. Neural Oracle Search (NOS), which selects the most likely hypothesis from an N-best hypothesis list by integrating information from multiple sources, such as the input acoustic representations, N-best hypotheses, additional first-pass statistics, and unpaired textual information through an external language model, has shown success in rescoring for RNN-T first-pass models. Multilingual first-pass speech recognition models often outperform their monolingual counterparts when trained on related or low-resource languages. In this paper, we investigate the use of the NOS rescoring model on a first-pass multilingual model and show that similar to the first-pass model, the rescoring model can be made multilingual. Our first-pass multilingual model does not require a language-id and we make a realistic assumption that an estimate of the language-id would be available for second-pass rescoring. We conduct comprehensive experiments on two sets of languages, one consisting of related low-resource languages, and the other with a high-resource language added to the first set to analyze the performance of the multilingual NOS rescorer under different settings. Our experimental results show that, multilingual NOS can improve the first-pass multilingual model resulting in average word error rate reduction of 9.4% in the first case, and 8.4% in the second, and out-performing the monolingual counterparts in both cases.

Index Terms: speech recognition, multilingual, RNN-T, N-best rescoring

1. INTRODUCTION

Recent progress in automated speech recognition (ASR) has seen the increased success of end-to-end approaches yielding good performance on a variety of tasks [1, 2, 3]. Large, complex, language models (LMs) such as maximum-entropy models [4], long short-term memory (LSTM) [5] LMs and transformer LMs [6, 7] have been used in either the first-pass decoding step or in the second-pass as an N-best/lattice rescorer to further improve performance and address the data-hunger of E2E models.

Multilingual ASR modeling is another approach for promoting data and parameter sharing and allowing for better generalization and robustness, in particular in low-resource languages [8]. In the E2E context, multilingual modeling has been explored for both streaming [9, 10, 11] and non-streaming [12, 13] applications. These approaches have not only allowed combination of data-scarce and high-resource languages but also allowed seamless language-switching for multilingual users. These models have now become the de facto approach for multilingual ASR [14, 15, 16]. Prior work in training multilingual representations and end-to-end models have shown that the best performing models require conditioning on language information. However, recent work has shown [9] that alternate normalization approaches can do away with this conditioning allowing for easy expansion to additional languages.

Second-pass rescoring strategies are normally language dependent, typically using language information to select the appropriate language model(LM) for rescoring. Approaches that use a unified vocabulary have shown WER reductions over first-pass multilingual models [17]. End-to-End models such as Connectionist Temporal Classification (CTC) [18] and Listen, Attend and Spell (LAS) [13] models have also been used for rescoring with some success [19]. An external LM (single or multilingual) built from unpaired text can be integrated into these models in the first-pass using LM fusion [20, 21] methodologies, or in a rescoring pass. Deliberation networks [22] use a second-pass decoding step that incorporates an additional attention over the whole sequence generated by the first-pass decoder, including words preceding and succeeding the current time step in the first-pass sequence.

Motivated by ranking and rescoring algorithms [23, 24, 25], a Neural Search Algorithm also referred to as Neural Oracle Search (NOS)[1] was proposed in [26]. NOS is a general framework to find the *best* sequence from among a set of N-best hypotheses (sequences). In this work, we extend NOS to be a multilingual model and demonstrate its effectiveness over multiple single language NOS models in the context of

[*] The first two authors have equal contributions. The rest of the list is sorted alphabetically.

[1] The algorithm directly predicts the index of the oracle path

multilingual ASR.

We show:

1. Lang-specific Neural Oracle Search models, give gains on top of a multilingual first-pass E2E streaming model with relative WER reductions of up to 12.5%.

2. Challenges in multilingual modeling can be addressed with a true, first-pass multilingual model with no language-id followed by a second-pass rescoring using a single, multilingual Neural Oracle Search model. We demonstrate an average Word Error Rate (WER) reduction of 9.4% relative across 5 languages, thus outperforming language-specific NOS models.

3. The multilingual Neural Oracle Search model does not require additional capacity and outperforms its language-specific counterparts.

The rest of the paper is organized as follows. We begin with a description of the proposed multilingual NOS model in Section 2. Section 3 presents the languages, associated training data and model architectures. We present results and discuss use of multilingual models in the first and second-pass and demonstrate the power of the proposed model in Section 4.

2. MULTILINGUAL NEURAL ORACLE SEARCH

In Section 2.1 we begin with a recap of NOS from [26]. We then show two ways of integrating a multilingual first-pass model with NOS. We first discuss integrating language-specific NOS models with a multilingual first-pass in Section 2.2, followed by a single multilingual NOS integrated with a multilingual first-pass in Section 2.3.

2.1. Background

NOS treats the oracle search problem as a sequence classification problem. The input to the algorithm is a sequence of acoustic features, $X = X_1, \cdots, X_T$, a list of up to N hypotheses, $H = H_1, \cdots, H_N$ obtained from beam-search of the first-pass model and external knowledge sources. The output is the index of the most likely hypothesis in the N hypothesis list. The input features used in this paper are encoder activations derived from the first-pass model. The NOS framework integrates information from multiple sources. Any number of external sources, such as general-purpose or domain-specific language models, named-entity detection scores or contextual language models can be combined using the NOS framework. In this paper, we restrict ourselves to three scores for each input pair (X, H) defined below.

1. $S_{\theta_1}(X|H_i)$ is the sequence level unnormalized likelihood score. This score can be decomposed into a prior score that predicts labels given the label history $y_{0:i-1}$. and a posterior score that predicts labels given label history and acoustic features using a cross-entropy objective. The sequence level unnormalized likelihood score $S_{\theta_1}(X|Y)$ for each hypothesis $Y = H_i$ is given by:

$$S_{\theta_1}(X|Y = y_{0:U}) \quad \propto \quad \log P(Y|X) - \log P(Y) \quad (1)$$

Note: For the rest of the paper we refer to the prior and posterior model that make up S_{θ_1} as the **NOS model**.

2. $S_{\theta_2}(H_i)$ is the score given by an external language model trained on text-only data. For the rest of the paper we assume that we have pre-trained language models available for each language.

3. $S_{\theta_3}(i)$ is the score of the hypothesis from the first-pass.

The overall sequence level score of each hypothesis is:

$$Score(H_i, X) = \lambda_1 S_{\theta_1}(X|H_i) + \lambda_2 S_{\theta_2}(H_i) + S_{\theta_3}(i) \quad (2)$$

The probability of H_i being the oracle is given by:

$$P(\text{Oracle} = i|X, H_{1:N}) = \frac{\exp(Score(H_i, X))}{\sum_j \exp(Score(H_j, X))} \quad (3)$$

Three different posterior smoothing functions were proposed in [26]. Model parameters are estimated using a cross-entropy objective between the estimated posterior and a ground truth distribution that assigns all the probability mass to the oracle hypothesis (Eq 3).

2.2. Monolingual NOS with a multilingual first-pass

The various component of the NOS model as described in Section 2.1 can be either language-specific or trained in a multilingual fashion. As mentioned before, we will assume that the external LMs (S_{θ_2}), are pre-trained and language-specific.

In this section, we first look at replacing the first-pass model (S_{θ_3}) with a multilingual version. The setup is shown on the left side of Figure 1. As can be seen, we use a multilingual first-pass model that is trained to recognize multiple languages. For each language, we also train language-specific NOS (prior and posterior) models (S_{θ_1}). Here, we will assume that we have access to an external language-id which will help us select the appropriate language-specific models for NOS and the external LM. Note that, we do not assume that the first-pass model has access to the language-id.

The first-pass model gives a set of hypotheses, which are then rescored by the appropriate language-specific NOS model and the language-specific external LM to give scores which are used to find the best hypothesis using Eq 3.

2.3. Multilingual NOS with a multilingual first-pass

We can further replace the language-specific NOS models with one multilingual NOS model as shown on the right

side of Figure 1. The multilingual NOS prior and posterior models are trained on data from all languages with each batch composed of the all languages sampled according to their natural distribution in the training data. In this setting we again assume we have access to an external language-id which we will only use to select the appropriate pre-trained external language-specific LM. We do not require to have the language-id available for use in either the multilingual first-pass model or the second-pass multilingual NOS model. Thus, the first-pass model provides a set of multilingual hypotheses, which are then rescored by the multilingual NOS model and the appropriate language-specific external LM to find the best hypothesis using Eq 3.

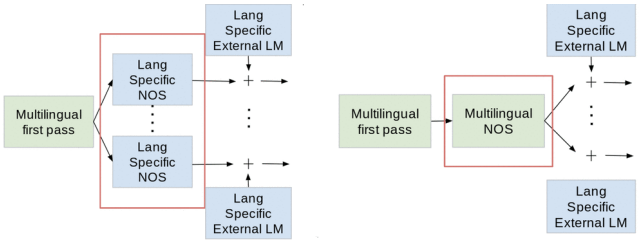

Fig. 1: (Left) Monolingual NOSs integrated with a multilingual first-pass model and monolingual external LMs; (Right) Multilingual NOS integrated with a multilingual first-pass model and monolingual external LMs.

3. EXPERIMENTAL SETUP

3.1. Languages and Data

We conduct experiments on two sets of languages. For the first set we use a set of related and low-resource languages from the Nordic/Germanic cluster of languages: Danish, Finish, Norwegian, Swedish and Dutch. For simplicity we call this set the "Nordics" languages. For the second set of experiments, we add a high-resource language (English) to the Nordics set. We will refer to this set as "Nordics++".

The amounts of training data vary per language, as can be seen in Table 1.

Language	da-dk	fi-fi	nb-no	nl-nl	sv-se	en-us
Train (H)	3.3k	3.8k	5.4k	8.4k	7.5k	52k

Table 1: Training data sizes (in hours) per language.

For all of these languages, the training and test data are anonymized and transcribed by humans.

3.2. Models

NOS Model: The NOS model includes a 2 layer unidirectional LSTM, with a 2 layer label synchronous attention mechanism for posterior scores, as well as a 2 layer unidirectional LSTM for prior scores. All LSTM layers have a dimension of 512 and the attention layers have a dimension of 128. The sequence level prior and posterior scores form $S_{\theta_1}(X|Y = H_i)$ given by Equation 1. There are 13M trainable parameters in NOS model.

First-pass Model: The first-pass model follows an encoder-decoder architecture detailed in [27] with the encoder replaced by a 17 layer Conformer encoder [28], and a 2 layer LSTM decoder trained using the RNN-T loss. For the experiments on "Nordics", we use a model dimension of 512 for the encoder, which leads to a total of 137M parameters, while for the experiments on "Nordics++", we use a model dimension of 768 for the encoder, which leads to a total of 278M parameters. The input acoustic features to the model are 80-dimensional log-Mel features stacked over three frames as described in [29]. The target vocabulary for all models is a wordpiece model containing 4096 wordpieces and generated using data from all languages. All models are trained with an effective batch size of 4096.

4. RESULTS

We train a multilingual first-pass model on data combined from all the languages. This model serves as our baseline model. As mentioned in 2.2, we train monolingual NOS prior and posterior models for each of the languages and use pre-trained language-specific external LMs. This setting is referred to as Mono NOS in Table 2 and Table 3. The multilingual NOS model is trained on the same data as the multilingual first-pass. The architecture of the multilingual NOS model is the same as the language-specific NOS model and uses the same number of total parameters (13M).

4.1. Rescoring for the "Nordics" languages

For this set of experiments we looked at the "Nordics" languages which is a set of related languages with different amounts of data. As can be seen in Table 2, we find that for each of the languages there are significant gains over the first-pass baseline when rescored with monolingual NOS models, with relative gains in range of 3.4% to 12.5%, with average gains of 6.8% relative. We further find that, when the multilingual first-pass is rescored with a multilingual NOS model, the performance is even better, with gains over the first-pass baseline in the range of 7.2% to 11.6% relative, with average gains of 9.4% relative. For almost all languages, we find that the multilingual NOS model performs better than the lang-specific NOS model for that language. This shows that, for these languages, the benefits of multilingual modeling carry over to second-pass rescoring as well.

4.2. Rescoring for the "Nordics++" languages

Here we discuss experiments where we added a high resource language (English) to the set of languages from the previous experiment. We refer to this set as "Nordics++". As mentioned in Section 3.2, here we use a first-pass with a larger

capacity (model dimension = 768). As in the previous experiment, even with monolingual NOS models we see significant gains, compared with the first-pass model, in the range of 4.0% to 9.0% relative with an average gain of 6.0% relative. When replacing the six language-specific NOS models with one multilingual NOS model, of the same capacity, we find that the gains, with respect to the first-pass model, are as good or better. The relative gains range from 5.5% to 10.6% with an average gain of 8.4% relative. It is worth noting that even in the case of the high resource language (English), multilingual NSA with no capacity increase performs as good as the monolingual NSA counterpart.

Language	First-Pass	+Mono NOS	+Multi NOS
da-dk	8.9	8.5	**8.3**
fi-fi	15.2	14.7	**14.0**
nb-no	11.4	**10.0**	10.1
nl-nl	10.1	9.4	**9.1**
sv-se	11.6	10.9	**10.4**

Table 2: WER Comparison of Mono NOS vs Multi NOS on a cluster of related languages (Nordic/Germanic)

Language	First-Pass	+Mono NOS	+Multi NOS
da-dk	8.5	7.9	**7.6**
fi-fi	15.0	14.5	**13.8**
nb-no	10.8	**9.8**	**9.8**
nl-nl	9.7	9.1	**8.8**
sv-se	10.9	10.2	**9.8**
en-us	5.6	**5.3**	**5.3**

Table 3: WER Comparison of Mono NOS vs Multi NOS on a cluster of related languages (Nordic/Germanic) + an unrelated high resource language (en-us)

4.3. Discussion

In order to have a better understanding about the relative gains and regressions of monolingual and multilingual NOS, we analyze the WER contributions of these two models on cases when the top hypothesis from the first pass is not the oracle and case when the top hypothesis from the first pass corresponds to the oracle. Note that when the top hypothesis in the first-pass is the oracle then NOS would incur errors if other hypotheses are scored higher in the second pass. On the other hand, the WER improves if NOS chooses the oracle in cases when the first-pass fails to do so. Table 4 shows the results of this analysis on Swedish from the "Nordics++" experiment.

From Table 4 we can see that when the top hypothesis generated by the first pass corresponds to the oracle, multilingual NOS tends to pick the oracle more frequently than monolingual NOS and thus incurs a lesser relative regressions (9.6% vs 13.4%). Note that, when the top hypothesis from the first pass is the oracle and NOS picks the oracle there are no (0%) relative gains or losses. When the top hypothesis generated by the first pass does not correspond to the oracle, the error reductions due to both monolingual and multilingual NOS are roughly similar. In addition, when the first pass does not pick the oracle correctly, multilingual NOS picks the oracle in the second pass more often than monolingual NOS. This breakdown suggests that in all conditions, multilingual NOS outperforms monolingual NOS. This analysis lends support to the suggestion in [26] that a gating function to decide when to apply the second-pass NOS model would result in further reductions in WER. In other words, the additional errors introduced by the second pass rescorer could potentially be avoided.

Relative WER Contribution	First-Pass picks oracle	First-Pass picks non-oracle
Mono NOS picks oracle	0%	-19.6%
Mono NOS picks non-oracle	+13.4%	-0.7%
Multi NOS picks oracle	0%	-18.7%
Multi NOS picks non-oracle	+9.6%	-1.3%

Table 4: Relative WER contributions of Mono NOS vs Multi NOS with respect to whether the First-Pass model picks the oracle. The results presented are for the Swedish testset from the "Nordics++" experiment. A postive number indicates additional errors incurred, while negative indicates gains and 0 indicates that NOS did not incur any additional gains or losses.

5. CONCLUSIONS

The majority of previous work on multilingual ASR models focus on the first-pass model. In this paper, we present a technique for training multilingual second-pass rescorers using the Neural Oracle Search. We use this technique on two sets of languages, one is a combination of low-resource/related languages, and one consisting of a high-resource/low-resource mix. In both cases, the multilingual NOS model results in consistent gains over the multilingual baseline, with average gains of up to 9.4% relative. In comparison to the monolingual counterparts, we see gains in almost all languages with replacing N models with one model of the same size. The external LM used in this paper is still a monolingual model. In the future, we plan to investigating the training and use of multilingual LMs to further improve low-resource languages, simplify model training, and remove reliance on a known or externally predicted language-id.

6. REFERENCES

[1] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[2] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *arXiv preprint arXiv:1706.02737*, 2017.

[3] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," 2018. [Online]. Available: https://arxiv.org/pdf/1712.01769.pdf

[4] F. Biadsy, M. Ghodsi, and D. Caseiro, "Effectively building tera scale MaxEnt language models incorporating non-linguistic signals," in *Interspeech*, 2017, pp. 2710–2714.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[7] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Rwth asr systems for librispeech: Hybrid vs attention," *arXiv preprint arXiv:1905.03072*, 2019.

[8] M. Harper, "The babel program and low resource speech technology," *Proc. of ASRU 2013*, 2013.

[9] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8239–8243.

[10] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, X. Hainan, H. Lu, H. Sak, I. Leal, N. Gaur, P. Moreno, and Q. Zhang, "Multilingual speech recognition with self-attention structured parameterization," in *Proc. INTERSPEECH*, 2020.

[11] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," *arXiv preprint arXiv:2002.02562*, 2020.

[12] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," 2014, pp. 5569–5573.

[13] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[14] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[15] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *arXiv preprint arXiv:1909.05330*, 2019.

[16] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. M. Mengibar, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of informed experts for multilingual speech recognition," in *ICASSP 2021, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021.

[17] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Multilingual acoustic and language modeling for ethio-semitic languages," *Proc. Interspeech 2020*, pp. 1047–1051, 2020.

[18] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[19] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.-C. Chiu, "Two-pass end-to-end speech recognition," *arXiv preprint arXiv:1908.10992*, 2019.

[20] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[21] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[22] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 1784–1794.

[23] W. Chen, T. yan Liu, Y. Lan, Z. ming Ma, and H. Li, "Ranking measures and loss functions in learning to rank," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., 2009, pp. 315–323. [Online]. Available: http://papers.nips.cc/paper/3708-ranking-measures-and-loss-functions-in-learning-to-rank.pdf

[24] F. Peng, S. Roy, B. Shahshahani, and F. Beaufays, "Search results based n-best hypothesis rescoring with maximum entropy classification," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 422–427.

[25] J. Liu and Y. Zhong, "N-best speech hypothesis reordering based on comprehensive information theory," in *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*. IEEE, 2003, pp. 29–32.

[26] E. Variani, T. Chen, J. Apfel, B. Ramabhadran, S. Lee, and P. Moreno, "Neural oracle search on n-best hypotheses," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7824–7828.

[27] Y. He, T. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. yiin Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," 2019. [Online]. Available: https://arxiv.org/abs/1811.06621

[28] A. Gulati, C.-C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, Eds., *Conformer: Convolution-augmented Transformer for Speech Recognition*, 2020.

[29] A. Waters, N. Gaur, P. Haghani, P. Moreno, and Z. Qu, "Leveraging language id in multilingual end-to-end speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 928–935.