GoldenRetriever: A Speech Recognition System Powered by Modern Information Retrieval

Yuanfeng Song[†‡], Di Jiang[†], Xiaoling Huang[†], Yawen Li[◇]
Qian Xu[†], Raymond Chi-Wing Wong[‡], and Qiang Yang[†‡]

[†]AI Group, WeBank Co., Ltd, Shenzhen, China
[‡]Department of CSE, The Hong Kong University of Science and Technology, Hong Kong, China
[◇]Beijing University of Posts and Telecommunications, Beijing, China
{songyf,raywong,qyang}@cse.ust.hk,{dijiang,smallhuang,qianxu}@webank.com

ABSTRACT

Existing Automatic Speech Recognition (ASR) systems usually generate the N-best hypotheses list first, and then rescore them with the language model score and the acoustic model score to find the best one. This procedure is essentially analogous to the working mechanism of modern Information Retrieval (IR) systems, which retrieve a relatively large amount of relevant candidates first, re-rank them, and output the top-N list. Exploiting their commonality, this demonstration proposes a novel system named GoldenRetriever that marries IR with ASR. GoldenRetriever transforms the problem of N-best hypotheses rescoring as a Learning-to-Rescore (L2RS) problem and utilizes a wide range of features beyond the language model score and the acoustic model score. In this demonstration, the audience can experience the great potential of marrying IR with ASR for the first time. GoldenRetriever should inspire more research on transferring the state-of-the-art IR techniques to ASR.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; • **Computing methodologies** → **Speech recognition**;

KEYWORDS

N-best rescoring, learning-to-rescore, speech recognition

ACM Reference Format:
Yuanfeng Song, Di Jiang, Xiaoling Huang, Yawen Li, Qian Xu, Raymond Chi-Wing Wong, Qiang Yang. 2020. GoldenRetriever: A Speech Recognition System Powered by Modern Information Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3394171.3414392

1 INTRODUCTION

Due to the ubiquity of mobile devices in daily life, speech communication has been gaining great momentum in recent years, which further promotes the need for high-quality Automatic Speech Recognition (ASR) systems. Existing rescoring approaches first evaluate each N-best hypothesis using an Acoustic Model (AM) and a Language Model (LM), and then utilize these scores to form the final ranking lists. Such a two-step approach suffers two drawbacks: 1) It aims to optimize an unnecessarily harder problem, namely predicting accurate grammatical legitimacy scores of the N-best hypotheses rather than directly predicting their partial orders with respect to a specific acoustic input. 2) Quite limited features are utilized, and most of the recent representation features given by advanced models in Natural Language Processing (NLP), such as BERT embedding [2], is hard to use under the existing rescoring pipeline.

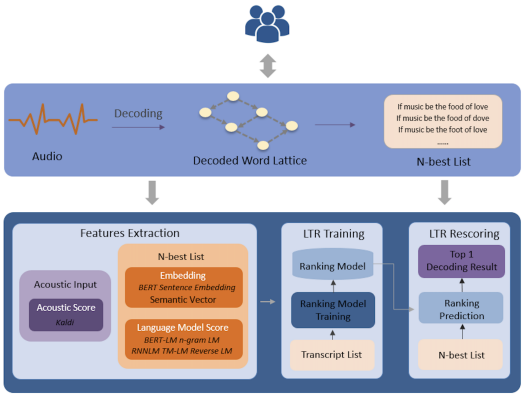

Figure 1: A System Overview of GoldenRetriever

In this demonstration, we present a novel system named *GoldenRetriever*, which is designed to empower ASR systems with state-of-the-art IR techniques to relieve above drawbacks. GoldenRetriever transforms the problem of N-best hypotheses rescoring as a Learning-to-Rescore (L2RS) [7] problem and utilizes a wide range of features beyond the LM and AM scores. The efficacy of L2RS relies in the design of the features. In GoldenRetriever, the features including n-gram LM, BERT sentence embedding, RNNLM score [4], reverse n-gram LM score, and AM score, are from the lexical level to the semantic level. Finally, GoldenRetriever learns a Ranking SVM model [1] which combines these features to formulate a rescoring model. Experimental results validate that L2RS can easily beat its rescoring counterparts such as RNNLM, EC-Classifier [5] and NS2TLM [8]. Through this demonstration, the user will have the ultimate opportunity to experience the mechanisms of L2RS techniques for ASR in a vivid and interactive approach. They can

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3414392

switch between the basic rescoring method and L2RS, and observe their real-time impact on the N-best lists and the final ASR results. To the best of our knowledge, this is the first system that utilizes L2RS as the rescoring method for ASR. We expect that it will inspire more research on transferring IR to advance the ASR field.

2 SYSTEM DESCRIPTION

ASR Pipeline An overview of the GoldenRetriever system is shown in Figure 1. GoldenRetriever follows the classical ASR pipeline and the final recognition result $\mathbf{w}^*$ for a given acoustic input $\mathbf{a}$ is defined as follows:

$$\mathbf{W} \propto \arg\max(\log P_{LM}(\mathbf{w}) + \log P_{AM}(\mathbf{a}|\mathbf{w})), \tag{1}$$

$$\mathbf{w}^* = \arg \max_{j \in [1,N] \& \mathbf{w}_j \in \mathbf{W}} f(\phi(\mathbf{a}, \mathbf{w_j})), \tag{2}$$

where P_{LM} is an LM score and P_{AM} is an AM score, and $\mathbf{W} = (\mathbf{w_1}, \cdots, \mathbf{w_N})$ is the N-best list. We use $f(\cdot)$ to denote the rescoring module, and $\phi(\mathbf{a}, \mathbf{w})$ for the feature-vector representation of pair $(\mathbf{a}, \mathbf{w})$. GoldenRetriever first generates the N-best list, then uses the L2RS technique for modeling the rescoring function $f(\cdot)$, which involves three steps: feature extraction, model training and reranking, as shown in Figure 1. During the L2RS prediction step, after the ASR system generates the N-best list $\mathbf{W}$, the final decoding result $\mathbf{w}^*$ can be obtained using Eq. (2).

Features The textual and acoustic features used in L2RS are from the lexical level to the semantic level, with details listed in Table 1.

Table 1: Features for Learning-to-Rescore

Feature	Detail
Trigram LM (lm)	A trigram LM trained using the transcripts
Reverse Trigram LM (rlm)	A reverse trigram LM trained using the transcripts
BERT Embedding (bert_se)	A 1024-dimension sentence embedding using bert-as-service toolkit [9]
RNNLM (rnnlm)	The LM score produced by an RNNLM trained using [4]
AM (am)	The AM score produced by Kaldi toolkit [6]

3 PERFORMANCE ANALYSIS

We proceed to show a quantitative evaluation result of the GoldenRetriever system on an industrial dataset. We build the whole system based on the open-sourced Kaldi toolkit [6]. The dataset used contains around 8000 hours of conversational speech collected from a real-life customer service scenario in Mandarin Chinese. We reserve 80% of data for the training set and split the rest for development (10%) and testing (10%). The whole system is deployed on a machine with 314GB memory, 72 Intel Core Processor (Xeon), Tesla K80 GPU and CentOS, and the evaluations are conducted under the same hardware configuration. The experimental results are shown in Table 2. We observed that L2RS can significantly improve the performance in terms of both NDCG@10 [3] and WER. Specifically, features such as BERT sentence embedding is quite effective, and the effectiveness of these embedding features are quite hard to evaluate under the traditional Kaldi rescoring pipeline.

Table 2: NDCG@10 and WER

Model	NDCG@10	WER
Kaldi Baseline	0.733	15.07%
L2RS(am+lm)	0.726	15.10%
L2RS(am+lm+rlm)	0.736	14.91%
L2RS(am+lm+rnnlm)	0.759	14.03%
L2RS(am+lm+rlm+rnnlm)	0.762	13.96%
L2RS(am+lm+rlm+rnnlm+bert_se)	0.792	13.41%

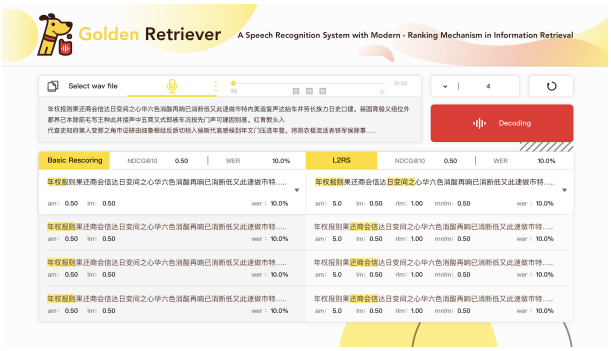

Figure 2: The Screenshot of GoldenRetriever

4 DEMONSTRATION OVERVIEW

The goal of the demonstration is to provide an interactive approach for users to experience the working mechanisms of the L2RS techniques mentioned above. Figure 2 shows a screenshot of the user interface with three basic functionalities.

Speech & Text Input: Two kinds of audio inputs are supported in GoldenRetriever: audio files and real-time speech recorded by the microphone provided in the system. In order to evaluate the WER and NDCG@n of different rescoring techniques, the groundtruth transcript is also required to be provided by the user.

Baseline Rescoring Result: When the above-mentioned audio and text files are ready, the user can click the "Decoding" button to start the decoding process. The N-best hypotheses together with the final decoded results will be presented to the corresponding area in an ordered list after decoding is complete, together with the NDCG@n value and the overall WER. The baseline Rescoring result section of the webpage shows the ranking list given by the baseline n-gram LM and the AM score. Each hypothesis is also provided with the LM score, AM score, and WER.

L2RS Rescoring Result: The L2RS result section shows the ranking list given by the L2RS approach. The user can compare the ranking results and clearly see the difference since the results given by these two approaches are simultaneously presented in a side-by-side fashion. Each hypothesis is also provided with the feature values discussed in Section 2 as well as its WER.

5 CONCLUSION

In this demonstration, we show a novel ASR system named GoldenRetriever which incorporates the Learning-to-Rescore mechanism for the rescoring step, and have proven its superiority compared with its rescoring counterparts.

ACKNOWLEDGEMENT

This research is partially supported by HKRGC GRF 16214017.

REFERENCES

[1] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. 2006. Adapting ranking SVM to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 186–193.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[3] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering* 16, 1 (2010), 100–103.

[4] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH2010*.

[5] Atsunori Ogawa, Marc Delcroix, Shigeki Karita, and Tomohiro Nakatani. 2018. Rescoring N-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6099–6103.

[6] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

[7] Yuanfeng Song, Di Jiang, Xuefang Zhao, Qian Xu, Raymond Chi-Wing Wong, Lixin Fan, and Qiang Yang. 2019. L2RS: A Learning-to-Rescore Mechanism for Automatic Speech Recognition. *arXiv preprint arXiv:1910.11496* (2019).

[8] Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, and Yushi Aono. 2018. Neural speech-to-text language models for rescoring hypotheses of dnn-hmm hybrid automatic speech recognition systems. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 196–200.

[9] Han Xiao. 2018. bert-as-service. https://github.com/hanxiao/bert-as-service.