

IMPROVING SPOKEN LANGUAGE UNDERSTANDING BY EXPLOITING ASR N-BEST HYPOTHESES

Mingda Li^{} [‡], Weitong Ruan[†], Xinyue Liu[†], Luca Soldaini[†], Wael Hamza[†], Chengwei Su[†]*

^{*} University of California, Los Angeles, USA

[†] Amazon Alexa AI, USA

ABSTRACT

In a modern spoken language understanding (SLU) system, the natural language understanding (NLU) module takes interpretations of a speech from the automatic speech recognition (ASR) module as the input. The NLU module usually uses the first best interpretation of a given speech in downstream tasks such as domain and intent classification. However, the ASR module might misrecognize some speeches and the first best interpretation could be erroneous and noisy. Solely relying on the first best interpretation could make the performance of downstream tasks non-optimal. To address this issue, we introduce a series of simple yet efficient models for improving the understanding of semantics of the input speeches by collectively exploiting the n -best speech interpretations from the ASR module.

Index Terms— ASR n-best hypotheses integration, spoken language understanding

1. INTRODUCTION

Currently, voice-controlled smart devices are widely used in multiple areas to fulfill various tasks, e.g. playing music, acquiring weather information and booking tickets. The SLU system employs several modules to enable the understanding of the semantics of the input speeches. When there is an incoming speech, the ASR module picks it up and attempts to transcribe the speech. An ASR model could generate multiple interpretations for most speeches, which can be ranked by their associated confidence scores. Among the n -best hypotheses, the top-1 hypothesis is usually transformed to the NLU module for downstream tasks such as domain classification, intent classification and named entity recognition (slot tagging). Multi-domain NLU modules are usually designed hierarchically [1]. For one incoming utterance, NLU modules will firstly classify the utterance as one of many possible domains and the further analysis on intent classification and slot tagging will be domain-specific.

In spite of impressive development on the current SLU pipeline, the interpretation of speech could still contain errors.

Sometimes the top-1 recognition hypothesis of ASR module is ungrammatical or implausible and far from the ground-truth transcription [2, 3]. Among those cases, we find one interpretation exact matching with or more similar to transcription can be included in the remaining hypotheses ($2^{nd} - n^{th}$).

To illustrate the value of the $2^{nd} - n^{th}$ hypotheses, we count the frequency of exact matching and more similar (smaller edit distance compared to the 1^{st} hypothesis) to transcription for different positions of the n -best hypotheses list. Table 1 exhibits the results. For the explored dataset, we only collect the top 5 interpretations for each utterance ($n = 5$). Notably, when the correct recognition exists among the 5 best hypotheses, 50% of the time (sum of the first row’s percentages) it occurs among the $2^{nd} - 5^{th}$ positions. Moreover, as shown by the second row in Table 1, compared to the top recognition hypothesis, the other hypotheses can sometimes be more similar to the transcription.

Table 1: Spoken recognition quality distribution of the n best hypotheses.

n Best Rank Position	2^{nd}	3^{rd}	4^{th}	5^{th}
Match	19%	14%	10%	7%
Prob (better than 1^{st} best)	22%	17%	16%	15%

Over the past few years, we have observed the success of reranking the n -best hypotheses [2, 4, 5, 6, 7, 8, 9, 10, 11] before feeding the best interpretation to the NLU module. These approaches propose the reranking framework by involving morphological, lexical or syntactic features [9, 10, 11], speech recognition features like confidence score [2, 5], and other features like number of tokens, rank position [2]. They are effective to select the best from the hypotheses list and reduce the word error rate (WER) [12] of speech recognition.

Those reranking models could benefit the first two cases in Table 2 when there is an utterance matching with transcription. However, in other cases like the third row, it is hard to integrate the fragmented information in multiple hypotheses.

This paper proposes various methods integrating n -best hypotheses to tackle the problem. To the best of our knowledge, this is the first study that attempts to collectively exploit the n -best speech interpretations in the SLU system. This paper serves as the basis of our n -best-hypotheses-based

[‡] This work was done while the first author was an intern at Amazon.

Table 2: Motivating example: comparison of ASR n -Best hypotheses with the corresponding transcription.

Transcription	1 st best	2 nd best	3 rd best
play muse	play news	play muse	play mus
track on bose	check on bowls	check on bose	track on bose
harry porter	how patter	how patter	harry power

SLU system, focusing on the methods of integration for the hypotheses. Since further improvements of the integration framework require considerable setup and descriptions, where jointly optimized tasks (e.g. transcription reconstruction) trained with multiple ways (multitask [13], multistage learning [14]) and more features (confidence score, rank position, etc.) are involved, we leave those to a subsequent article.

This paper is organized as follows. Section 2 introduces the Baseline, Oracle and Direct models. Section 3 describes proposed ways to integrate n -best hypotheses during training. The experimental setup and results are described in Section 4. Section 5 contains conclusions and future work.

2. BASELINE, ORACLE AND DIRECT MODELS

2.1. Baseline and Oracle

The preliminary architecture is shown in Fig. 1. For a given transcribed utterance, it is firstly encoded with Byte Pair Encoding (BPE) [15], a compression algorithm splitting words to fundamental subword units (*pairs of bytes* or *BPs*) and reducing the embedded vocabulary size. Then we use a BiLSTM [16] encoder and the output state of the BiLSTM is regarded as a vector representation for this utterance. Finally, a fully connected Feed-forward Neural Network (FNN) followed by a softmax layer, labeled as a multilayer perceptron (MLP) module, is used to perform the domain/intent classification task based on the vector.

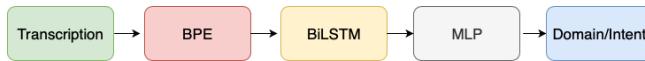


Fig. 1: Baseline pipeline for domain or intent classification.

For convenience, we simplify the whole process in Fig.1 as a mapping BM (Baseline Mapping) from the input utterance S to an estimated tag’s probability $p(\tilde{t})$, where $p(\tilde{t}) \leftarrow BM(S)$. The *Baseline* is trained on transcription and evaluated on ASR 1st best hypothesis ($S = \text{ASR } 1^{\text{st}} \text{ best}$). The *Oracle* is trained on transcription and evaluated on transcription ($S = \text{Transcription}$). We name it Oracle simply because we assume that hypotheses are noisy versions of transcription.

2.2. Direct Models

Besides the Baseline and Oracle, where only ASR 1-best¹ hypothesis is considered, we also perform experiments to utilize

¹We use ASR n -best hypotheses or n -bests to denote the top n interpretations of a speech, and the 1,5-best standing for the top 1 or 5 hypotheses.

ASR n -best hypotheses during evaluation. The models evaluating with n -bests and a BM (pre-trained on transcription) are called *Direct Models* (in Fig. 2):

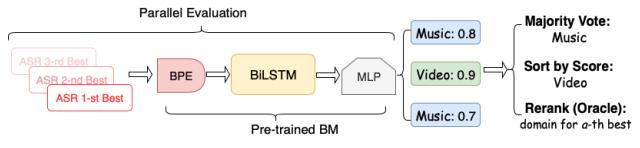


Fig. 2: Direct models evaluation pipeline.

- **Majority Vote.** We apply the BM model on each hypothesis independently and combine the predictions by picking the majority predicted label, i.e. Music.
- **Sort by Score.** After parallel evaluation on all hypotheses, sort the prediction by the corresponding confidence score and choose the one with the highest score, i.e. Video.
- **Rerank (Oracle).** Since the current rerank models (e.g., [2, 4, 5]) attempt to select the hypothesis most similar to transcription, we propose the Rerank (Oracle), which picks the hypothesis with the smallest edit distance to transcription (assume it is the a -th best) during evaluation and uses its corresponding prediction.

3. INTEGRATION OF N-BEST HYPOTHESES

All the above mentioned models apply the BM trained on one interpretation (transcription). Their abilities to take advantage of multiple interpretations are actually not trained. As a further step, we propose multiple ways to integrate the n -best hypotheses during training. The explored methods can be divided into two groups as shown in Fig. 3. Let H_1, H_2, \dots, H_n denote all the hypotheses from ASR and $bp_{H_k, i} \in BPs$ denotes the i -th pair of bytes (BP) in the k th best hypothesis. The model parameters associated with the two possible ways both contain: embedding e_{bp} for pairs of bytes, BiLSTM parameters θ and MLP parameters W, b .

3.1. Hypothesized Text Concatenation

The basic integration method (*Combined Sentence*) concatenates the n -best hypothesized text. We separate hypotheses with a special delimiter (<SEP>). We assume BPE totally produces m BPs (delimiters are not split during encoding). Suppose the n^{th} hypothesis has j pairs. The entire model can be formulated as:

$$(h_1, \dots, h_m) \leftarrow BiLSTM_\theta(bp_{H_1, 1}, \dots, bp_{<\text{sep}>}, \dots, bp_{H_n, j}) \quad (1)$$

$$p(\tilde{t}) = \sigma(W[h_{1b}, h_{m_f}] + b) \quad (2)$$

In Eqn. 1, the connected hypotheses and separators are encoded via BiLSTM to a sequence of hidden state vectors. Each hidden state vector, e.g. h_1 , is the concatenation of forward h_{1f} and backward h_{1b} states. The concatenation of the

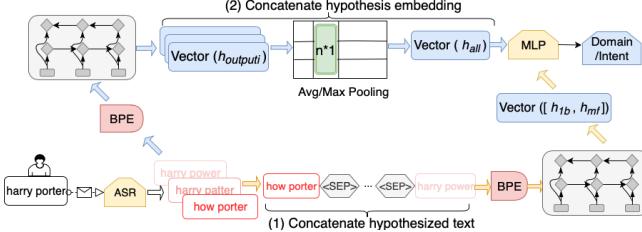


Fig. 3: Integration of n -best hypotheses with two possible ways: 1) concatenate hypothesized text and 2) concatenate hypothesis embedding.

last state of the forward and backward LSTM forms the output vector of BiLSTM (concatenation denoted as $[,]$). Then, in Eqn. 2, the MLP module defines the probability of a specific tag (domain or intent) \tilde{t} as the normalized activation (σ) output after linear transformation of the output vector.

3.2. Hypothesis Embedding Concatenation

The concatenation of hypothesized text leverages the n -best list by transferring information among hypotheses in an embedding framework, BiLSTM. However, since all the layers have access to both the preceding and subsequent information, the embedding among n -bests will influence each other, which confuses the embedding and makes the whole framework sensitive to the noise in hypotheses.

As the second group of integration approaches, we develop models, *PoolingAvg/Max*, on the concatenation of hypothesis embedding, which isolate the embedding process among hypotheses and summarize the features by a pooling layer. For each hypothesis (e.g., i^{th} best in Eqn. 3 with j pairs of bytes), we could get a sequence of hidden states from BiLSTM and obtain its final output state by concatenating the first and last hidden state (h_{output_i} in Eqn. 4). Then, we stack all the output states vertically as shown in Eqn. 5. Note that in the real data, we will not always have a fixed size of hypotheses list. For a list with $r (< n)$ interpretations, we get the embedding for each of them and pad with the embedding of the first best hypothesis until a fixed size n . When $r \geq n$, we only stack the top n embeddings. We employ h_{output_1} for padding to enhance the influence of the top 1 hypothesis, which is more reliable. Finally, one unified representation could be achieved via *Pooling* (Max/Avg pooling with n by 1 sliding window and stride 1) on the concatenation and one score could be produced per possible tag for the given task.

$$(h_{H_i,1}, \dots, h_{H_i,j}) \leftarrow BiLSTM_\theta(bp_{H_i,1}, \dots, bp_{H_i,j}) \quad (3)$$

$$h_{output_i} = [h_{H_i,1b}, h_{H_i,jf}] \quad (4)$$

$$h_{outputs} = \left\{ \begin{array}{ll} \left\{ \begin{array}{l} h_{output_1} \\ \dots \\ h_{output_r} \end{array} \right\} & r - bests \\ \left\{ \begin{array}{l} h_{output_1} \\ \dots \end{array} \right\} & Padding \text{ with } h_{output_1} \end{array} \right\} \quad (5)$$

$$h_{all} = Pooling(h_{outputs}) \quad (6)$$

$$p(\tilde{t}) = \sigma(Wh_{all} + b) \quad (7)$$

4. EXPERIMENT

4.1. Dataset

We conduct our experiments on $\sim 8.7M$ annotated anonymised user utterances. They are annotated and derived from requests across 23 domains.

4.2. Performance on Entire Test Set

Table 3 shows the relative error reduction (RErr)² of Baseline, Oracle and our proposed models on the entire test set ($\sim 300K$ utterances) for multi-class domain classification. We can see among all the direct methods, predicting based on the hypothesis most similar to the transcription (Rerank (Oracle)) is the best.

Table 3: Micro and Macro F1 score for multi-class domain classification.

Category	Model	RErr(%)
	Baseline	0.00
Integration	PoolingAvg	14.29
	PoolingMax	13.20
	Combined Sentence	11.67
Direct	Sort by Score	1.85
	Majority Vote	1.64
	Rerank (Oracle)	3.71
	Oracle	27.04

As for the other models attempting to integrate the n -bests during training, PoolingAvg gets the highest relative improvement, 14.29%. It also turns out that all the integration methods outperform direct models drastically. This shows that having access to n -best hypotheses during training is crucial for the quality of the predicted semantics.

4.3. Performance Comparison among Various Subsets

Table 4: Performance comparison for the subset ($\sim 19\%$) where ASR first best disagrees with transcription.

Category	Model	RErr(%)
	Baseline	0.00
Integration	PoolingAvg	24.67
	PoolingMax	26.23
	Combined Sentence	19.23
Direct	Sort by Score	9.95
	Majority Vote	7.59
	Rerank (Oracle)	7.25
	Oracle	53.02

To further detect the reason for improvements, we split the test set into two parts based on whether ASR first best agrees with transcription and evaluate separately. Comparing Table 4 and Table 5, obviously the benefits of using multiple hypotheses are mainly gained when ASR 1st best disagrees

²The RErr for a model m is calculated by comparing the relative difference between $100\% - MicroF1_m$ and $100\% - MicroF1_{Baseline}$.

Table 5: Performance comparison for the subset ($\sim 81\%$) where ASR first best agrees with transcription.

Category	Model	RErr(%)
	Baseline	0.00
Integration	PoolingAvg	3.56
	PoolingMax	-0.38
	Combined Sentence	4.50
Direct	Sort by Score	-8.269
	Majority Vote	-3.19
	Rerank (Oracle)	0.00
	Oracle	0.00

with the transcription. When ASR 1st best agrees with transcription, the proposed integration models can also keep the performance. Under that condition, we can still improve a little (3.56%) because, by introducing multiple ASR hypotheses, we could have more information and when the transcription/ASR 1st best does not appear in the training set's transcriptions, its n -bests list may have similar hypotheses included in the training set's n -bests. Then, our integration model trained on n -best hypotheses as well has clue to predict. The series of comparisons reveal that our approaches integrating the hypotheses are robust to the ASR errors and whenever the ASR model makes mistakes, we can outperform more significantly.

4.4. Improvements on Different Domains and Different Numbers of Hypotheses

Among all the 23 domains, we choose 8 popular domains for further comparisons between the Baseline and the best model of Table 3, PoolingAvg. Fig. 4 exhibits the results. We could find the PoolingAvg consistently improves the accuracy for all 8 domains.

In the previous experiments, the number of utilized hypotheses for each utterance during evaluation is five, which means we use the top 5 interpretations when the size of ASR recognition list is not smaller than 5 and use all the interpretations otherwise. Changing the number of hypotheses while evaluation, Fig. 5 shows a monotonic increase with the access to more hypotheses for the PoolingAvg and PoolingMax (Sort by Score is shown because it is the best achievable direct model while the Rerank (Oracle) is not realistic). The growth becomes gentle after four hypotheses are leveraged.

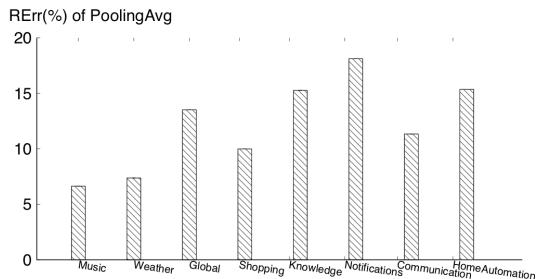


Fig. 4: Improvements on important domains.

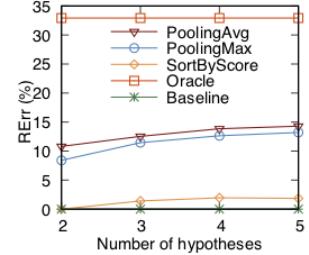


Fig. 5: The influence of different amount of hypotheses.

4.5. Intent Classification

Table 6: Intent classification for three important domains.

Domain	Metric	Shopping	Knowledge	Communication
Baseline		0.0	0.0	0.0
Oracle	RErr (%)	47.63	40.28	32.89
PoolingAvg		25.55	25.00	11.92

Since another downstream task, intent classification, is similar to domain classification, we just show the best model in domain classification, PoolingAvg, on domain-specific intent classification for three popular domains due to space limit. As Table 6 shows, the margins of using multiple hypotheses with PoolingAvg are significant as well.

5. CONCLUSIONS AND FUTURE WORK

This paper improves the SLU system robustness to ASR errors by integrating n -best hypotheses in different ways, e.g. the aggregation of predictions from hypotheses or the concatenation of hypothesis text or embedding. We can achieve significant classification accuracy improvements over production-quality baselines on domain and intent classifications, 14% to 25% relative gains. The improvement is more significant for a subset of testing data where ASR first best is different from transcription. We also observe that with more hypotheses utilized, the performance can be further improved. In the future, we aim to employ additional features (e.g. confidence scores for hypotheses or tokens) to integrate n -bests more efficiently, where we can train a function f to obtain a weight for each hypothesis embedding before pooling. Another direction is using deep learning framework to embed the word lattice [17] or confusion network [18, 19], which can provide a compact representation of multiple hypotheses and more information like times, in the SLU system.

6. ACKNOWLEDGEMENTS

We would like to thank Junghoo (John) Cho for proofreading.

7. REFERENCES

- [1] Gokhan Tur and Renato De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.
- [2] Fuchun Peng, Scott Roy, Ben Shahshahani, and Françoise 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.
- [3] Preethi Jyothi, Leif Johnson, Ciprian Chelba, and Brian Strope, “Large-scale discriminative language model reranking for voice-search,” in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012, pp. 41–49.
- [4] Eugene Charniak and Mark Johnson, “Coarse-to-fine n-best parsing and maxent discriminative reranking,” in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 173–180.
- [5] Fabrizio Morbini, Kartik Audhkhasi, Ron Artstein, Maarten Van Segbroeck, Kenji Sagae, Panayiotis Georgiou, David R Traum, and Shri Narayanan, “A reranking approach for recognition and classification of speech input in conversational dialogue systems,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 49–54.
- [6] Erinç Dikici, Murat Semerci, Murat Saraclar, and Ethem Alpaydin, “Classification and ranking approaches to discriminative language modeling for asr,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 291–300, 2012.
- [7] Hasim Sak, Murat Saralar, and Tunga Güngör, “Discriminative reranking of asr hypotheses with morphological and n-best-list features,” *2011 IEEE Workshop on Automatic Speech Recognition —& Understanding*, pp. 202–207, 2011.
- [8] Hasim Sak, Murat Saraclar, and Tunga Güngör, “On-the-fly lattice rescoring for real-time automatic speech recognition,” in *Eleventh annual conference of the international speech communication association*, 2010.
- [9] Hasim Sak, Murat Saraclar, and Tunga Gungor, “Discriminative reranking of ASR hypotheses with morphological and n-best-list features,” in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011*, 2011, pp. 202–207.
- [10] Michael Collins, Brian Roark, and Murat Saraclar, “Discriminative syntactic language modeling for speech recognition,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 507–514.
- [11] Ho Yin Chan and Phil Woodland, “Improving broadcast news transcription by lightly supervised discriminative training,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 1, pp. I–737.
- [12] Takanobu Oba, Takaaki Hori, and Atsushi Nakamura, “An approach to efficient generation of high-accuracy and compact error-corrective models for speech recognition,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.
- [13] Rich Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [14] Pinghua Gong, Jieping Ye, and Changshui Zhang, “Multi-stage multi-task feature learning,” *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2979–3010, 2013.
- [15] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” *arXiv preprint arXiv:1508.07909*, 2015.
- [16] Mike Schuster and Kuldip K Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [17] Xunying Liu, Yongqiang Wang, Xie Chen, Mark JF Gales, and Philip C Woodland, “Efficient lattice rescoring using recurrent neural network language models,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4908–4912.
- [18] Dilek Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, and Gokhan Tur, “Beyond asr 1-best: Using word confusion networks in spoken language understanding,” *Computer Speech & Language*, vol. 20, no. 4, pp. 495–514, 2006.
- [19] Gokhan Tur, Jerry Wright, Allen Gorin, Giuseppe Riccardi, and Dilek Hakkani-Tür, “Improving spoken language understanding using word confusion networks,” in *Seventh International Conference on Spoken Language Processing*, 2002.