On The Diversity of ASR Hypotheses In Spoken Language Understanding

Surya Kant Sahu
Skit.ai, India
surya.oju@pm.me

Swaraj Dalmia
Skit.ai, India

Abstract

In Conversational AI, an Automatic Speech Recognition (ASR) system is used to transcribe the user's speech, and the output of the ASR is passed as an input to a Spoken Language Understanding (SLU) system, which outputs semantic objects (such as intent, slot-act pairs, etc.). Recent work, including the state-of-the-art methods in SLU utilize either Word lattices or N-Best Hypotheses from the ASR. The intuition given for using N-Best instead of 1-Best is that the hypotheses provide extra information due to errors in the transcriptions of the ASR system, i.e., the performance gain is attributed to the word-error-rate (WER) of the ASR. We empirically show that the gain in using N-Best hypotheses is related to not WER but to the diversity of hypotheses. Code and datasets are available at this URL.

1 Introduction

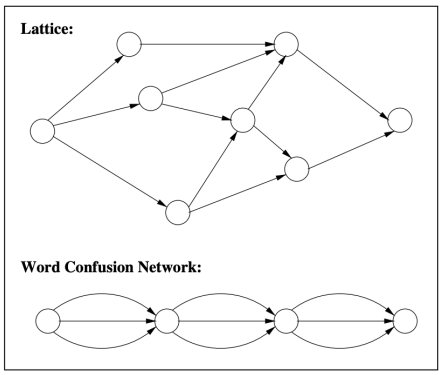
In Conversational AI [1], an Automatic Speech Recognition (ASR) system first transcribes a user's speech. A Spoken Language Understanding (SLU) system parses the speech into semantic objects such as intents and slots. These intents or slots are used by dialog policy to produce the bot's response.

As the ASR is not perfect, the transcripts are noisy and contain errors, so the SLU system needs to be aware of and robust to the errors caused by ASR. Prior work enables this "awareness of ASR errors" by providing extra information such as Word Lattices or Word Confusion Networks [2] or N-Best lists [3] in case WCNs are not available. Prior work [4, 5] argues that using WCNs is better when compared to using 1-Best hypothesis as the oracle path has the lowest word error with respect to the ground truth transcript; similar arguments are made in the case of N-Best hypotheses [3, 6]. In this work, we argue that this assumption that the benefit of using WCNs or N-Best hypotheses is only due to the errors made by the ASR system might not be the entire truth. We empirically show that the diversity of the hypotheses plays a crucial role in determining the gain in performance when using 1-Best v/s. N-Best.

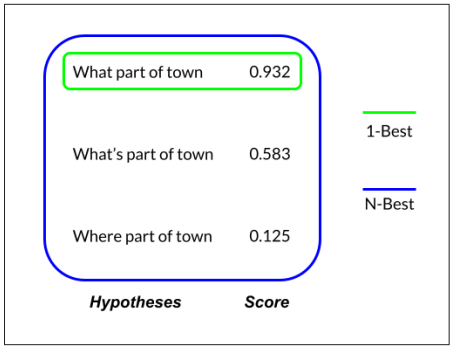
This work investigates the conditions under which N-Best ASR Hypotheses-based methods are more beneficial and show that the diversity in ASR Hypotheses is an important factor. The contributions of this work are summarized below:

- Propose novel metrics to measure the diversity of ASR hypotheses.
- Show that WER of the 1-Best ASR hypothesis is not indicative of how useful the non-1-Best hypotheses are.
- Show that the ASR hypotheses' diversity might indicate the gain in performance.

I Can't Believe It's Not Better Workshop at NeurIPS 2022.



(a) Visual representation of word lattices and word confusion networks. Each transition adds a word to the hypothesis sub-sequence. This figure is taken from [2].



(b) In N-Best lists, the hypotheses are ranked according to a *score*, which is usually a function of the log probability and the length of the hypothesis. The 1-Best hypothesis is the hypothesis with the best score, whereas the N-best hypothesis refers to the set of top-N hypotheses.

Figure 1: Illustrations of various data structures relating to ASR hypotheses.

2 Related Work

2.1 Notation

Let D and $|D|$ denote the dataset, and the number of examples in the dataset, respectively. N is the number of ASR hypotheses considered, x_i^j is the j-th-best hypothesis in the i-th instance in D. y_i is the corresponding intent/slot label. x_i^* is the "gold" or ground-truth transcript obtained from the ASR dataset. $|x_i|$ denotes the number of tokens in x_i. We denote $\mathcal{L}_{ce}$ as the Cross-Entropy loss function. $f(., \theta)$ is a neural network parameterized by θ, a set of parameters.

Throughout the work we discuss the word-error-rate (WER), which is defined as below:

$$WER_i = \frac{S + D + I}{N} \tag{1}$$

Here, S, D, and I are the numbers of substitution, deletion, and insertion errors. N is the total number of tokens in x_i^*. WER is computed between x_i^1 and x_i^*. WER is averaged across the dataset D, and we denote the averaged WER as simply WER.

2.2 Word Confusion Networks and Word Lattices

Word lattices [7] and Word Confusion Networks (WCNs) [2, 8] as shown in Figure 1a provides a graphical representation of hypotheses, where each path from the start to the end node represents 1 hypothesis. Lattices are a normalised and a more compact topology of WCNs; while WCNs retain the topology of the original search space.

Liu et al. [8] show that using WCN as input yields better results for intent performance (2% to 4%) compared to the N-Best hypotheses and Word Lattices. There are similar improvements over 1-Best and N-Best on using word lattices as discussed in [7].

2.3 N-Best ASR Transformer

Ganesan et al. [3] propose a method that concatenates N-Best ASR hypotheses, illustrated in Fig. 1b, that feeds to a pretrained language model (LM). This simple method outperformed state-of-the-art methods by leveraging information contained in non-1-Best alternatives while having several benefits over prior baselines, such as leveraging pretrained LMs, and plug-and-play support for third-party ASR services.

Hypotheses	what part of town **what's** part of town **where** part of town	i wonder transaction problem **one a** transaction problem i **wondered** transaction problem	i want to listen to the seventies music i **went** to listen to seventies music i want to listen to seventies **and** music
# new tokens	2	3	2
Dataset	DSTC-2	In-house	SNIPS-TTS

Figure 2: Examples of Hypotheses from each of the datasets used in this paper. In the first row, the first line is the 1-Best hypothesis, second and third lines are 2^{nd}-best and 3^{rd}-best hypotheses, respectively. *# new tokens* denoted the number of tokens present in non-1-Best hypothesis not present in the 1-Best hypothesis (denoted by **boldface**). We note that DSTC-2 and SNIPS-TTS have highly similar hypotheses on qualitative observation, followed by the In-house dataset.

The input is constructed as follows:

$$\hat{x_i} = \begin{cases} [CLS] \oplus \bigoplus_{j=1}^{N} \{x_i^j \oplus [SEP]\} & \text{if } N > 1 \\ [CLS] x_i^1 [SEP] & \text{if } N = 1 \end{cases} \tag{2}$$

where $\bigoplus$ is the string concatenation operation, and $[CLS]$ and $[SEP]$ are special tokens denoting classifier and separator tokens.

The loss used is $\mathcal{L}_{ce}(\hat{x}, y, \theta)$.

3 Experiments

Following [3], the proposed methods in this section utilize a pretrained XLM-RoBERTa (base). The set of parameters is denoted by θ.

We train models in cases when a) top-N ASR hypotheses are available, b) only 1-Best ASR hypothesis is available and for completeness, c) when only ground-truth transcripts are available. We report Macro F1 scores on ASR hypotheses as inputs. The code to reproduce results on the public datasets will be provided as supplementary material.

3.1 Datasets

N depends on the ASR used to generate the hypotheses. For example, suppose the beam-width is 2. In that case, the number of hypotheses available is $N = 2$, i.e., N is the property of the dataset and is not treated as a hyperparameter for the sake of this study.

Table 1: Results for SLU. Macro F1 scores are computed after 5 independent trials. We note that the difference in F1 scores between 1-best and N-best is high in our In-house dataset, followed by DSTC-2 and SNIPS-TTS.

Dataset	**Method**	**ASR F1**
DSTC-2	Gold	0.595 ± 0.022
	1Best	0.579 ± 0.023
	NBest [3]	0.602 ± 0.024
In-house	Gold	0.439 ± 0.018
	1Best	0.451 ± 0.013
	NBest [3]	0.523 ± 0.023
SNIPS-TTS	Gold	0.817 ± 0.011
	1Best	0.923 ± 0.004
	NBest [3]	0.949 ± 0.004

Table 2: Various statistics related to ASR Hypotheses of datasets used in the paper. Δ_b^a is the percentage difference in ASR F1 when using methods a and b. We note that Δ seems related to Jaccard-Index and Gestalt-PM, i.e., hypotheses diversity, and that using N-best alternatives is more beneficial when hypotheses are diverse.

Dataset	WER	Jaccard Index	Gestalt PM	Δ_{NBest}^{1Best}
SNIPS-TTS	0.443	0.639	0.882	2.82
DSTC-2	0.29	0.6	0.799	3.97
In-house	0.439	0.523	0.757	15.96

DSTC-2: We build an act-slot-pair multi-label classification dataset using the original act-slot-value DSTC-2 dataset [9]. The resulting dataset consists of 10.8K and 9.1K training and testing examples with 21 labels (act-slot pairs). We discard the "value" labels as this requires a change in the model architecture. We use $N = 5$.

SNIPS-TTS: We use the 7-class classification dataset introduced by [10]. However, few train and test set examples do not have N-Best alternatives available. We filter such examples. The train and test sets have 11K and 0.6K examples, respectively. We use $N = 9$.

In-house: This multi-class classification dataset consists of utterances collected using a deployed voice assistant (VA) in the banking domain. The VA handles queries about blocking debit/credit cards, checking bank balances, reporting fraud, etc. Human annotators label the speech audio with transcriptions, and ASR hypotheses are tagged with one of 42 intents. We perform an inner-join of these datasets, resulting in a dataset of 3.4K utterances, which we split into training and testing sets of 2.7K and 0.68K examples. We use $N = 5$.

In Figure 2, we illustrate some examples of each of the datasets used. We observe that DSTC-2 and SNIPS-TTS have highly similar ASR alternatives. In the next section, we define metrics that quantify this observation.

4 Diversity of ASR Hypotheses

In Table 1, we note that the gain in F1 scores due to using N-best hypotheses is much lower when compared with the gains in our In-house dataset. We believe this is because the hypotheses generated are less diverse in DSTC-2 and SNIPS-TTS, and $\{2, 3, \ldots, N\}$-best hypotheses do not provide much extra information to the SLU model.

We compute various statistics based on token-level and sequence-level string-similarity functions to support the above claim. For a string-similarity function $s(u, v)$ computes a scalar value that denotes the similarity between two strings u and v.

$$S = \sum_{i=1}^{|D|} \sum_{j=2}^{N} \frac{s(x_i^1, x_i^j)}{|D|(N-1)} \tag{3}$$

S is the averaged statistic across D. S denotes how much, on average, the non-1-Best hypotheses differ from the 1-Best hypothesis. It is considered a measure of diversity in ASR hypotheses.

The string-similarity functions s considered are listed below:

- *Gestalt-PM*: Gestalt Pattern Matching [11], also known as the Ratcliff-Obershelp algorithm for sequence matching. A higher score means higher similarity and vice-versa.
- *Jaccard-Index*: A token-level set-similarity measure [12]. A higher score means higher similarity and vice-versa.

Table 2 contains pair-wise string-similarity statistics between 1-Best ASR Hypothesis and others, averaged across all examples in the train and test sets. SNIPS-TTS has a higher similarity of

hypotheses, followed by DSTC-2 and In-house, across both string-similarity functions. SNIPS-TTS has the highest WER of 0.443 while having the least diversity in alternatives. The In-house dataset on the other hand, has a similar WER; however, the hypotheses are much more diverse. We emphasize that the WER of ASR seems unrelated to Δ_{NBest}^{1Best} and the diversity of hypotheses. Δ_b^a is the percentage difference in ASR F1 when using methods a and b.

In hindsight, it would seem obvious that if the hypotheses contained very similar tokens, it would not be useful in helping the SLU "guess" the correct transcription and classify the utterance correctly. However, to our knowledge, this intuition was never empirically proven to be valid.

5 Limitations

The number of datasets on which the experiments were carried out may be too few (3) to make a concrete claim on the importance of diversity. However, we argue that this is enough evidence to show that the intuition used by prior work, which relies on the WER of the 1-best hypothesis, is false.

6 Conclusion

In this work, we first benchmark N-best ASR Transformer on three different datasets with varying word-error-rates. We show that for the datasets we tried, the gain in performance when using N-best hypotheses is unrelated to WER. Instead, we show that the diversity of alternatives is an important factor. To support this, we introduce novel metrics to quantify the ASR hypotheses' diversity and show a positive correlation between the gain in performance when using N-best hypotheses and the diversity of alternatives.

References

[1] Jianfeng Gao, Michel Galley, and Lihong Li. Neural approaches to conversational ai. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[2] Dilek Z. Hakkani-Tür, Frédéric Béchet, Giuseppe Riccardi, and Gökhan Tür. Beyond asr 1-best: Using word confusion networks in spoken language understanding. *Comput. Speech Lang.*, 20:495–514, 2006.

[3] Karthik Ganesan, Pakhi Bamdev, B. Jaivarsan, Amresh Venugopal, and Abhinav Tushar. N-best asr transformer: Enhancing slu performance using multiple asr hypotheses. In *ACL*, 2021.

[4] Andreas Stolcke, Yochai Konig, and Mitch Weintraub. Explicit word error minimization in n-best list rescoring. *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 1997.

[5] Vaibhava Goel, William J. Byrne, and Sanjeev Khudanpur. Lvcsr rescoring with modified loss functions: a decision theoretic perspective. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1:425–428 vol.1, 1998.

[6] Mingda Li, Weitong Ruan, Xinyue Liu, Luca Soldaini, Wael Hamza, and Chengwei Su. Improving spoken language understanding by exploiting asr n-best hypotheses. *ArXiv*, abs/2001.05284, 2020.

[7] Faisal Ladhak, Ankur Gandhe, Markus Dreyer, Lambert Mathias, Ariya Rastrow, and Björn Hoffmeister. Latticernn: Recurrent neural networks over lattices. In *Interspeech*, pages 695–699, 2016.

[8] Chen Liu, Su Zhu, Zijian Zhao, Ruisheng Cao, Lu Chen, and Kai Yu. Jointly encoding word confusion network and dialogue context with bert for spoken language understanding. *ArXiv*, abs/2005.11640, 2020.

[9] Matthew Henderson, Blaise Thomson, and Jason D Williams. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272, 2014.

[10] Chao-Wei Huang and Yun-Nung Chen. Learning asr-robust contextualized embeddings for spoken language understanding. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8009–8013, 2020.

[11] July 01 John W. Ratclif and John W. Ratclif. Pattern matching: The gestalt approach, Jul 1988.

[12] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.