N-best Response-based Analysis of Contradiction-awareness in Neural Response Generation Models

Shiki Sato[1], Reina Akama[1,3], Hiroki Ouchi[2,3], Ryoko Tokuhisa[1], Jun Suzuki[1,3], Kentaro Inui[1,3]

[1] Tohoku University [2] Nara Institute of Science and Technology [3] RIKEN

Overview

- We analyze contextual **contradiction-awareness** of response generation models focusing on **consistency of n-best candidates**
- **Beam search has limitation** on avoiding contradiction and **unlikelihood training reduce** it

Background: Avoiding contradiction with contradiction detector

Context
Aren't you in South Carolina?
I'm in North Carolina.

Our focus / Previous focus

n-best
1st *South.* — contradictory
2nd *Yes.* — contradictory
3rd *Not really.* — noncontradictory

Contradiction detector

All candidates affect whether final output is contradictory

→ Need to observe **all candidates' consistency**

Proposed framework

Entailing NLI data
Premise: *I'm in North Carolina.*
Hypothesis: *I'm in North Carolina.*

EntQ: Should be answered *Yes.*
I'm in North Carolina.
Are you in North Carolina?

Contradicting NLI data
Premise: *I'm in North Carolina.*
Hypothesis: *I'm in South Carolina.*

CntQ: Should be answered *No.*
I'm in North Carolina.
Aren't you in South Carolina?

Yes-no questions allow for **clear** judge of responses

1. Prepare stimulus inputs

I'm in North Carolina.
Aren't you in South Carolina?

2. Detect noncontradictory candidate

System A: 1st *South.* 2nd *Yes.* 3rd *Not really.*

System B: 1st *I'm not.* 2nd *No!* 3rd *North*

yes-no classifier — Categorize into *yes*/*no*

All n-bests should have ✔ to always avoid contradiction

$$\textbf{Certainty} = \frac{\text{Number of } n-\text{bests having at least one ✔}}{\text{Number of } n-\text{bests}}$$

✔'s proportion in n-best should be **high** to provide many available candidates — Leads to better output

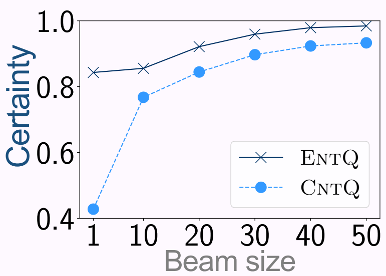
$$\textbf{Variety} = \text{Average}\left(\frac{\text{Number of ✔ in the } n-\text{best}}{\text{Size of } n-\text{best having at least one ✔}}\right)$$

3. Compute scores

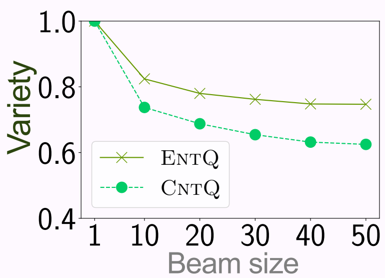
	System A	System B
Certainty	3/3= 1.00	1/3= 0.33
Variety	(.33+.33+.33)/3= 0.33	(1.00)/1= 1.00

Experiments: Proposed framework reveals n-best's properties

Using **beam search**,



Increases as beam size increases



Decreases as beam size increases

↔ **trade-off**

Using **newer techniques**,

Technique	Certainty EntQ	Certainty CntQ	Variety EntQ	Variety CntQ
Beam search	.856	.768	.824	.737
Diverse beam search [Vijayakumar+'16]	.999	.981	.758	.478
Nucleus sampling [Holtzman+'20]	1.00	.994	.755	.462
Unlikelihood training [Li+'20]	**.910**	**.937**	**.969**	**.968**

Unlikelihood training improves both scores

Settings
Generation model: Blender 3B [Roller+'21] (See our paper for results of DialoGPT [Zhang+'20])
Inputs: 2,000 EntQ/CntQ from Multi-Genre NLI [Williams+'18], Yes-no classifier: RoBERTa [Liu+'19] fine-tuned on Circa [Louis+'20]