Evaluating Dialogue Generation Systems via Response Selection

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

[1]Tohoku University [2]RIKEN

{shiki.sato,reina.a,jun.Suzuki,inui}@ecei.tohoku.ac.jp
hiroki.ouchi@riken.jp

Our test set available at https://github.com/cl-tohoku/eval-via-selection

Overview

Motivation

Comparing the performance of Dialogue Generation Systems (DGS):

- With a high correlation with humans
- At low cost

Approach

Response Selection (RS) with **well-chosen false candidates**

Context	Candidates
How is he?	**A. He is fine.**
	B. She is fine
	C. Is he fine?

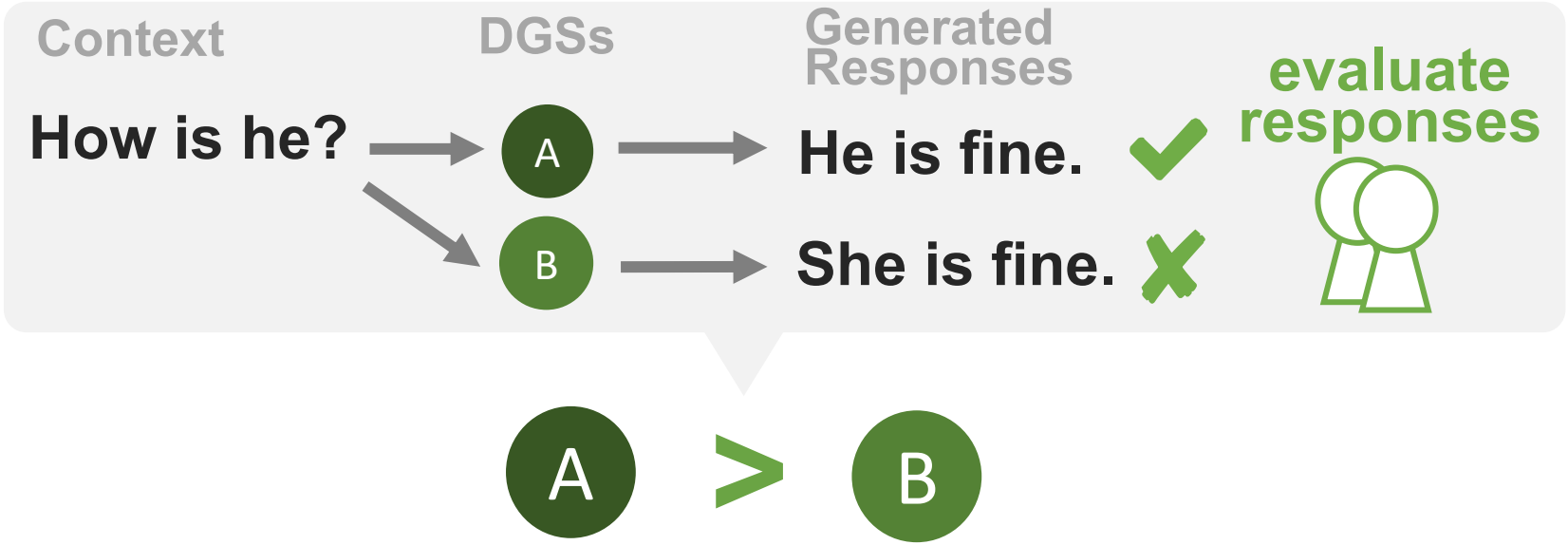
similar to GOLD but unacceptable

Contributions

1. Development of a RS test set with well-chosen candidates
2. Our comparison method **correlates with human judgements**

Necessity of Automatic Evaluation

DGSs can be compared by **human evaluation**



Cons

Takes high cost

➡ Cannot evaluate a lot of DGSs

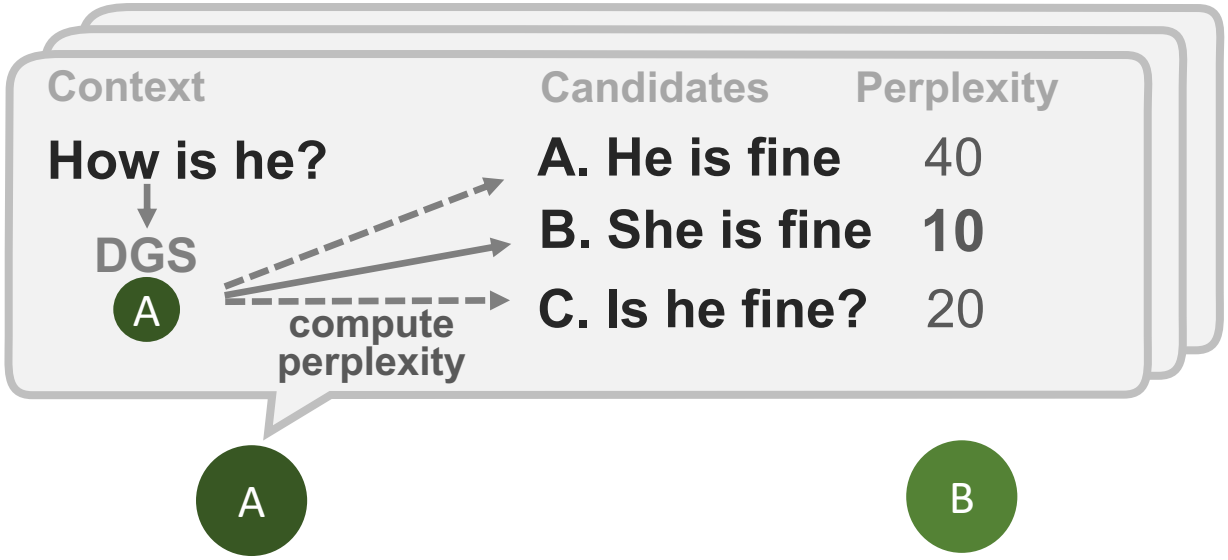
Necessity of Automatic Evaluation

Automatic evaluation drives research as preliminary evaluation

Comparing DGSs via RS

Method

We focus on comparing DGSs by **RS accuracy**



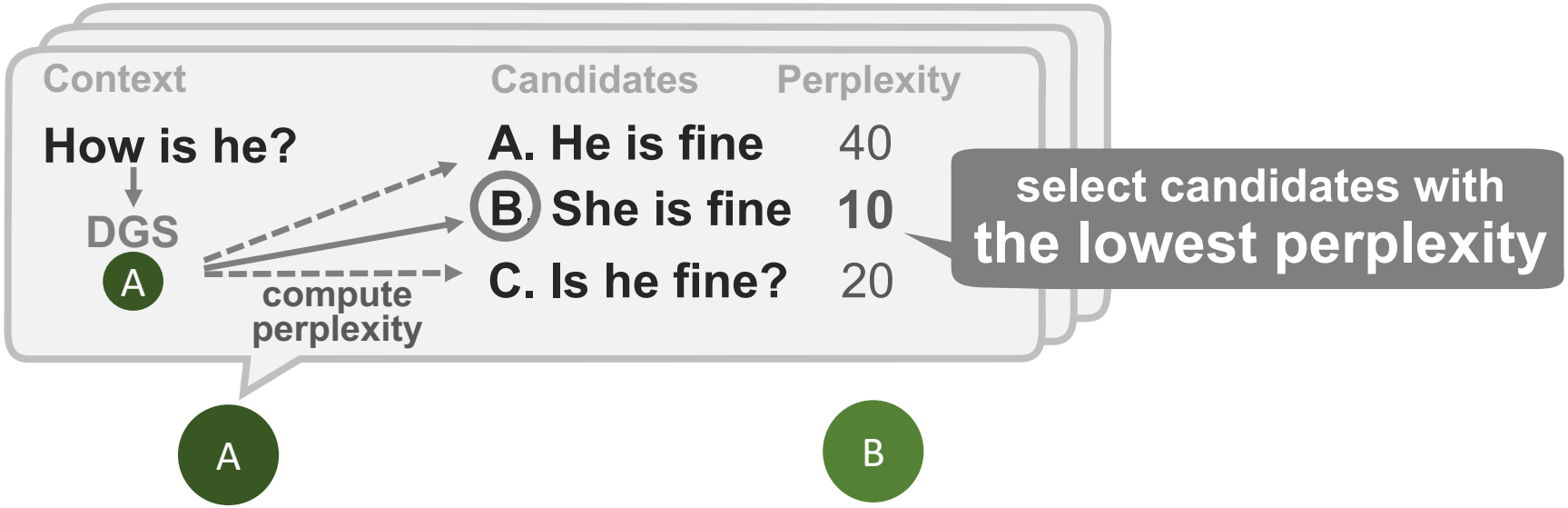
Pros

RS can remedy **one-to-many** problem

Comparing DGSs via RS

Method

We focus on comparing DGSs by **RS accuracy**



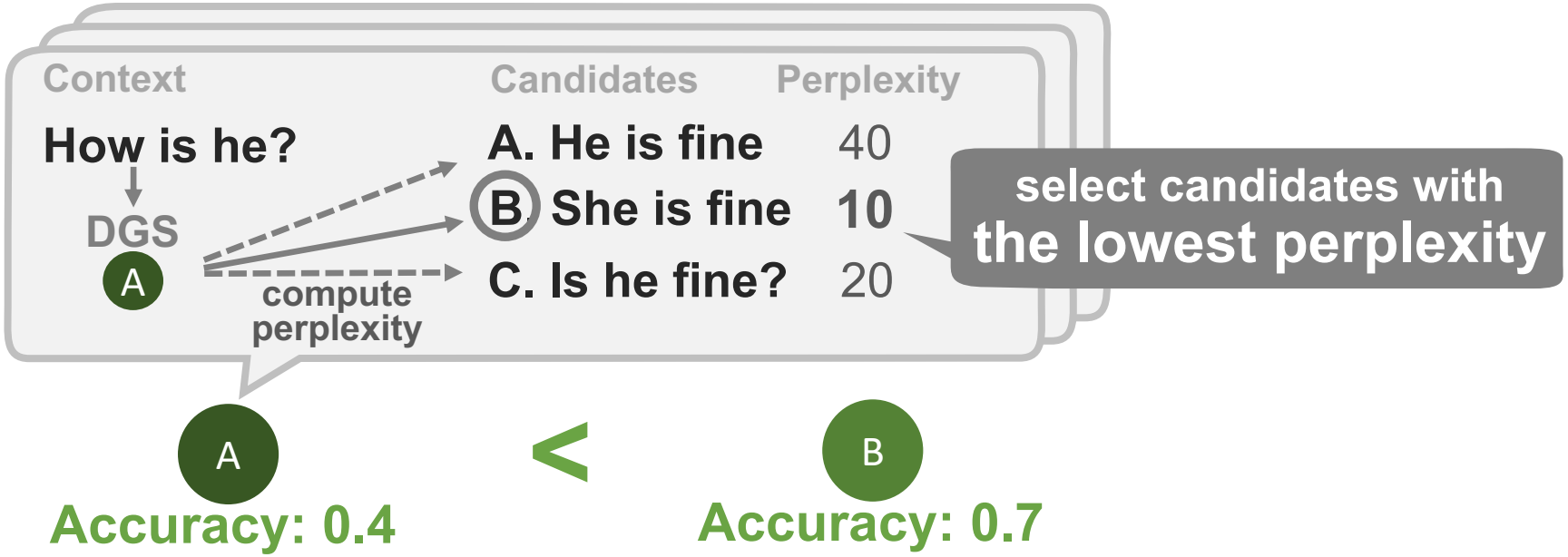
Pros

RS can remedy **one-to-many** problem

Comparing DGSs via RS

Method

We focus on comparing DGSs by **RS accuracy**



Pros

RS can remedy **one-to-many** problem

Problems on RS false candidates

Undesirable false candidates are sampled by **random-sampling**

Context	Candidates
How is he?	A. He is fine.
randomly sampled	B. That is a car
	C. I don't know

Problems on RS false candidates

Undesirable false candidates are sampled by **random-sampling**

Context	Candidates
How is he?	A. He is fine.
randomly sampled	B. That is a car
	C. I don't know

(i) Containing unrelated words

Problems on RS false candidates

Undesirable false candidates are sampled by **random-sampling**

Context	Candidates	
How is he?	A. He is fine.	
randomly sampled	B. That is a car	(i) Containing unrelated words
	C. I don't know	(ii) Acceptable response

Contributions

1. Development of a RS test set with **well-chosen** false candidates

   "**well-chosen**" ?

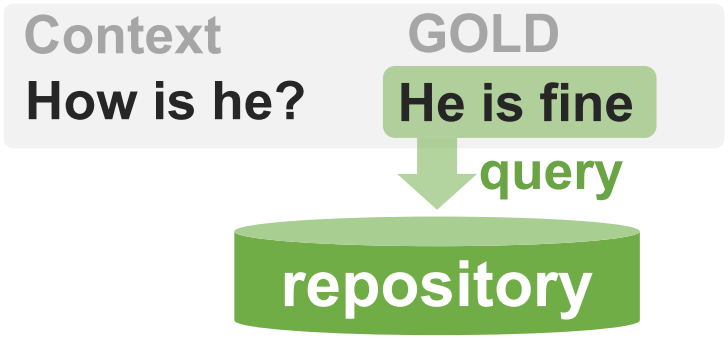
   - **Similar to the GOLD**
   - **But unacceptable**

2. Our comparison method correlates with human judgements

Test Set Construction Method

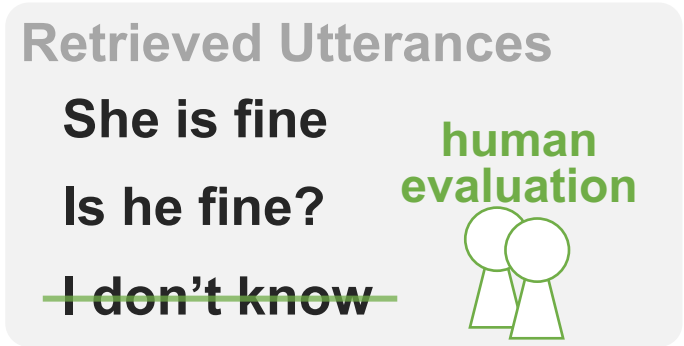
Collect false candidates in two steps:

1. Retrieve utterances **similar to GOLD**



Collect hardly distinguishable candidates

2. Filter out **acceptable utterances**



Retain only **unacceptable** candidates

Example of Our RS Test Set

Context

A: Excuse me. Could you please take a picture of us with this *camera*?

B: Sure. Which button do I press to shoot?

A: This one.

Chosen Candidates

1. Do I have to *focus* it?
2. But I do have ninja *focus*.
3. Do not lose your *focus*!
4. Could he not *focus* on that?

Containing "***focus***" related to "***camera***"

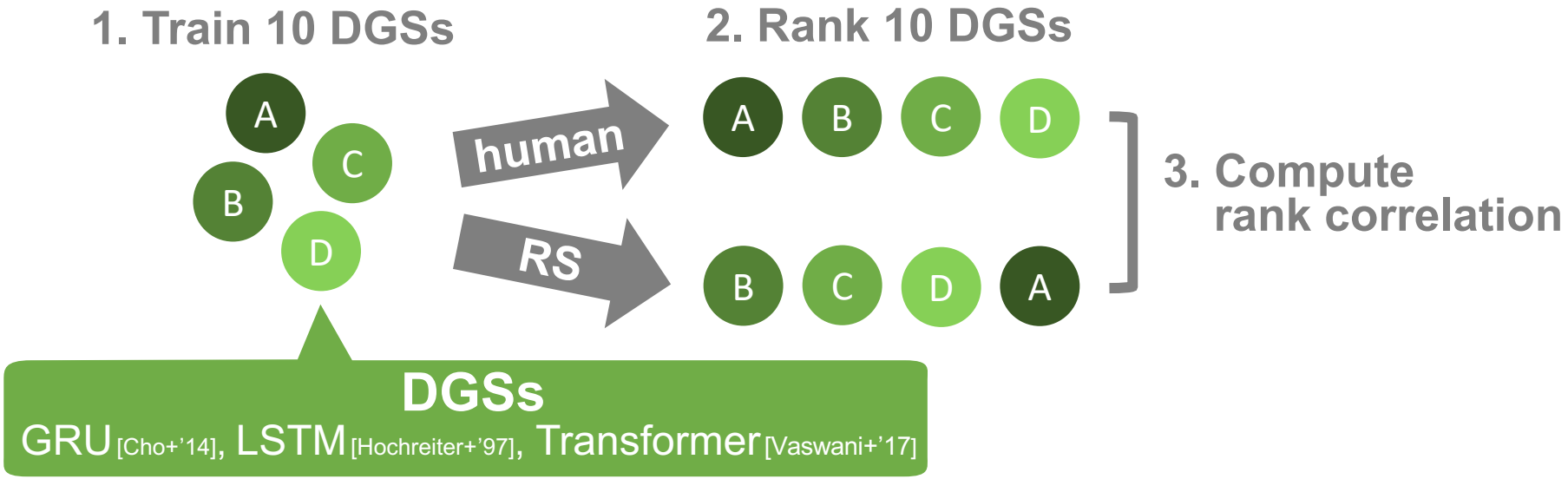
1,019 questions available at:
https://github.com/cl-tohoku/eval-via-selection

Experiments

Can our method compare DGSs like humans?

Experimental Procedure

Compute the similarities between [system ranking by **humans**
system ranking by **RS**



Results: Correlation between DGS Rankings

Correlation with ranking of DGSs by humans

Metrics	Spearman	p-value
BLEU-1	-0.36	0.30
BLEU-2	0.085	0.82
METEOR	0.073	0.84
ROUGE-L	0.35	0.33
Our Method	0.48	0.19

word overwrap-based metrics

Results: Correlation between DGS Rankings

Correlation with ranking of DGSs by humans

Metrics	Spearman	p-value
BLEU-1	-0.36	0.30
BLEU-2	0.085	0.82
METEOR	0.073	0.84
ROUGE-L	0.35	0.33
Our Method	**0.48**	**0.19**

word overwrap-based metrics

Correlates with humans more strongly than word overwrap-based metrics

Conclusion

Motivation

Comparing the performance of Dialogue Generation Systems (DGS):

- With high correlation with humans
- At low cost

Approach

Response Selection with well-chosen false candidates

Results

Our comparison method **correlates with human judgements**

Thank You for Listening!


shiki.sato@ecei.tohoku.ac.jp

Our test set available at https://github.com/cl-tohoku/eval-via-selection