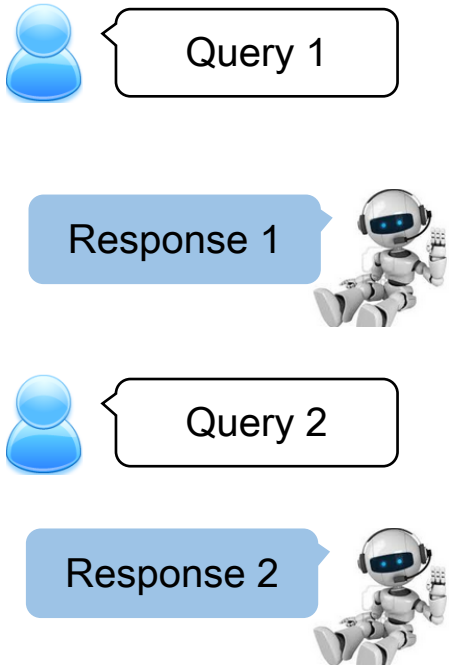
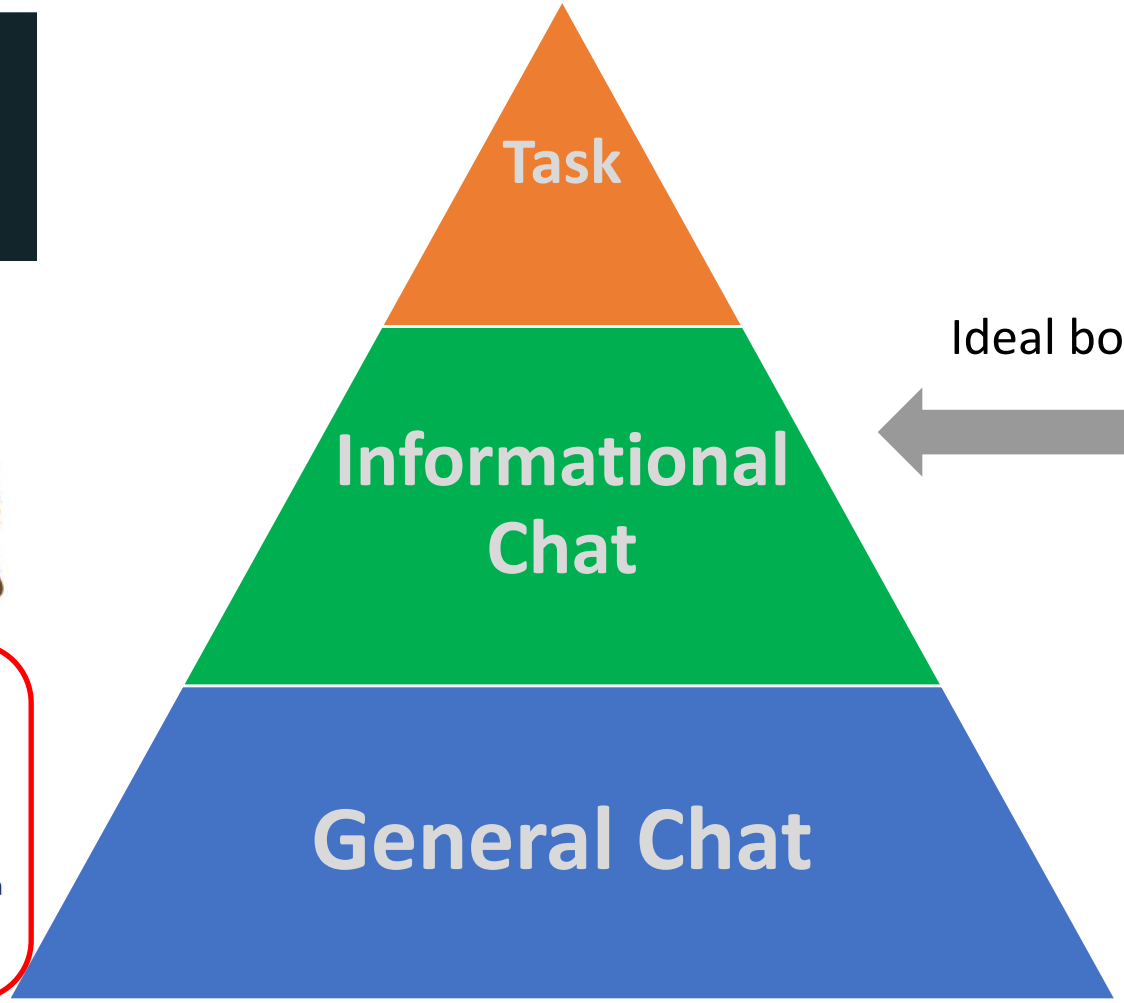
Response Editing

Yu Wu

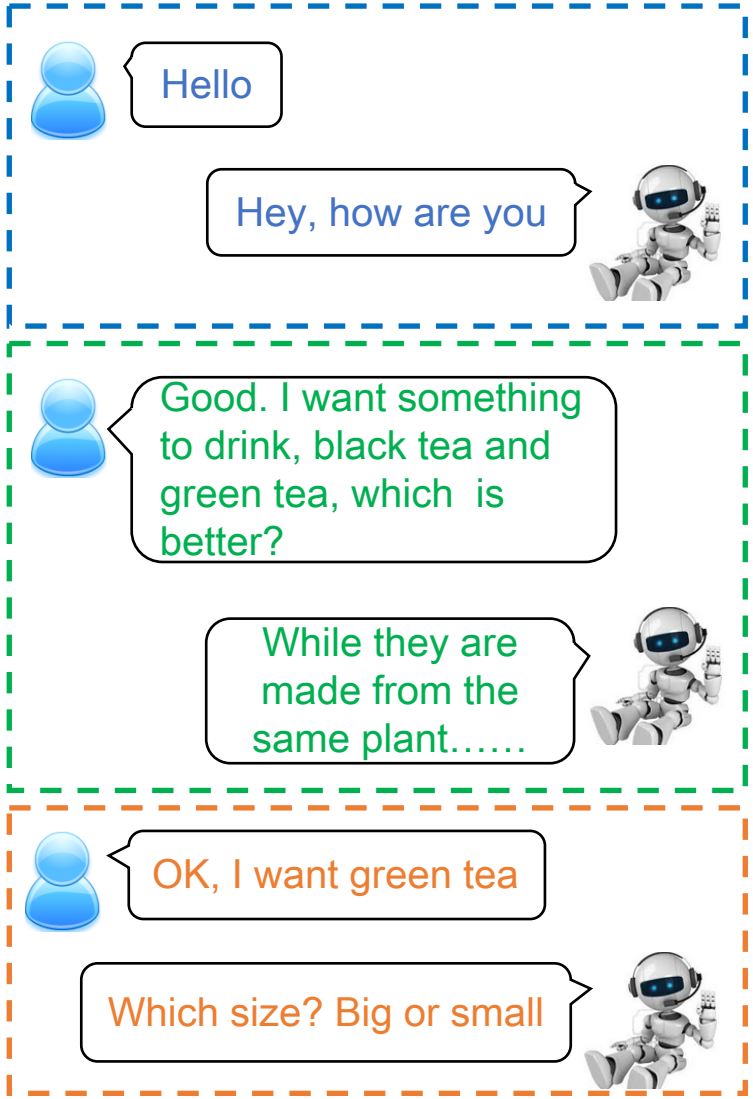
An Era of Conversational Agents/Bots



An Ideal bot



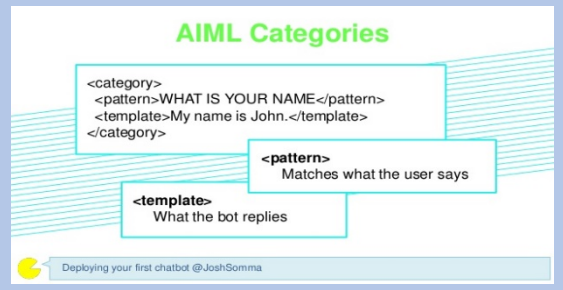
Ideal bot



Three typical approaches

Template based

- Respond users with hand-crafted features.



- Reliable and controllable
- Hard to scalable

Retrieval based

- Select a proper response from an index

- Fluent and informative responses
- Easy to implement
- Heavily depend on a pre-defined index

Generation based

- Generate a response with NLG techniques

- Flexible
- Suffer from safe response problem
- Require more resource

Ensemble Approaches

- Input:
  - User issued query
  - Some retrieved query-response pairs.
- Model:
  - Generate a response with the query and pairs.
- Output:
  - An appropriate response.

Previous ensemble approaches

- **Multiple encoders.**
- **Copy words in prototypes.**
- **Still use query as a source language**

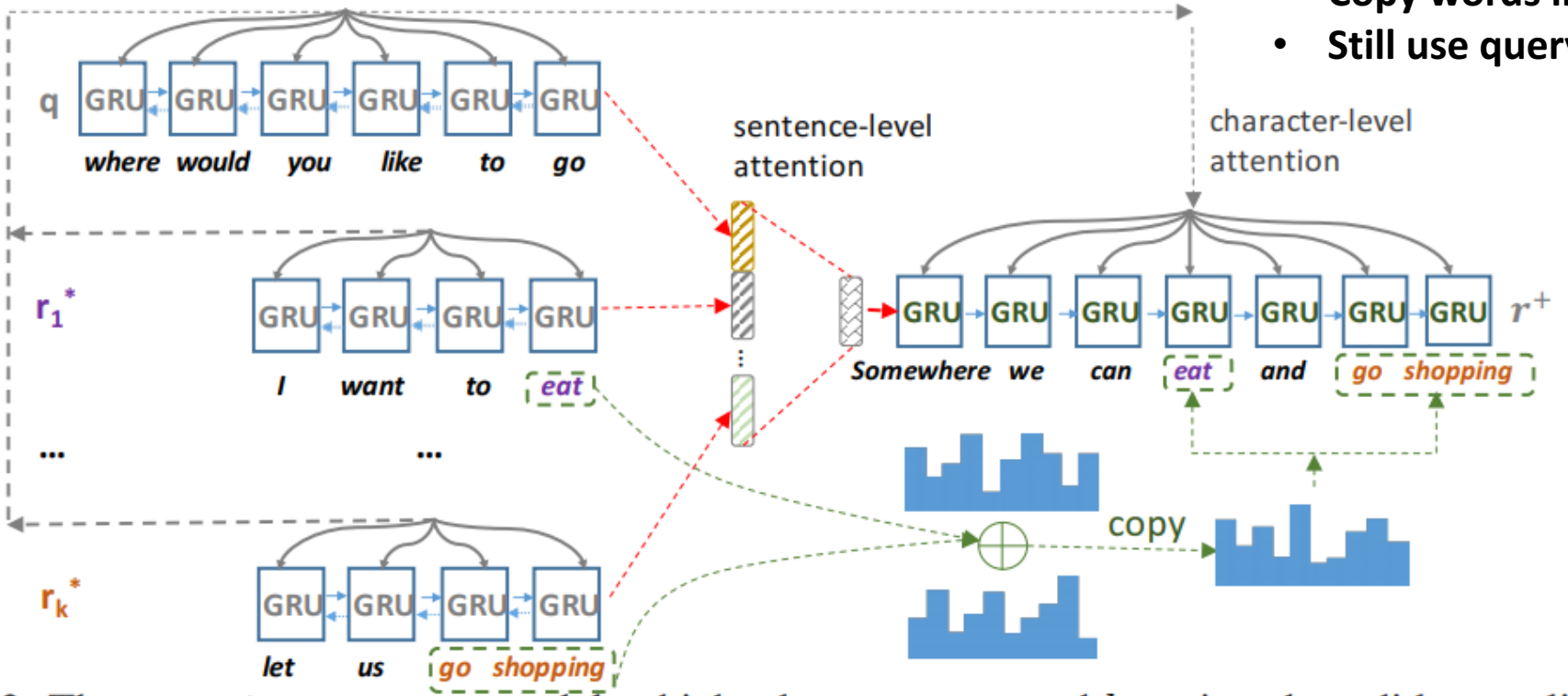


Figure 2: The `multi-seq2seq` model, which takes a query q and k retrieved candidate replies r^* as the input and generate a new reply r^+ as the output.

AN ENSEMBLE OF RETRIEVAL-BASED AND GENERATION-BASED HUMAN-COMPUTER CONVERSATION SYSTEMS. Song et al. IJCAI 2018

Previous ensemble approaches

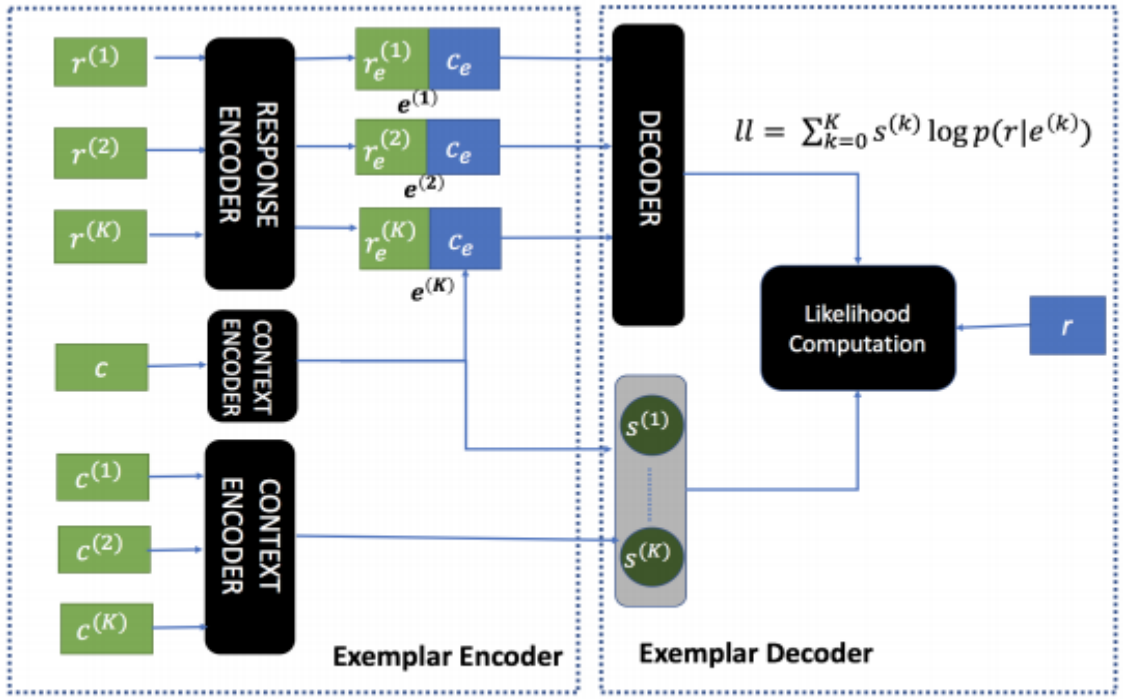


Figure 1: A schematic illustration of the EED network. The input context-response pair is (c, r), while the exemplar context-response pairs are $(c^{(k)}, r^{(k)})$, $1 \leq k \leq K$.

- **Multiple encoders.**
- **Weight prototypes with a matching model.**
- **Still use query as a source language**

Exemplar Encoder-Decoder for Neural Conversation Generation. Pandey et al. ACL 2018

Motivation

- Prototype controls rough semantic.
- Context difference determines how to revise the response.
- Ensemble retrieval and generation approaches.

Context	My friends and I ~~went to some~~ vegan place for ~~dessert~~ yesterday.
Prototype context	My friends and I had Tofu and vegetables at a vegan place nearby yesterday.
Prototype response	**Raw green vegetables** are very **beneficial** for your health.
Revised response	**Desserts** are very **bad** for your health.

Table 1: An example of context-aware prototypes editing. Underlined words mean they do not appear in the original context, while ~~words with strikethrough~~ mean they are not in the prototype context. Words in bold represent they are modified in the revised response.

Advantages

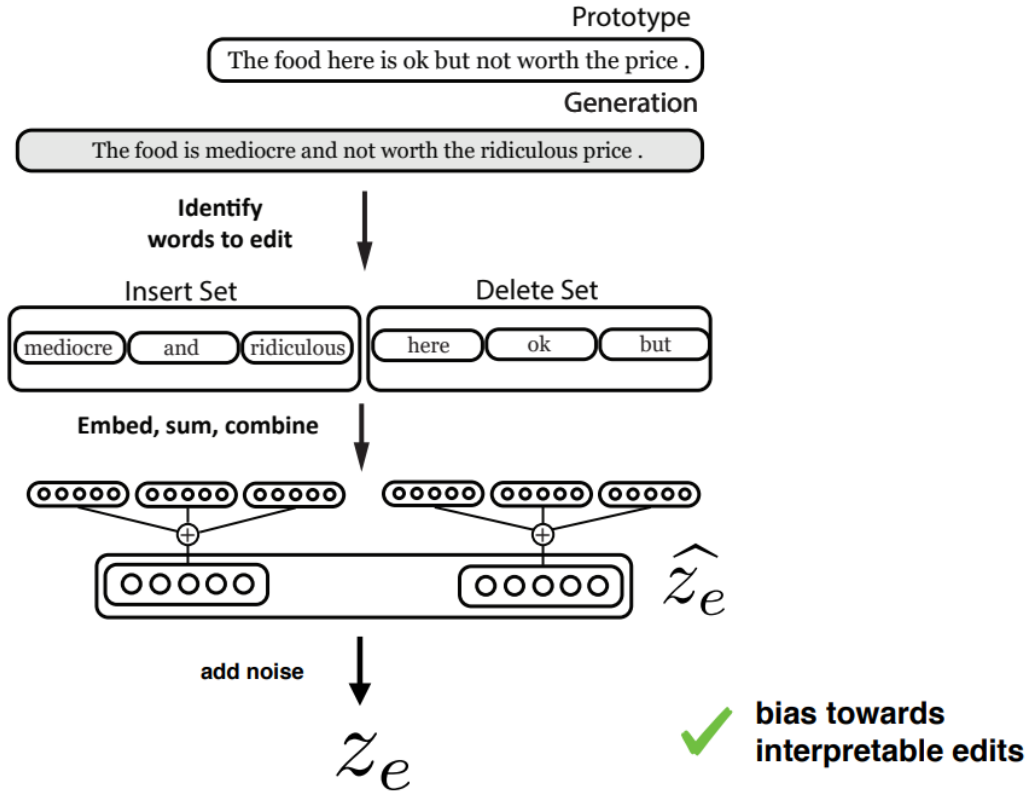
- Prototype response provides a good start-point for our editing model.
  - Informative and fluent.
- We regard prototype response and revised response as a source language and a target language respectively.
  - Easy to learn alignment

Context	My friends and I ~~went to some~~ vegan place for ~~dessert~~ yesterday.
Prototype context	My friends and I had Tofu and vegetables at a vegan place nearby yesterday.
Prototype response	**Raw green vegetables** are very **beneficial** for your health.
Revised response	**Desserts** are very **bad** for your health.

Table 1: An example of context-aware prototypes editing. Underlined words mean they do not appear in the original context, while ~~words with strikethrough~~ mean they are not in the prototype context. Words in bold represent they are modified in the revised response.

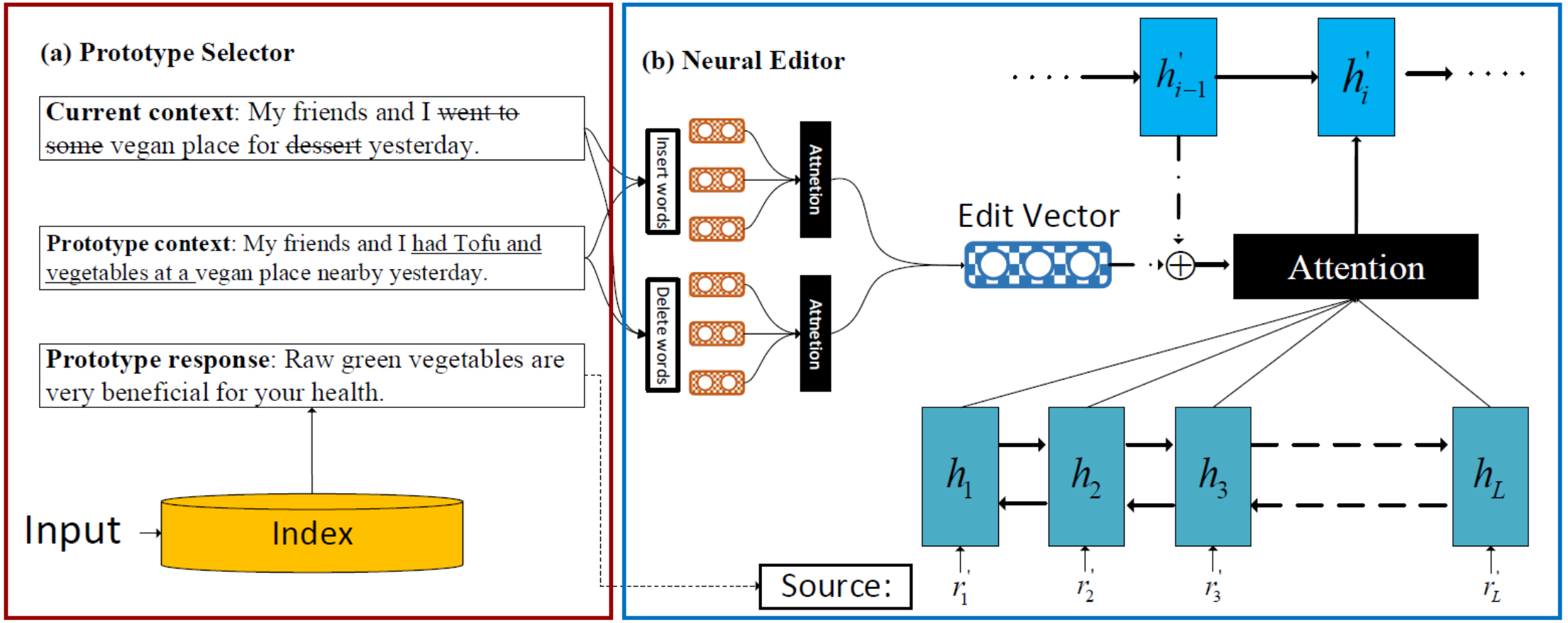
Background

- Build a training set of lexically similar sentence pairs (x,x')
- For each pair (x,x')
  - Identify words that differ between x and x'.
  - Embed those into a vector.
  - Add noise to get edit vector z.
  - Train s2s mapping (z,x) -> y

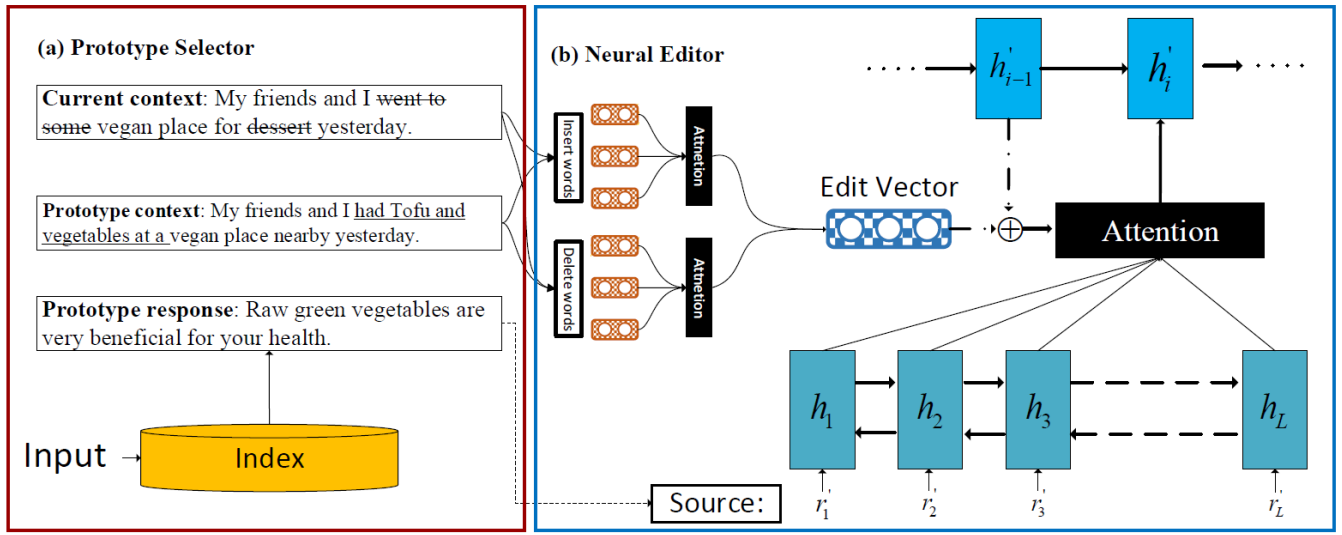


Generating Sentences by Prototype Editing. Guu et al. 2018 TACL

Model Overview



Model Overview



$Edit_{vec} = [I; D]$ (Concatenate two vectors)

Append the edit vector to each hidden vector to compute the output word.

$$p(output) = f(edit\_vec, hidden)$$

Insertion vector: $I = \sum_{w_i \in Ins} \alpha_i \cdot w_i$, where w_i is a word embedding and α_i is its corresponding weight.

Deletion vector: $D = \sum_{w_i \in Del} \beta_i \cdot w'_i$, where w'_i is a word embedding and β_i is its corresponding weight.

Dataset: Douban Corpus

- Training data: Given (C,R), search similar (C',R') with response similarity.
  - $Jaccard(R, R') \in (0.3, 0.7)$
- Test data: Given C, search similar (C',R') with context similarity.

	train	val	test
context-response pairs for s2s	20M	10k	10k
context-response pairs for editing	20M (Samle from 40M)	10k	10k
Vocabulary	30000		

Baseline Methods

- S2SA: the standard S2S model with an attention mechanism. We use the implementation with Blocks https://github.com/mila-udem/blocks
- S2SA-MMI: the model proposed by Li et al. (Li et al.2015). We implement this baseline by the code published by the authors at
- CVAE: recent work for response generation with a conditional variational auto-encoder (Zhao, Zhao, and Esḱenazi).
- Retrieval models: Retrieval-default uses results given by Lucene. Retrieval-rerank ranks top 20 results given by Lucene with LSTM.
- Ensemble models: It has multiple encoders to encode prototype response information, and generate the response with the concatenation result of context representation and prototype representation.

Automatic Evaluation

Table 2: Automatic evaluation results. Numbers in bold mean that improvement from the model on that metric is statistically significant over the baseline methods (t-test, p-value < 0.01). κ denotes Fleiss Kappa (Fleiss 1971), which reflects the agreement among human annotators.

	Relevance			**Diversity**		**Originality**	**Fluency**				
	Average	Extrema	Greedy	Distinct-1	Distinct-2	Not appear	+2	+1	0	Avg	κ
S2SA	0.346	0.180	0.350	0.032	0.087	0.208	94.0%	5.2%	0.8%	1.932	0.89
S2SA-MMI	0.379	0.189	0.385	0.039	0.127	0.297	91.6%	7.6%	0.8%	1.908	0.83
CVAE	0.360	0.183	0.363	0.062	0.178	0.745	83.8%	12.7%	3.5%	1.803	0.84
Retrieval-default	0.288	0.130	0.309	**0.098**	**0.549**	0.000	92.8%	6.8%	0.4%	1.924	0.88
Retrieval-Rerank	0.380	0.191	0.381	0.067	0.460	0.000	91.7%	7.8%	0.5%	1.912	0.88
Ensemble-default	0.352	0.183	0.362	0.035	0.097	0.223	91.3%	7.2%	1.5%	1.898	0.84
Ensemble-Rerank	0.372	0.187	0.379	0.040	0.135	0.275	91.3%	7.2%	1.5%	1.898	0.84
Edit-default	0.297	0.150	0.327	0.071	0.300	0.796	89.6%	9.2%	1.2%	1.884	0.87
Edit-1-Rerank	0.367	0.185	0.371	0.077	0.296	0.794	93.2%	5.6%	1.2%	1.920	0.79
Edit-N-Rerank	**0.386**	**0.203**	**0.389**	0.068	0.280	**0.860**	87.2%	10.8%	2.0%	1.852	0.85

- We evaluate models from relevance, diversity, originality and fluency.
- Our model shows good results on relevance and originaliy

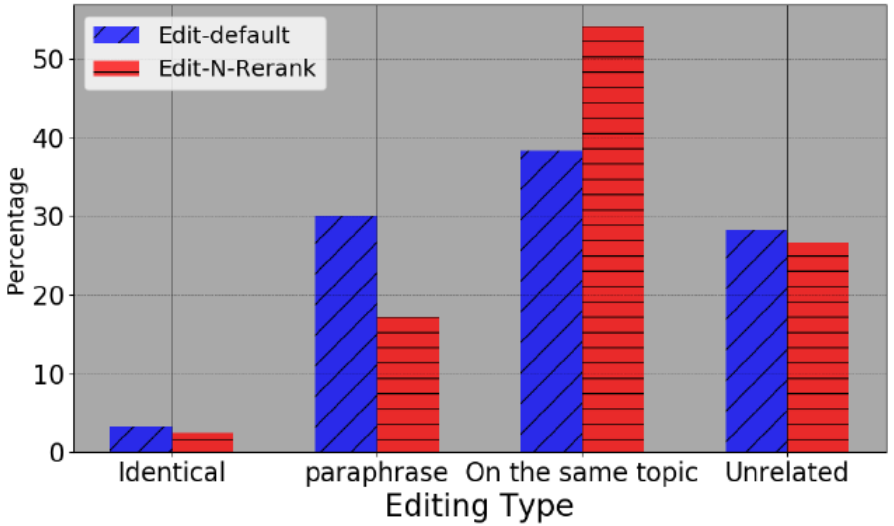
Human side by side evaluation

Table 3: Human side-by-side evaluation results. Fleiss Kappa is denoted as κ. If a row name is "a v.s.b", win means that "a" is better than "b" according to human evaluation.

	Loss	Tie	Win	κ
Ed-Default v.s. R-Default	23.6	41.2	35.2	0.54
Ed-Default v.s. Ens-Default	29.4	47.8	22.8	0.59
Ed-1-Reran v.s. R-Rerank	29.2	45.3	25.5	0.60
Ed-N-Reran v.s. Ens-Rerank	24.9	42.1	33.0	0.55
Ed-N-Reran v.s. R-Rerank	25.2	44.8	30.0	0.62

Editing analysis

- Identical: our generation result is identical with the prototype response .
- Paraphrase: our generation result is a paraphrase of the prototype response.
- On the same topic: our generation results talk about the same topic with the prototype.
- Unrelated: : our generation result and prototype response are completely unrelated.



Case 1:
Context: 身 在 国外，寂寞 无聊 就 化妆 // I am abroad. If I feel bored, I try some makeup.

Prototype Context: 无聊 就 玩 // If you feel bored, go to play.
Prototype Response: 嗯 // Well..
Revised Response: 我 也 喜欢 化妆// I love make up as well .

Case 2:
Context: 我 比较 常 吃 辛拉面 // I often eat spice noodles.

Prototype Context: 我 在 台湾 时候 常 吃 辛拉面 和 卤肉饭 // When I lived at Taiwan, I often eat spicy Noodles and braised pork rice..
Prototype Context: 我 也 喜欢 卤肉饭 // I love braised pork rice as well. ..
Revised Response: 我也喜欢 // I love it as well. (In Chinese, model just deletes the phrase ``braised pork rice" without adding any word.)

Case 3:
Context: 纹身 有 没有 办法 全部 弄 干净 // Is there any way to get all tattoos clean?

Prototype Context: 把 药 抹 头发 上 能 把 头发 弄 干净 么// Is it possible to clean your hair by wiping the medicine on your hair?.
Prototype Context: 抹 完 真的 头发 干净 很多// After wiping it, hair gets clean
Revised Response: 抹 完 纹身 就 会 掉 很多// After wiping it, most of tattoos will be cleaned.

Summary and Future work

- Summary
  - This paper proposes a new **paradigm**, prototype-then-edit, for response generation.
  - We elaborate a simple but effective **context-aware editing model** for response generation.

- Future work
  - More powerful context-aware editing models.

We are hiring

- Position
  - Research intern of MSRA NLC Group
- Research Interest
  - Machine Reading Comprehension
  - Machine Translation
  - Semantic Parser
  - Fundamental NLP
- Requirement:
  - At least 6 months internship