Improving Neural Conversational Models with Entropy-Based Data Filtering

Richard Csaky[1], Patrik Purgai[1], Gabor Recski[1,2]
[1]Department of Automation and Applied Informatics, Budapest University of Technology and Economics
[2]Apollo.AI

BACKGROUND

- In dialog data targets to the same input vary semantically (*one-to-many*) [Wei et al., 2017].
- Generic responses that appear in a diverse set of contexts (*many-to-one*) [Wu et al., 2018].

Previous approaches to these issues:

- Feeding **extra information** to dialog models [Li et al., 2016b].
- **Augmenting** the **model** or decoding process [Shao et al., 2017].
- Modifying the **loss** function [Li et al., 2016a].

METHODS

`IDENTITY` **approach:**

- **Filter utterances** from datasets in the *one-to-many*, *many-to-one* categories.
- Remove **high entropy** utterances (paired with **diverse sources/targets**), based on the conditional probabilities of utterance pairs in the data (Figure 1).
- 3 filtering ways: `SOURCE` (utterance pairs with a high entropy source), `TARGET` (pairs with a high entropy target), `BOTH` (union of `SOURCE` and `TARGET`).

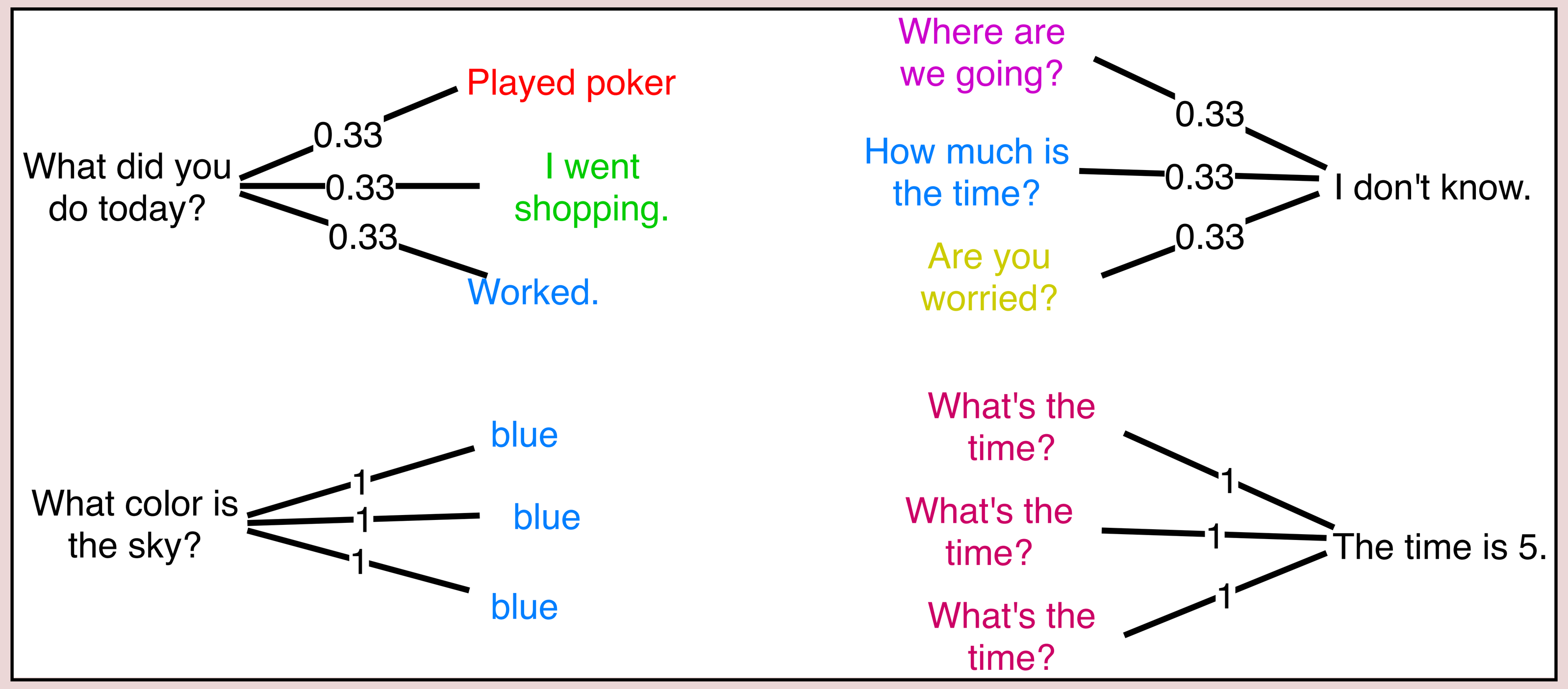

Figure 1: A high/low entropy (top/bottom) source utterance (left) and response (right). Numbers represent conditional probabilities.

`SENT2VEC` **and** `AVG-EMBEDDING` **approach:**

- **Cluster utterances** with Mean Shift [Fukunaga and Hostetler, 1975]. Sentence representations: `SENT2VEC` [Pagliardini et al., 2018], `AVG-EMBEDDING` [Arora et al., 2017].
- **Entropy** at the **cluster level**, filtering clusters instead of individual utterances.
- A **high entropy** cluster groups similar utterances paired with **diverse sources/targets** (Figure 2).

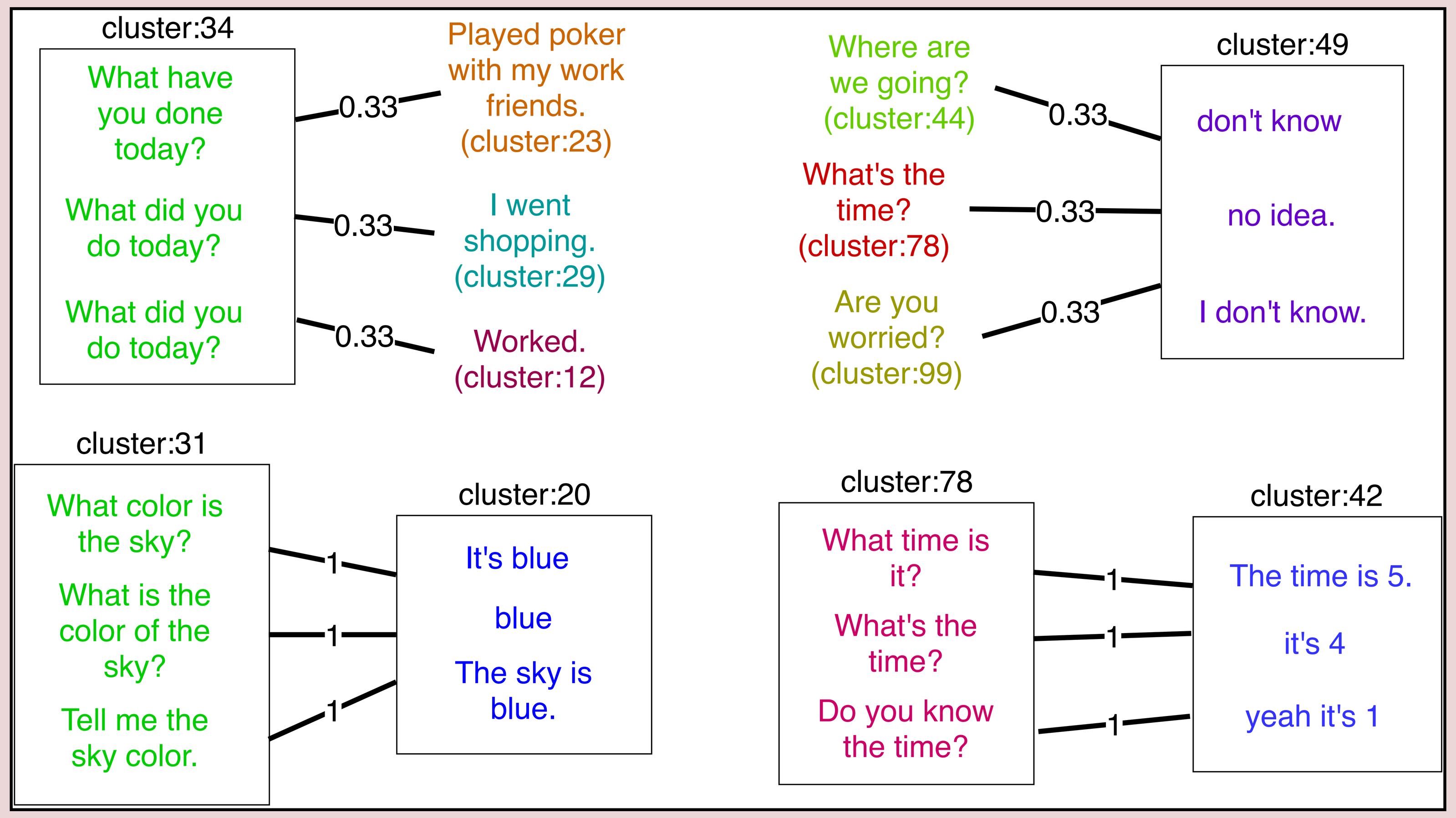

Figure 2: A high/low entropy (top/bottom) source cluster (left) and target cluster (right). Numbers represent conditional probabilities.

Filtering generic **utterances** from data using **entropy-based** methods improves **response quality.**

17 metrics on **3 datasets.**

Overfitting on cross-entropy **loss**

=

Better on automatic **metrics**

Paper: `arxiv.org/abs/1905.05471`

Filtering Code:

`github.com/ricsinaruto/Seq2seqChatbots`

Evaluation Code:

`github.com/ricsinaruto/dialog-eval`

METRICS

- **Length**: Number of words in the response.
- **Entropy**: Per-word, per-bigram and utterance entropy of responses [Serban et al., 2017]. We also introduce the KL divergence between model and ground truth response sets.
- **Embedding**: Embedding *average*, *extrema*, *greedy* metrics measuring the similarity between response and target word embeddings [Liu et al., 2016].
- **Coherence**: Similarity between input and response word embeddings [Xu et al., 2018].
- **Distinct**: *Distinct-1* and *distinct-2* measure the ratio of unique unigrams/bigrams to the total number of unigrams/bigrams in a set of responses [Li et al., 2016a].
- **BLEU**: N-gram overlap between response and target [Papineni et al., 2002].

Experimental setup:

- **Model**: `transformer` [Vaswani et al., 2017].
- **Dataset**: DailyDialog [Li et al., 2017]. Evaluations on Twitter and Cornell data in the paper.
- **Data filtered**: 5-15% depending on filtering method.
- **Decoding**: Greedy, better than beam search on all metrics [Tandon et al., 2017].

- Many automatic **metrics** correlate badly with **human judgment** [Liu et al., 2016].
- Responses at the validation **loss minimum** are often qualitatively **worse than** after **overfitting** [Csaky, 2019, Tandon et al., 2017].
- We observed that all **metrics** perform much **better after** the model **overfitted** according to the loss function (Figure 3). **Metrics saturate** and don't decrease even after **640 epochs**.

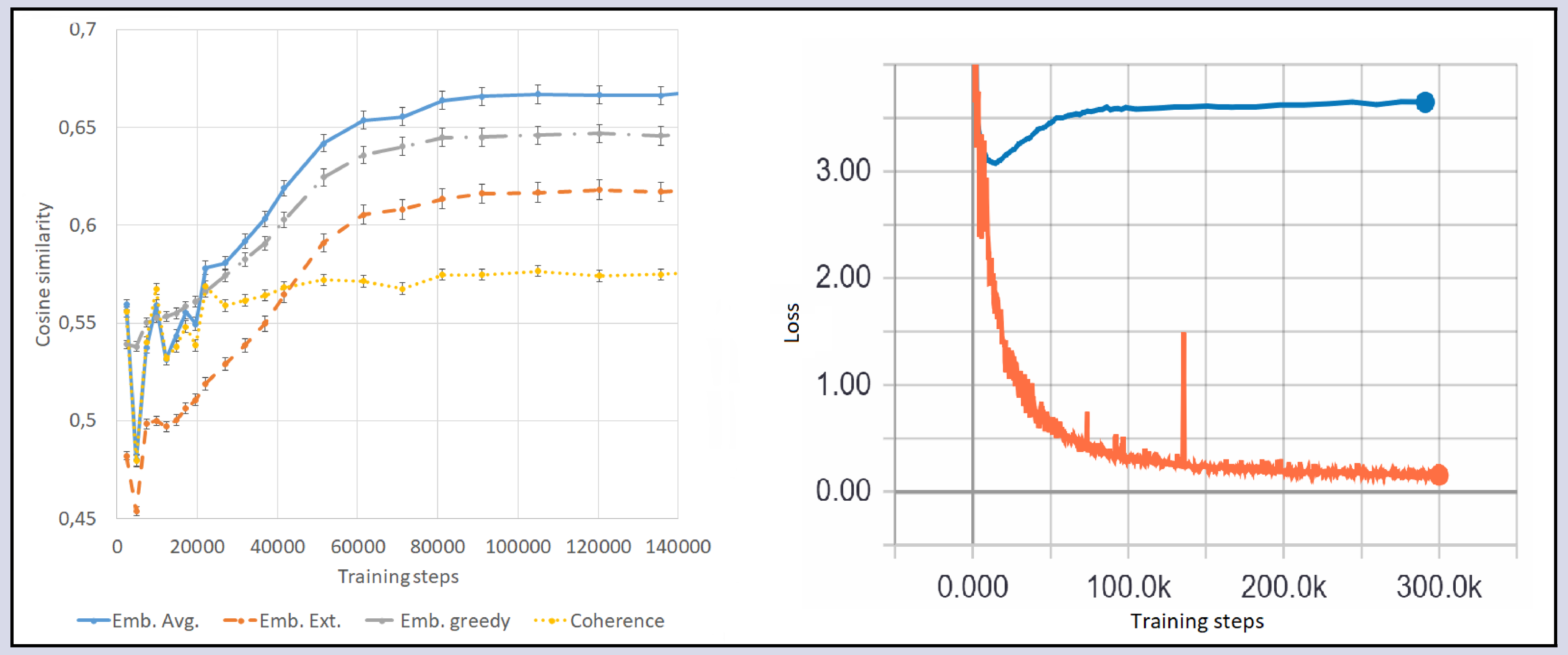

Figure 3: Embedding metrics and coherence on validation data (left) and training and validation loss (right) as a function of the training evolution of `transformer` on unfiltered data.

EXPERIMENTS

- Metrics on the unfiltered test set after 150 epochs of training.
- `TRF` = baseline `transformer`, **ID** = `IDENTITY`, **AE** = `AVG-EMBEDDING`, **SC** = `SENT2VEC`.
- `SOURCE`, `TARGET`, `BOTH` filtering denoted by initials.
- **GT** = ground truth responses, **RT** = random responses from the training set.
- The 17 metrics from left to right: response length, unigram and bigram entropy, unigram and bigram utterance entropy, unigram and bigram KL divergence, embedding *average*, *extrema* *greedy*, coherence, *distinct-1* and *distinct-2*, BLEU-1, BLEU-2, BLEU-3, BLEU-4.

		$\lvert U\rvert$	H^u_w	H^b_w	H^u_u	H^b_u	D^u_{kl}	D^b_{kl}	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		11.5	7.98	**13.4**	95	142	.0360	.182	.655	**.607**	**.640**	.567	**.0465**	**.297**	**.333**	.333	.328	.315
ID	B	**13.1**	**8.08**	**13.6**	**107**	**162**	.0473	.210	**.668**	**.608**	**.638**	**.598**	.0410	.275	**.334**	**.340**	**.339**	**.328**
	T	12.2	8.04	**13.6**	100	150	**.0335**	**.181**	**.665**	**.610**	**.640**	.589	.0438	.289	**.338**	**.341**	**.339**	**.328**
	S	12.3	7.99	**13.5**	101	153	.0406	.187	.662	**.610**	**.641**	.578	.0444	.286	**.339**	**.342**	**.338**	**.326**
AE	B	11.9	7.98	**13.5**	98	147	.0395	.197	.649	.600	.628	.574	.0434	.286	.318	.321	.318	.306
	T	*12.5*	7.99	**13.5**	*102*	*155*	.0436	.204	.656	.602	.634	.580	.0423	.279	.324	.327	.325	.313
	S	12.1	7.93	**13.4**	99	148	.0368	.186	.658	.605	**.636**	.578	.0425	.278	.325	.328	.324	.311
SC	B	12.8	**8.07**	**13.6**	105	159	.0461	.209	.655	.600	.629	.583	.0435	.282	.322	.328	.327	.316
	T	**13.0**	**8.06**	**13.6**	**107**	**162**	.0477	.215	.657	.602	.632	.585	.0425	.279	.324	.330	.329	.318
	S	12.1	7.96	**13.4**	100	150	*.0353*	.183	.657	**.606**	**.638**	.576	.0443	.286	.331	.333	.329	.317
RT		13.5	8.40	14.2	116	177	.0300	.151	.531	.452	.481	.530	.0577	.379	.090	.121	.130	.125
GT		14.1	8.39	13.9	122	165	0	0	1	1	1	.602	.0488	.362	1	1	1	1

Top 20 high entropy source utterances found by `IDENTITY`:

yes. | thank you. | why? | here you are. | ok. | what do you mean? | may i help you? | can i help you? | really? | sure. | what can i do for you? | why not? | what? | what happened? | anything else? | thank you very much. | what is it? | i see. | no. | thanks.