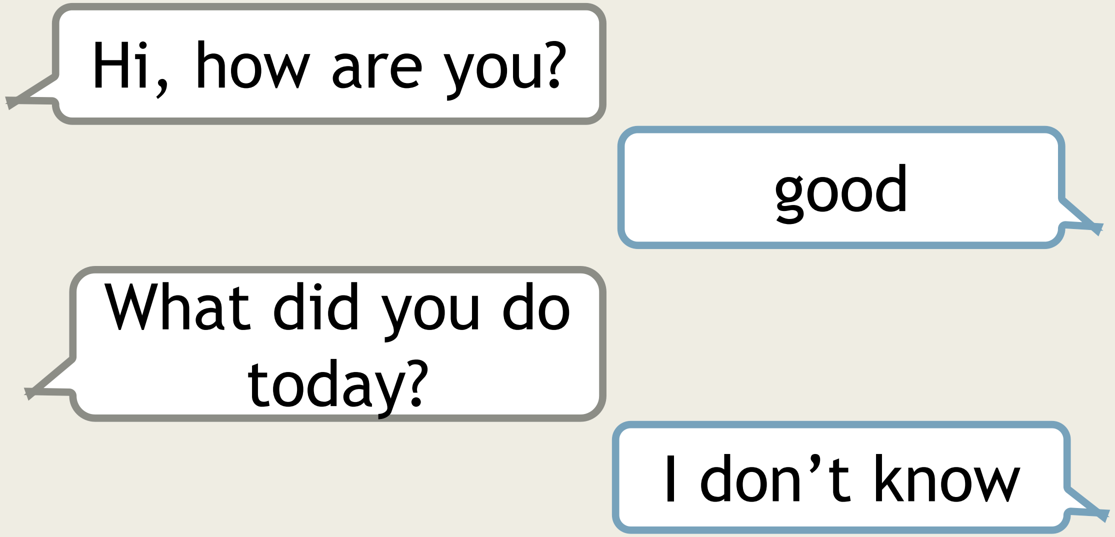
IMPROVING NEURAL CONVERSATIONAL MODELS WITH ENTROPY-BASED DATA FILTERING

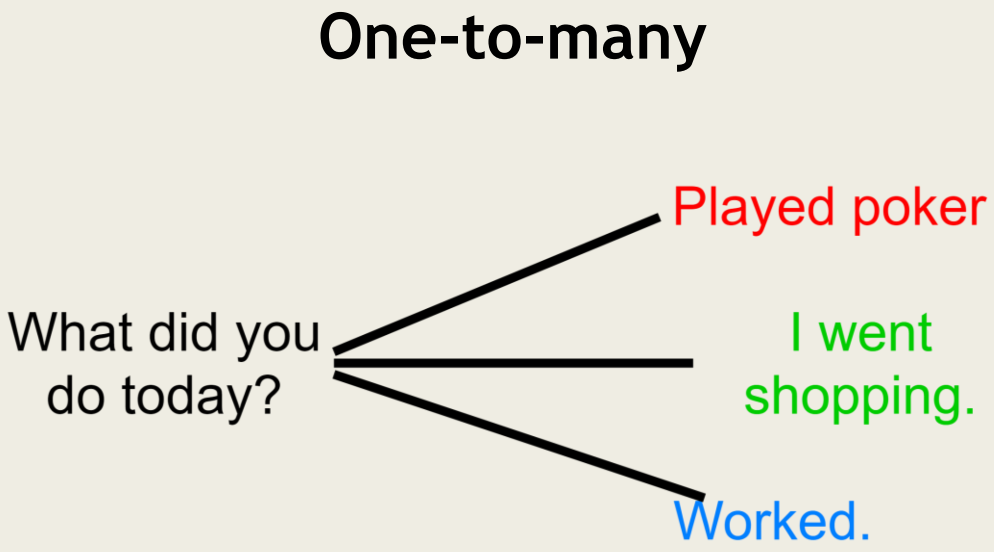
Richard Csaky[1], Patrik Purgai[1], Gabor Recski[1,2]
[1]Budapest University of Technology
[2]Sclable AI

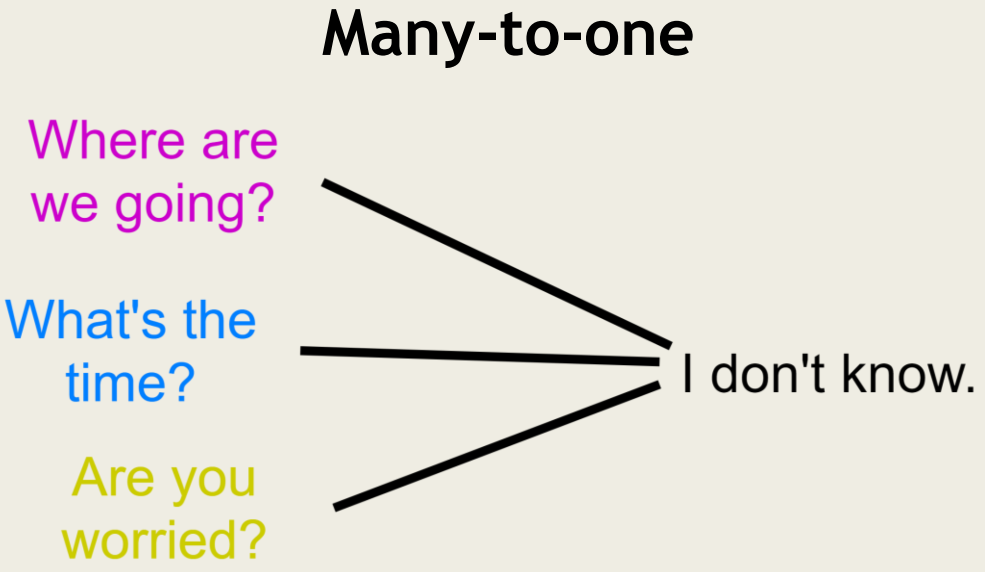
Introduction

- Takeaways
  - ***Better responses*** *by* ***filtering*** *training data*
  - ***Overfitting*** = *better on automatic* ***metrics***



Problem formulation





<u>Previous approaches</u>:

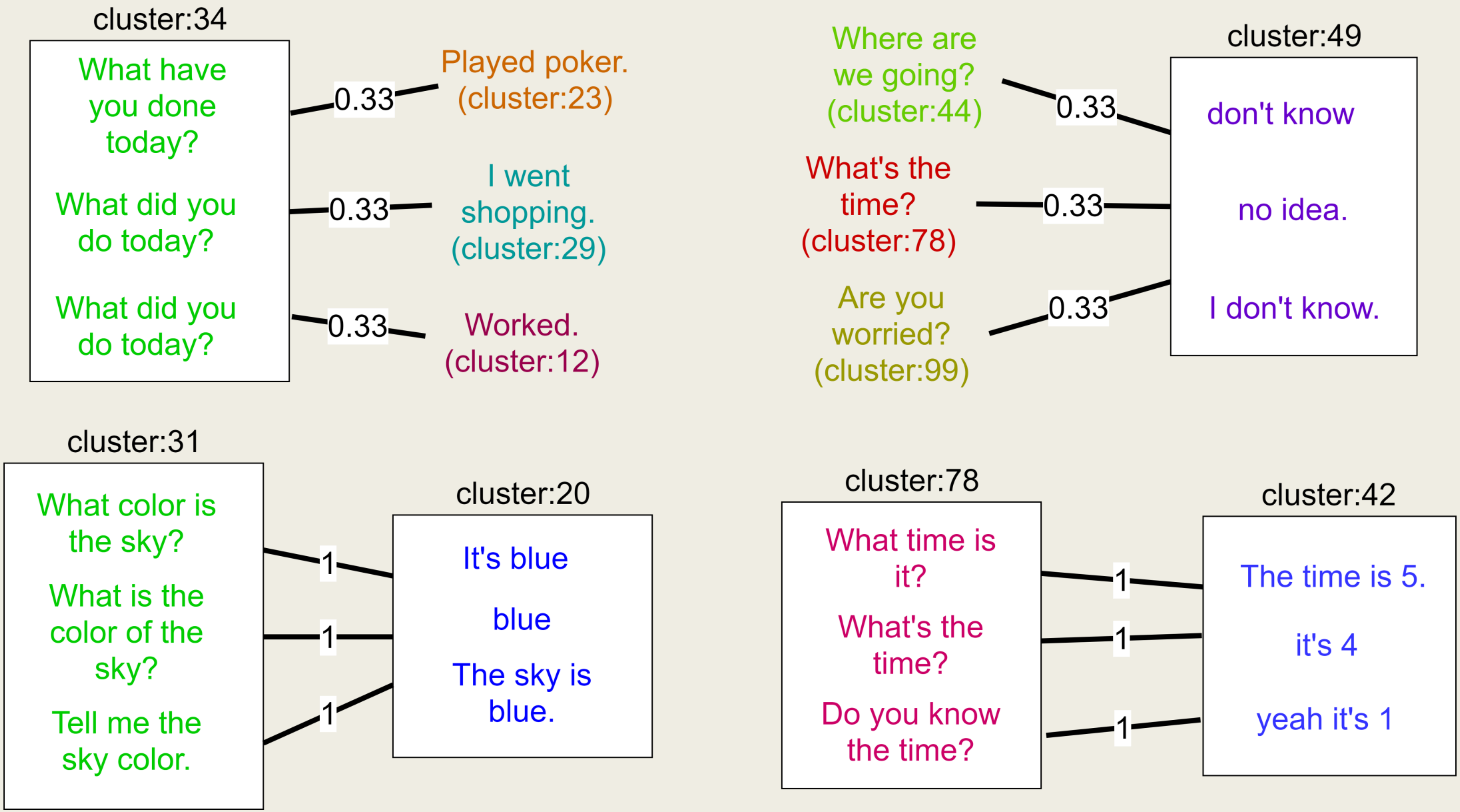
- Feeding **extra information** to dialog models [1]
- **Augmenting** the **model** or decoding process [2]
- Modifying the **loss** function [3]

Methods (Identity)

- Filter **high-entropy** utterances
- 3 filtering ways: SOURCE, TARGET, BOTH

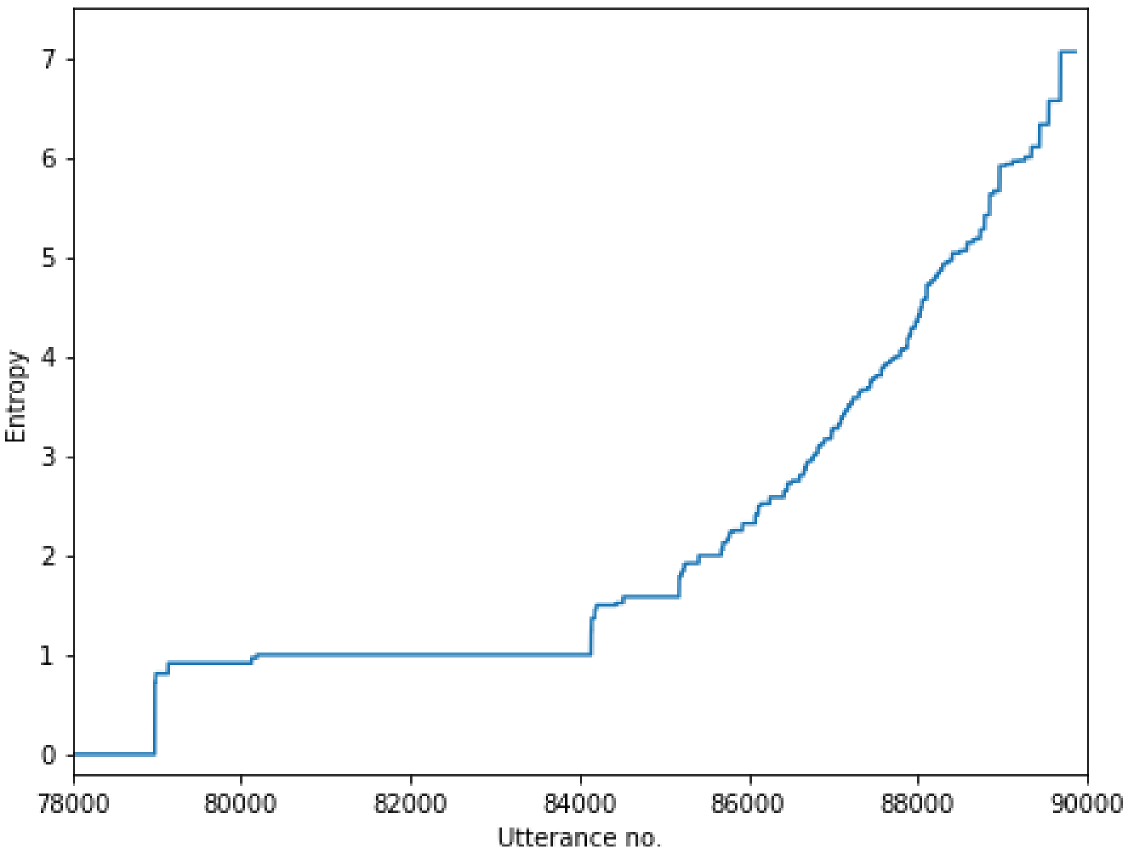
Methods (Clustering)

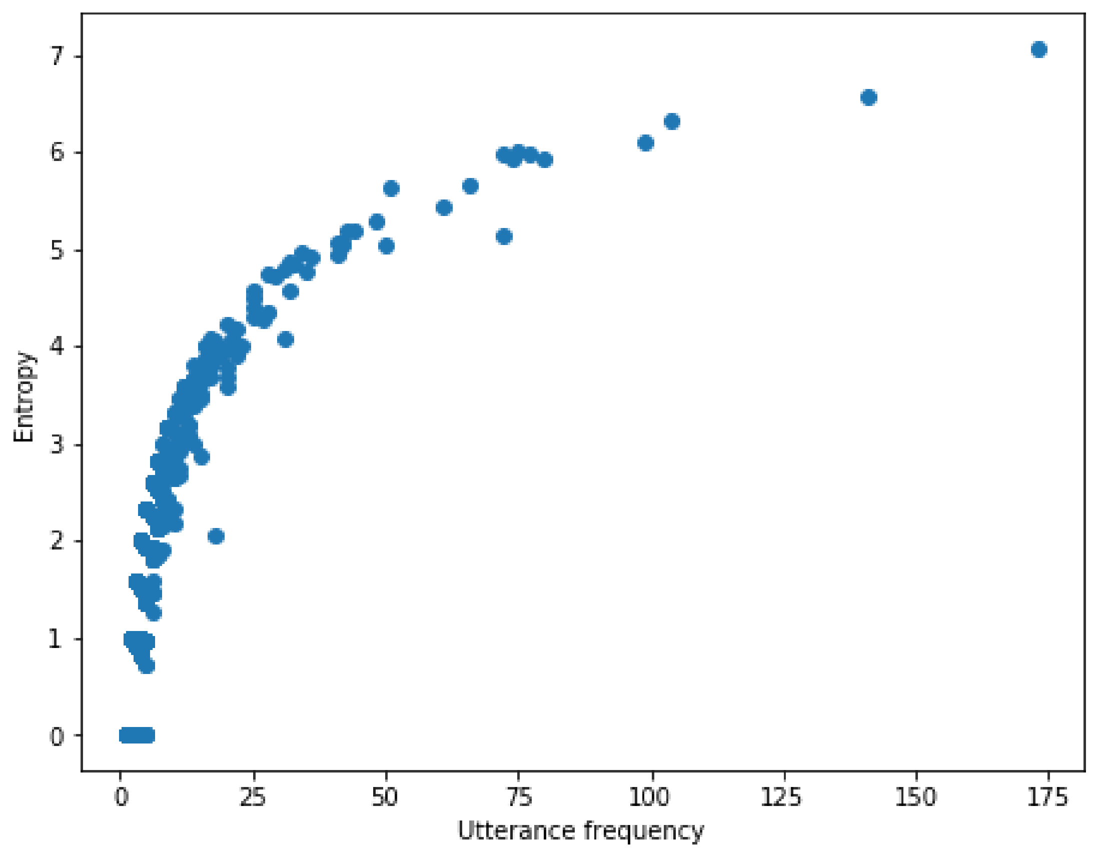
- SENT2VEC [4] and AVG-EMBEDDING [5]
- **Mean Shift** clustering algorithm [6]



Data

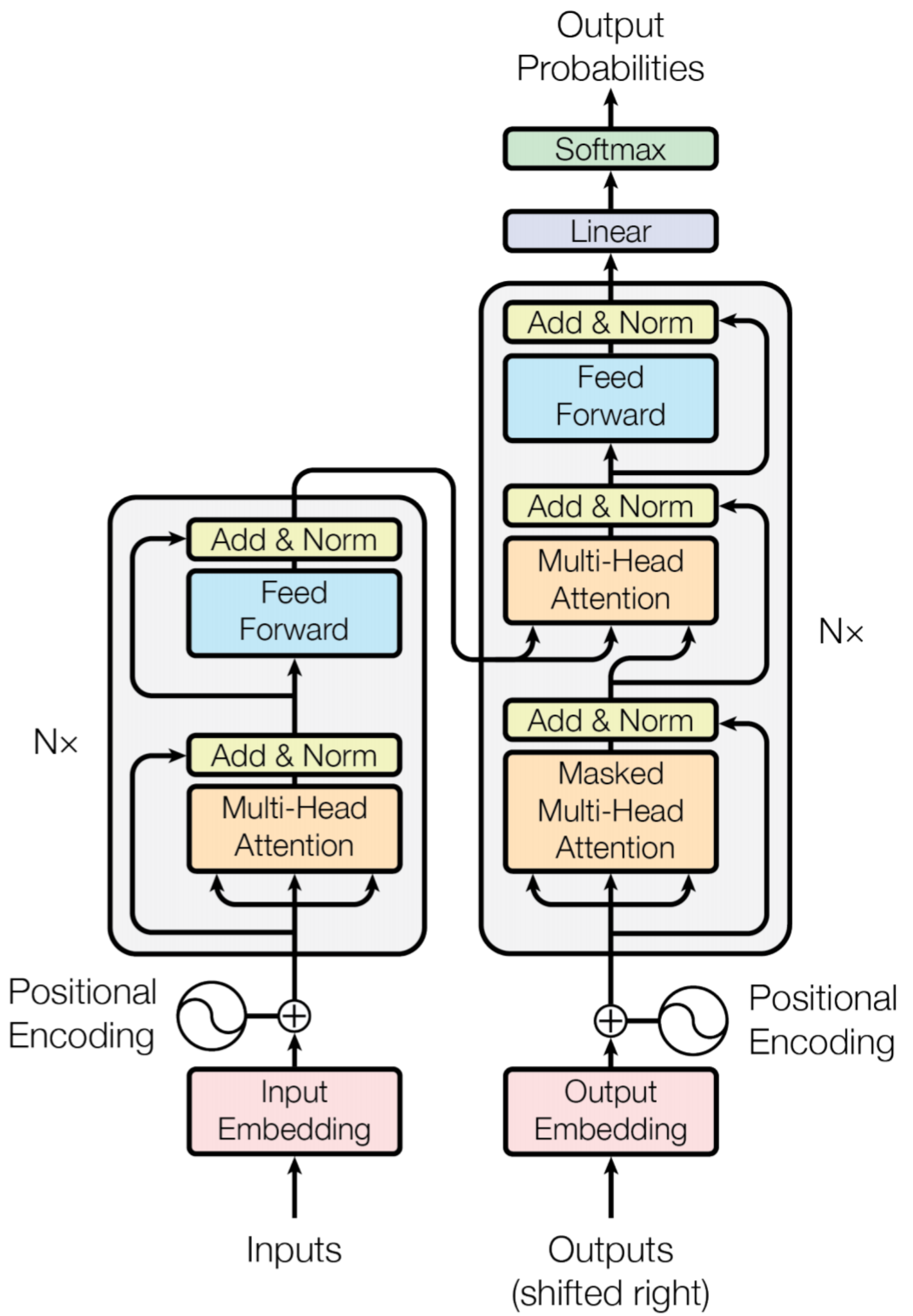
- DailyDialog (~90.000 pairs) [7]
- Remove 5-15% of utterances
- High entropy utterances:
  - *yes* | *thank you* | *why?* | *ok* | *what do you mean?* | *sure*



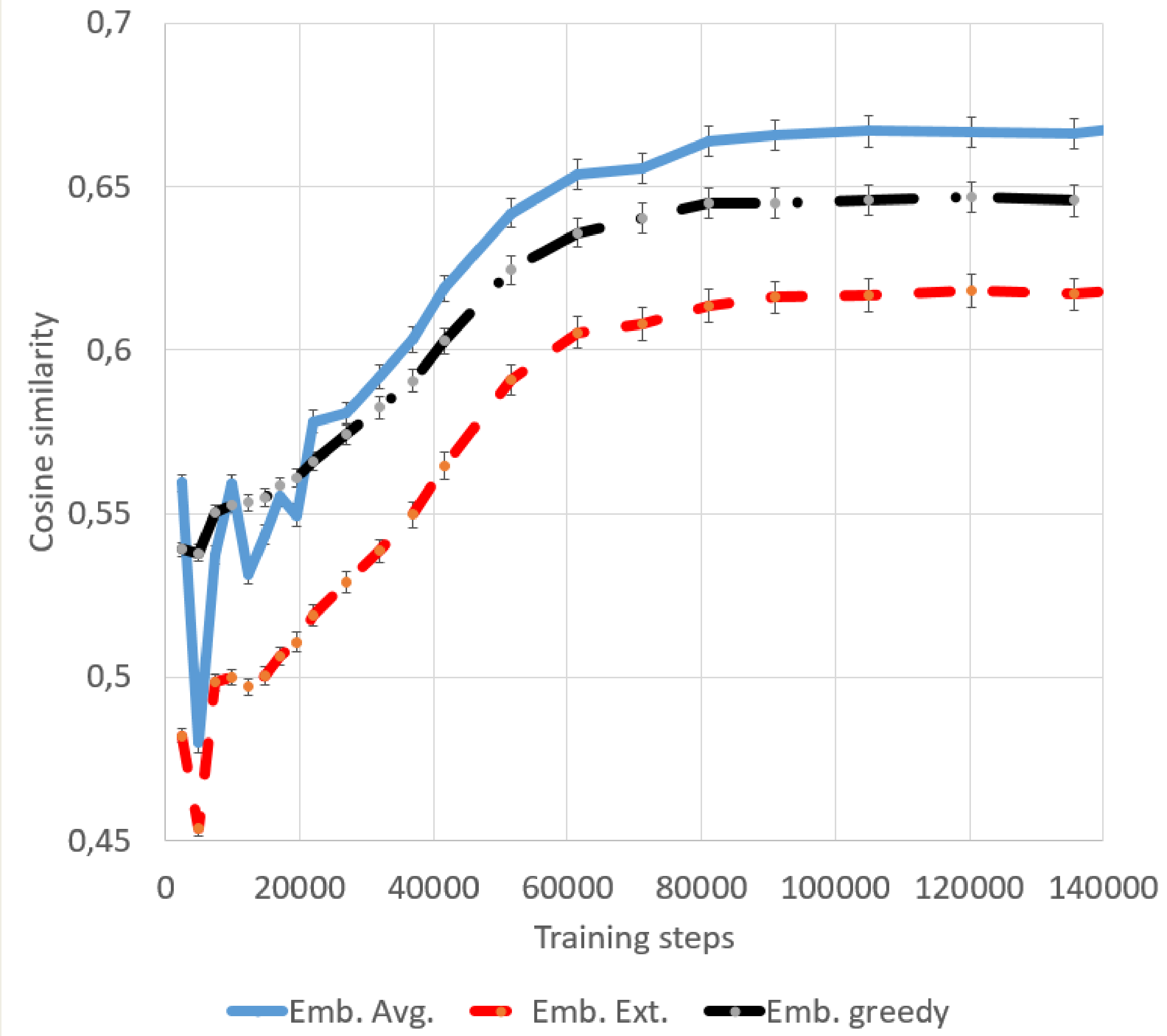


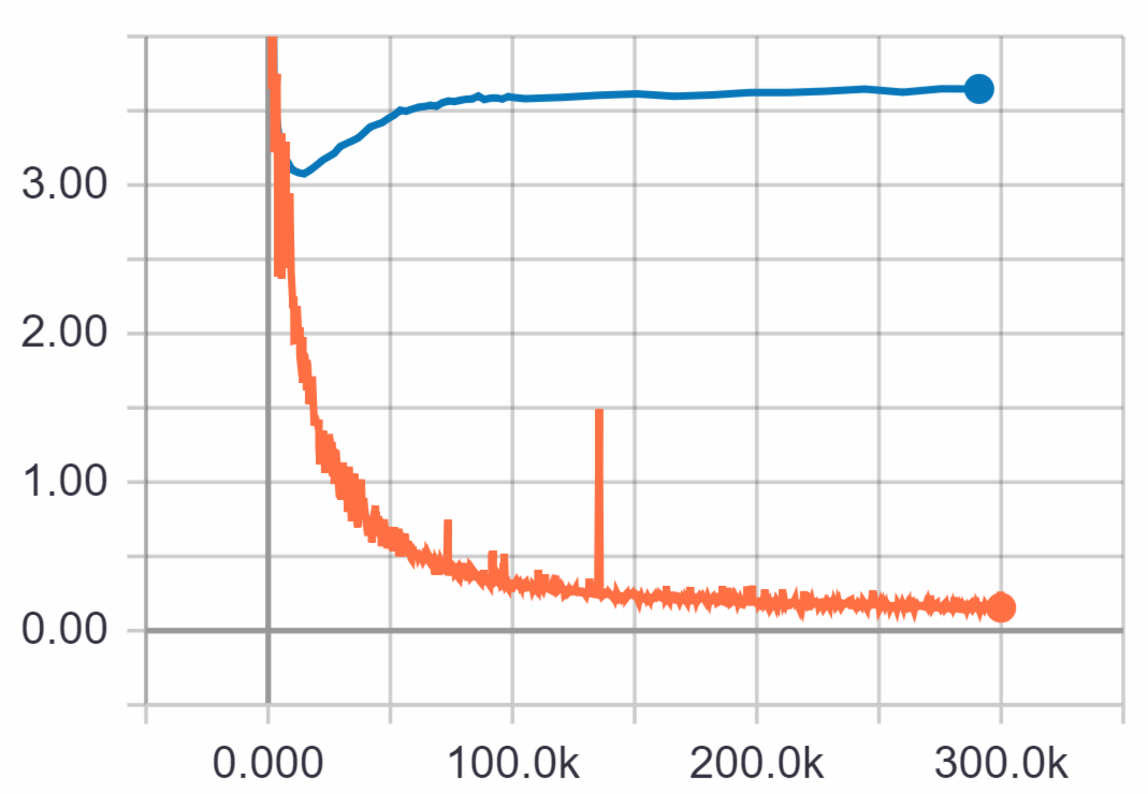
Setup

- Response length
- Word / utterance entropy [8]
- KL-divergence
- Embedding metrics [9]
- Coherence [10]
- Distinct-1, -2 [11]
- BLEU-1, -2, -3, -4 [12]



Evaluation Metrics





Results (at loss minimum)

		$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		8.6	7.30	12.2	63.6	93	.330	.85	.540	.497	.552	.538	**.0290**	.149	.142	.135	.130	.119
ID	B	9.8	7.44	12.3	71.9	105	.315	.77	.559	**.506**	.555	.572	.0247	.138	.157	.151	.147	.136
	T	*10.9*	**7.67**	**12.7**	**83.2**	**121**	**.286**	**.72**	**.570**	**.507**	.554	**.584**	.0266	**.150**	**.161**	**.159**	**.156**	**.146**
	S	9.4	7.19	11.9	66.4	98	.462	1.08	.540	.495	.553	.538	.0262	.130	.139	.133	.128	.117
AE	B	7.9	7.25	12.0	57.7	83	.447	1.05	.524	.486	.548	.524	.0283	.132	.128	.121	.115	.105
	T	8.6	7.26	12.1	61.4	90	.425	1.12	.526	.492	.548	.529	.0236	.115	.133	.127	.121	.111
	S	*9.0*	7.21	11.9	*65.1*	*95*	.496	1.16	.536	.490	.548	.538	.0232	.109	.134	.130	.126	.116
SC	B	10.0	7.40	12.3	72.6	108	.383	.97	.544	.497	.549	.550	.0257	.131	.145	.142	.138	.128
	T	**11.2**	*7.49*	*12.4*	**82.2**	**122**	.391	.97	*.565*	*.500*	.552	*.572*	.0250	.132	*.153*	*.153*	*.152*	*.142*
	S	**11.1**	7.15	11.9	74.4	114	.534	1.27	.546	*.501*	**.560**	.544	.0213	.102	.144	.139	.135	.125

Results (after overfitting)

		$\lvert U \rvert$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	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

Results (other datasets)

Cornell-Movie Dialog Corpus

		$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		8.1	6.55	10.4	54	75	2.29	3.40	**.667**	.451	.635	**.671**	4.7e-4	1.0e-3	**.108**	.120	.120	.112
ID	B	7.4	**6.67**	**10.8**	50	69	**1.96**	**2.91**	.627	**.455**	.633	.637	**2.1e-3**	**7.7e-3**	.106	.113	.111	.103
	T	**12.0**	6.44	10.4	**74**	**106**	2.53	3.79	.646	**.456**	**.637**	.651	9.8e-4	3.2e-3	**.108**	**.123**	**.125**	**.118**

Twitter dataset

		$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		20.6	6.89	**11.4**	121	177	**2.28**	**3.40**	.643	.395	.591	.659	**2.1e-3**	**6.2e-3**	.0519	.0666	.0715	.0693
ID	B	20.3	**6.95**	**11.4**	119	171	2.36	**3.41**	**.657**	.394	.595	**.673**	1.2e-3	3.4e-3	**.0563**	**.0736**	.0795	.0774
	T	**29.0**	6.48	10.7	**157**	**226**	2.68	3.69	.644	**.403**	**.602**	.660	1.4e-3	4.6e-3	**.0550**	**.0740**	**.0819**	**.0810**

Conclusion

- **Better responses** by **filtering** training data
- **Overfitting** = better on automatic **metrics**

Thanks for your attention!

- github.com/ricsinaruto/NeuralChatbots-DataFiltering
  - ***code/utils/filtering_demo.ipynb***
- github.com/ricsinaruto/dialog-eval
- ricsinaruto.github.io
  - *Paper, Poster, Blog post, Slides*

References

[1] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model.
[2] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models.
[3] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models.
[4] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features.
[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings.
[6] Keinosuke Fukunaga and Larry Hostetler. 1975. The estimation of the gradient of a density function, with applications in pattern recognition.
[7] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.
[8] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues.
[9] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.
[10] Xinnuo Xu, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity.
[11] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models.
[12] Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.