THE GUTENBERG DIALOGUE DATASET

Richard Csaky[1,2], Gábor Recski[3]

[1]Budapest University of Technology

[2]University of Oxford

[3]TU Wien

INTRODUCTION

- New dialogue dataset extracted from books
- 14.8M utterances in English
  - Smaller datasets in French, German, Dutch, Spanish, Italian, Hungarian, Portuguese

OUTLINE

1. Current dialogue dataset trade-offs
2. Extraction pipeline
3. Error analysis
4. Trainings and results

Worse quality

Dataset	Size	Source	Quality
DailyDialog (Li et al., 2017b)	90k	ESL websites	auto-extracted
Wizard-of-Wikipedia (Dinan et al., 2019)	100k	crowdsourcing	human-written
Document-grounded (Zhou et al., 2018)	100k	crowdsourcing	human-written
Persona-Chat (Zhang et al., 2018)	150k	crowdsourcing	human-written
Self-dialogue (Fainberg et al., 2018)	150k	crowdsourcing	human-written
Cornell Movie Corpus (Danescu-Niculescu-Mizil and Lee, 2011)	300k	movie scripts	auto-extracted
Self-feeding chatbot (Hancock et al., 2019)	500k	human-bot dialogues	partly human-written
Twitter corpus[7]	5M	Twitter posts/replies	auto-extracted
Gutenberg Dialogue Dataset	15M	Books	auto-extracted
Opensubtitles (Henderson et al., 2019)	320M	movie subtitles	auto-extracted
Reddit (Henderson et al., 2019)	730M	Reddit threads	auto-extracted

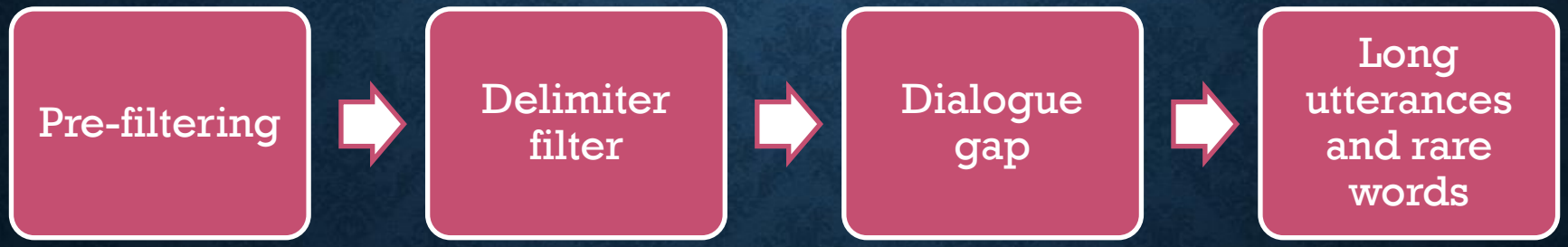
Bigger

DIALOGUE DATASETS

[7]https://github.com/Marsan-Ma-zz/chat_corpus

EXTRACTION PIPELINE

- Project Gutenberg [1]: 60.000 online books in the public domain
- Identifying dialogues and changes between speakers



NEW LANGUAGES

- 0-50 line of python:
  - Specifying conversational delimiters
  - Minimal dialogue and turn segmentation
- Checking pipeline output and refining for quality

	#U	$\|U\|$	#D	$\|D\|$
English	14 773 741	22.17	2 526 877	5.85
German	226 015	24.44	43 440	5.20
Dutch	129 471	24.26	23 541	5.50
Spanish	58 174	18.62	6 912	8.42
Italian	41 388	19.47	6 664	6.21
Hungarian	18 816	14.68	2 826	6.66
Portuguese	16 228	21.40	2 233	7.27

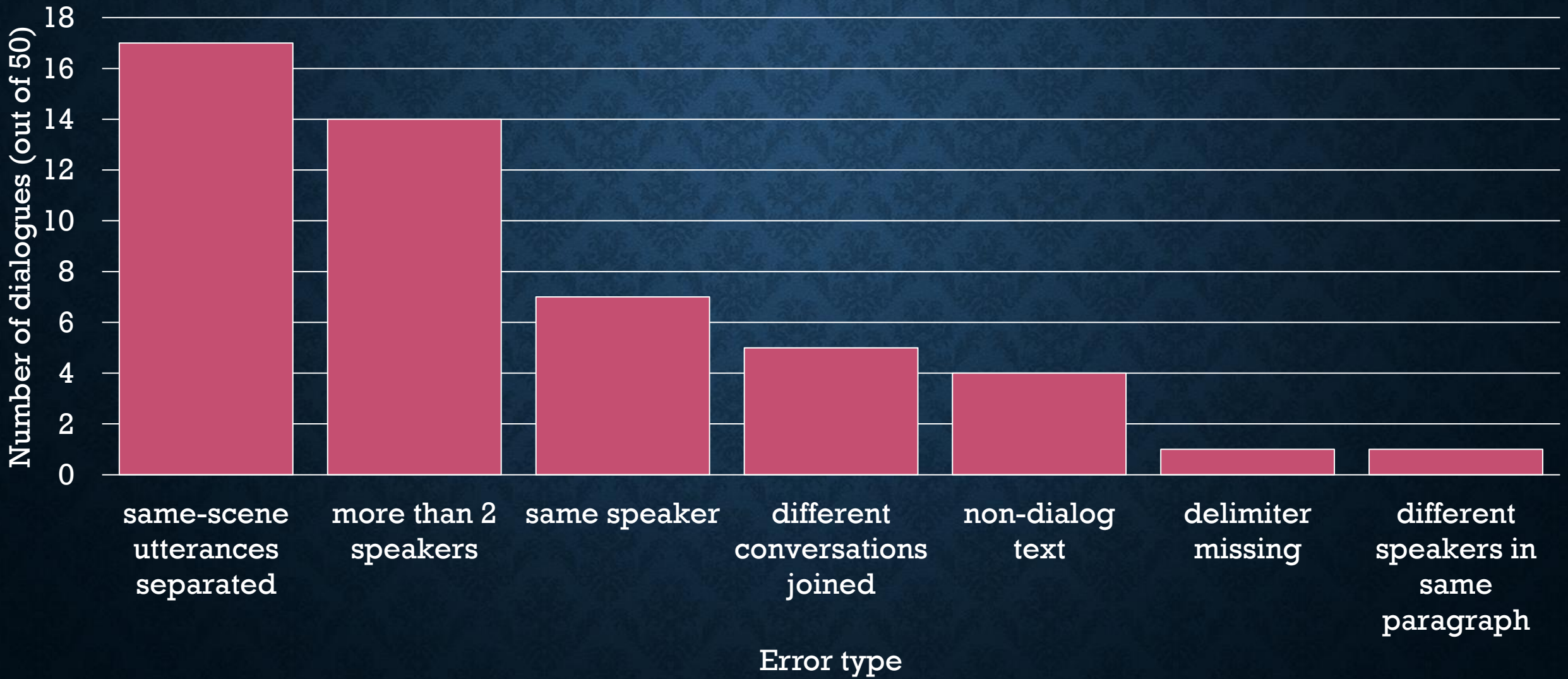
DATASET STATISTICS

in a voice of deep anguish he said,—
“She can sleep—she can sleep—no ghostly vision scares slumber from her eyes—while—”
He shuddered, and passed a step or two on, then pausing again, he said,
“Oh, if she, the young and innocent...”

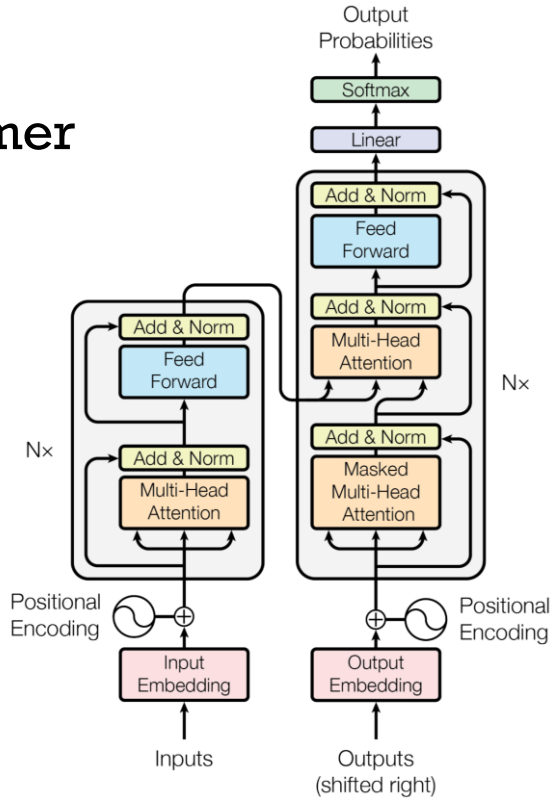
And he was singing, too, as he went on with his task; sometimes–
"Play on, minstrèl, play on, minstrèl, My lady is mine only girl;"

ERROR TYPES

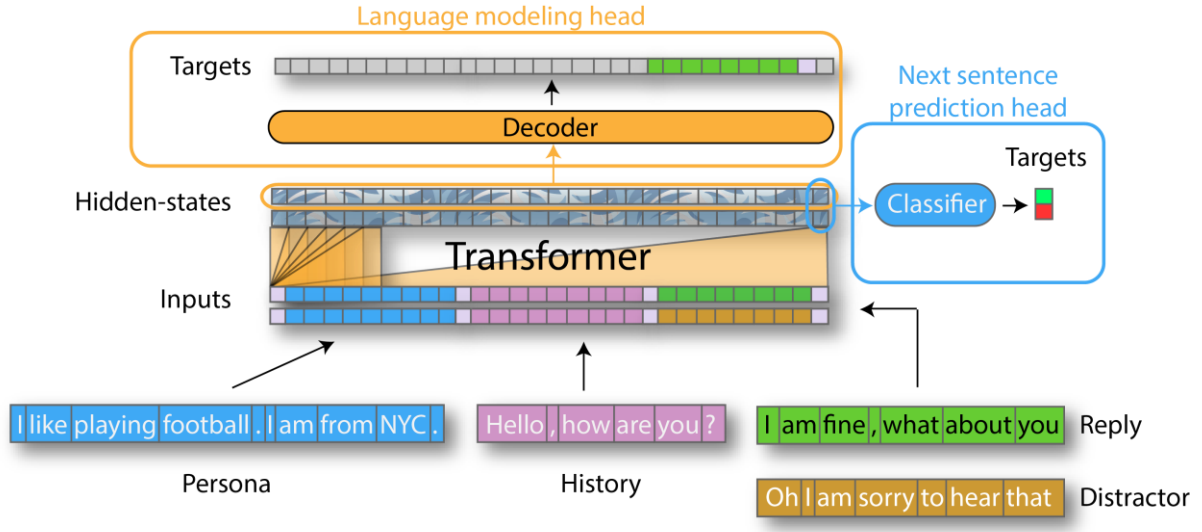
ERROR STATISTICS



Transformer



GPT2



TRAINING SETUP

METRICS [2]

github.com/ricsinaruto/dialog-eval

- Response length
- Word / utterance entropy
- KL-divergence
- Embedding metrics
- Coherence
- Distinct-1, 2
- BLEU-1, 2, 3, 4

			$\|U\|$	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
Transformer	ZS	G	**8.3**	**6.99**	**11.9**	**57.7**	**80**	**1.00**	**2.24**	**.493**	.540	**.545**	**.574**	**.0154**	**.077**	.091	.092	.091	.084
		O	6.6	6.70	11.5	45.2	67	2.00	2.85	.471	**.556**	.542	.476	.0004	.001	**.094**	**.098**	**.095**	**.088**
	FT	G	**11.0**	6.48	10.4	68.2	92	**1.28**	**2.15**	**.513**	**.575**	**.571**	**.593**	**.0104**	**.048**	**.165**	**.163**	**.164**	**.155**
		O	10.6	6.37	10.1	68.3	98	2.58	2.66	.431	**.575**	.532	.444	.0011	.002	.148	.151	.154	.146
		B	**11.1**	**6.88**	**11.0**	**76.5**	**110**	**1.28**	2.21	.508	.570	.562	.559	.0047	.018	**.164**	**.163**	**.165**	**.156**
GPT2	ZS	G	**9.5**	**7.62**	**13.1**	**72.7**	**101**	.56	1.15	**.510**	**.501**	**.531**	**.551**	.0206	.160	**.092**	**.104**	**.107**	**.101**
		O	6.0	7.35	12.6	44.9	60	**.44**	**1.11**	.478	.491	.519	.537	**.0294**	**.186**	.072	.074	.072	.066
	FT	G	**11.0**	7.45	**11.8**	**82.6**	**116**	.27	.64	**.536**	**.559**	**.558**	**.590**	.0182	.129	**.157**	**.159**	**.162**	**.153**
		O	10.5	7.41	11.6	78.1	108	.32	.71	.531	**.558**	.555	.583	.0205	.129	.153	.154	.155	.146
		B	10.3	**7.50**	**11.8**	77.9	108	**.25**	**.61**	.533	.554	.553	.587	**.0219**	**.136**	.151	.154	.155	.146

PRETRAINING RESULTS

- Scenarios: Zeroshot (ZS) and Finetuned (FT)
- Pre-trained on Gutenberg (G) vs Opensubtitles (O) vs only Personachat (B)
- Tested on Personachat test set

		$\|U\|$	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
EN	G	**8.8**	**7.77**	13.4	**69**	**105**	.331	.707	**.494**	.468	.518	**.529**	.0034	.037	.0806	**.0879**	**.0883**	**.0828**
	O	6.1	7.68	13.4	47	68	**.292**	**.689**	.472	**.475**	**.522**	.519	**.0048**	**.045**	**.0867**	.0855	.0810	.0739
DE	G	**7.4**	7.98	13.9	**60**	**84**	**.194**	**.500**	**.536**	.581	**.581**	**.576**	**.0387**	**.241**	.0803	.0813	.079	.0734
	O	6.4	**8.12**	**14.3**	52	72	.269	.635	.524	.581	.579	.566	.0329	.236	**.0825**	**.0864**	**.083**	**.0769**
NL	G	**6.8**	7.81	13.8	**53**	**76**	**.214**	**.624**	.503	.526	.581	.541	**.0453**	**.282**	**.0858**	.0854	.083	.077
	O	5.8	7.79	**14.0**	45	64	.388	.922	.504	.524	.580	.543	.0382	.252	.0850	**.0869**	**.084**	.077
ES	G	**8.0**	7.16	12.1	**58**	**83**	.373	.744	**.452**	.471	**.524**	.473	.056	.242	**.0883**	**.0839**	**.0788**	**.0723**
	O	5.8	**7.76**	**13.4**	46	61	**.198**	**.621**	.438	.466	.516	**.507**	**.093**	**.397**	.0840	.0771	.0716	.0642
IT	G	**6.9**	7.59	12.7	**51**	**69**	**.183**	**.331**	**.452**	.486	.544	**.490**	.131	.451	**.0732**	**.0746**	**.0708**	**.0658**
	O	4.9	**7.89**	**13.6**	39	49	.266	.987	.434	.485	.538	.473	**.155**	**.558**	.0676	.0638	.0604	.0551
HU	G	4.59	7.62	**13.2**	34.3	38	**.176**	.530	**.410**	.452	.520	.447	**.120**	**.463**	.086	.075	.0677	.0609
	O	**5.56**	**7.73**	13.0	**42.1**	**44**	.278	.538	.401	.447	**.529**	.442	.111	.419	**.106**	**.100**	**.0937**	**.0848**
PT	G	**8.4**	7.44	12.6	**63**	**88**	.189	**.495**	.455	.409	.552	.474	.184	.575	.0886	.0933	.093	**.087**
	O	6.3	7.62	**13.0**	49	61	.226	.671	.443	.407	.544	.488	**.210**	**.627**	.0816	.0812	.078	.072

GUTENBERG VS OPENSUBTITLES

- Tested on concatenated test set from both datasets

CONCLUSION

- Large, good-quality dialogue dataset
  - Extracted from books
  - 7 languages
- Community contributions are welcome
  - Improve quality
  - Extend to other languages

THANK YOU FOR YOUR ATTENTION!

- github.com/ricsinaruto/gutenberg-dialog
  - Download datasets/trainings and open-source pipeline
- https://ricsinaruto.github.io/chatbot.html
  - Online chatbot demo for almost all trainings

Work partly supported by Project FIEK 16-1-2016-0007, financed by the FIEK 16 funding scheme of the Hungarian National Research, Development and Innovation Office (NKFIH). Recski was partly supported by BRISE-Vienna (UIA04-081), a European Union Urban Innovative Actions project.

References

[1] https://www.gutenberg.org/
[2] Richard Csaky, Patrik Purgai, Gábor Recski. 2019. Improving Neural Conversational Models with Entropy-Based Data Filtering
(Li et al., 2017b) Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.
(Dinan et al., 2019) Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents
(Zhou et al., 2018) Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations.
(Zhang et al., 2018) Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?
(Fainberg et al., 2018) Joachim Fainberg, Ben Krause, Mihai Dobre, Marco Damonte, Emmanuel Kahembwe, Daniel Duma, Bonnie Webber, and Federico Fancellu. 2018. Talking to myself: self-dialogues as data for conversational agents
(Danescu-Nicolescu-Mizil and Lee, 2011) Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.
(Hancock et al., 2019) Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot!
(Henderson et al., 2019) Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšic, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets.
Transformer image: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need
GPT2 image: Huggingface. 2019. How to build a State-of-the-Art Conversational AI with Transfer Learning