



A Dataset for Building Code-Mixed Goal-Oriented Conversation Systems

Suman Banerjee · Nikita Moghe · Siddhartha Arora · Mitesh M. Khapra

Department of Computer Science and Engineering,
Indian Institute of Technology Madras,
Chennai - 600036, India



Objectives

- Introduce a code-mixed version of modified DSTC2 dataset in 4 languages (Hindi, Bengali, Gujarati and Tamil).
- Analyze the amount of code-mixing in the dataset.
- Report the results of generation based simple baseline models.

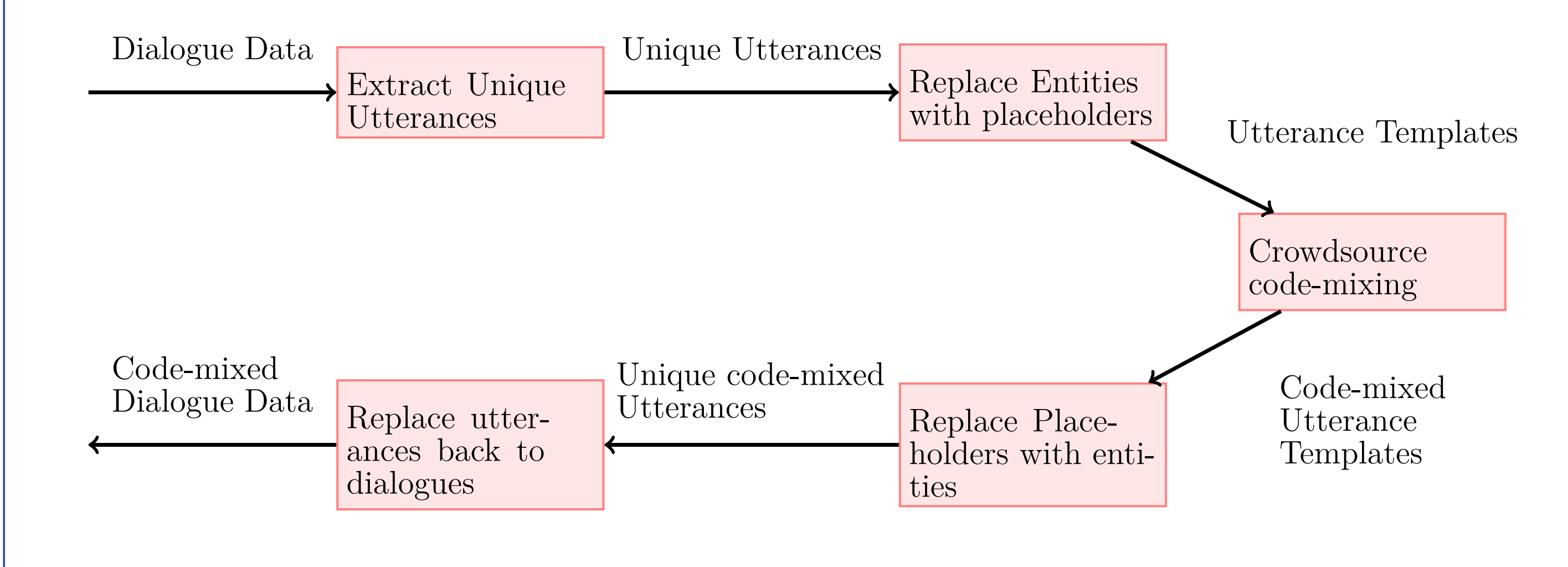
Motivation

- Increasing demand for goal-oriented conversation systems in daily activities.
- Existing systems are limited to English due to the unavailability of other datasets.
- Code-mixing is a predominant phenomenon in multilingual regions of the world.
- Need to fill this gap by introducing code-mixed conversation datasets
- These can facilitate the development of code-mixed goal-oriented conversation systems.

Example

User	Bot
1. <SILENCE>	Hello, Cambridge restaurant system mein aapka swagat hai. Aap paa sakte hai restaurant by area, price range ya food type. Aap ko kaise help kar sakta hu main ?
2. Main ek moderately priced restaurant dhund raha hu.	Aap kis type ka food like karenge?
3. matter nahi karta	town ka kaun se part mein chahiye aapko?
4. south part of town.	api call R cuisine south moderate
pizza.hut cherry.hinton R.post.code pizza.hut.cherry.hinton.post.code pizza.hut.cherry.hinton R.cuisine.italian pizza.hut.cherry.hinton R.location.south pizza.hut.cherry.hinton R.phone pizza.hut.cherry.hinton.phone pizza.hut.cherry.hinton R.address pizza.hut.cherry.hinton.address pizza.hut.cherry.hinton R.price.moderate pizza.hut.cherry.hinton R.rating.3 restaurant.alimentum R.post.code restaurant.alimentum.post.code restaurant.alimentum R.cuisine.european restaurant.alimentum R.location.south restaurant.alimentum R.phone restaurant.alimentum.phone restaurant.alimentum R.address restaurant.alimentum.address restaurant.alimentum R.price.moderate restaurant.alimentum R.rating.10	
5. <SILENCE>	restaurant.alimentum ek.accha restaurant hain in the south of town jo modern.european food serve karta hain.
6. kya mujhe address mil sakta hain ?	Sure, restaurant.alimentum restaurant.alimentum.address.pe.hain.
7. aur phone number kya hain?	restaurant.alimentum.ka.number.hain restaurant.alimentum.phone.
8. okay, thank you, good bye.	Welcome, alvida.

Dataset Creation



Dataset Statistics

	Hindi	Bengali	Gujarati	Tamil
Vocabulary Size	1676	1372	1858	2185
Code-Mixed English Vocabulary	386	360	387	424
Native Language Vocabulary	739	477	912	1214
Others Vocabulary	551	535	559	547
Unique Utterances	6549	6274	6417	6666
Utterances with code-mixed words	5750	5703	5643	5632
Pure Native Language utterances	348	210	340	420
Pure English utterances	451	361	434	614
Average length of utterances	8.16	7.74	8.04	6.78
Average # of code-mixed utterances per dialog	12.11	14.28	11.80	12.96

Code-Mixing Index

$$C_u(x) = K \cdot \left\{ \frac{N(x) - \max_{L_i \in \mathcal{L}} \{t_{L_i}\}}{N(x)} + \frac{P(x)}{N(x)} \right\} : (if N(x) > 0)$$

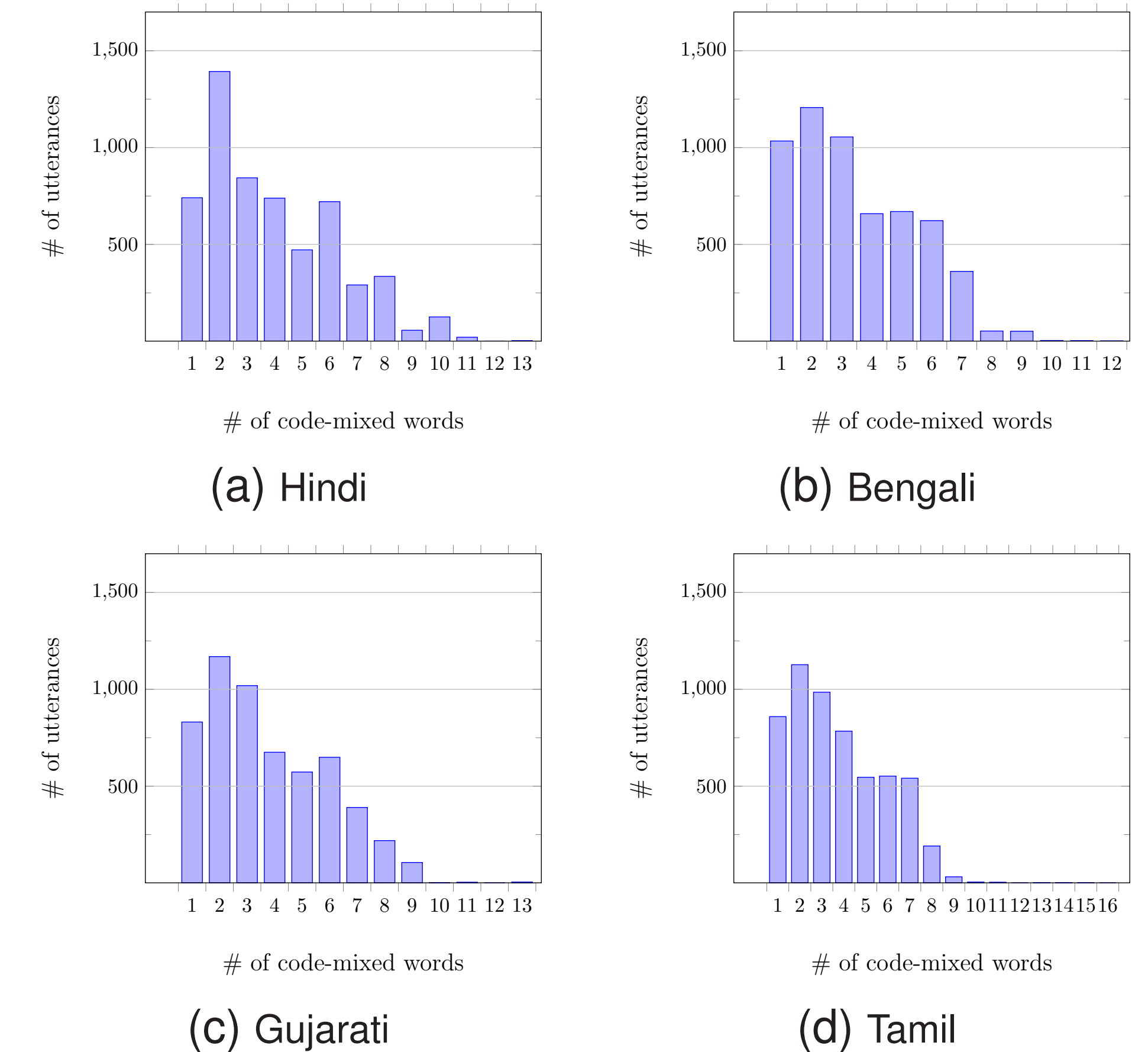
$$C_{avg} = \text{Average of } C_u(x)$$

$$native(x) = \begin{cases} t_{L_n} & : t_{L_n} > 0 \\ N(x) & : t_{L_n} = 0 \end{cases}$$

$$C_c = \frac{100}{U} \left[\frac{1}{2} \sum_{i=1}^U \left(1 - \frac{native(x) + P(x)}{N(x)} + \delta(x) \right) + \frac{5}{6} S \right]$$

Language-pair	En-Be	En-Hi	En-Hi	En-Hi	En-Hi	En-Hi	En-Hi	En-Be	En-Gu	En-Ta
Dataset	TW	TW	FB	FB+TW	Vyas	Hi-DSTC2	Be-DSTC2	Gu-DSTC2	Ta-DSTC2	
I-index	-	-	-	-	-	0.04	0.04	0.03	0.03	
C _{avg}	8.34	21.19	3.92	11.82	11.44	32.12	31.80	31.66	29.54	
δ	22.09	30.99	6.70	17.81	53.50	26.38	29.06	24.50	38.32	
C _c	25.14	64.38	16.76	38.53	31.31	73.31	76.27	71.63	80.49	

of utterances vs # of code-mixed words



Human Evaluation

Languages	Colloquialism	Intelligibility	Coherent
Hindi-English	4.20	4.06	4.21
Bengali-English	4.07	4.05	4.11
Gujarati-English	3.66	3.6	3.76
Tamil-English	4.17	3.96	3.93

Baseline Results

SEQ2SEQ WITH ATTENTION					
Metrics	En-DSTC2	Hi-DSTC2	Be-DSTC2	Gu-DSTC2	Ta-DSTC2
BLEU-4	56.6	54.0	56.8	53.8	62.1
ROUGE-1	67.2	62.9	67.4	64.7	67.8
ROUGE-2	55.9	52.4	57.5	54.8	56.3
ROUGE-L	64.8	61.0	65.1	62.6	65.6
Per response acc.	46.0	48.0	50.4	47.6	49.3
Per dialog acc.	1.4	1.2	1.5	1.5	1.3

HRED					
Metrics	En-DSTC2	Hi-DSTC2	Be-DSTC2	Gu-DSTC2	Ta-DSTC2
BLEU-4	57.8	54.1	56.7	54.1	60.7
ROUGE-1	67.9	63.3	67.1	65.3	67.1
ROUGE-2	57.5	52.6	56.9	55.2	55.6
ROUGE-L	65.7	61.5	64.8	63.2	65.1
Per response acc.	48.8	47.2	47.7	47.9	47.8
Per dialog acc.	1.4	1.5	1.6	1.6	1.0

References

- Mihail Eric and Christopher Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 468-473.
- Bjorn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz, Slovenia, May 23-28, 2016*.