Natural Language Processing for Multilingual Task-Oriented Dialogue

Tutorial Abstract

Evgeniia Razumovskaia[1], **Goran Glavaš**[2], **Olga Majewska**[1]
Edoardo Maria Ponti[4,5], **Ivan Vulić**[1,6]
[1] Language Technology Lab, University of Cambridge
[2] Center for Artificial Intelligence and Data Science, University of Würzburg
[4]Mila – Quebec Artificial Intelligence Institute [5]McGill University [6]PolyAI Limited
{er563,om304,iv250}@cam.ac.uk goran@informatik.uni-mannheim.de
edoardo-maria.ponti@mila.quebec

1 Motivation and Objectives

Enabling machines to intelligently *converse* with humans in order to solve particular well-defined *tasks* is in the core focus of task-oriented dialogue (ToD) systems and their development (Kim and Banchs, 2014; Li et al., 2018; Henderson et al., 2019; Zang et al., 2020). Such systems have wide applications in a multitude of domains such as hospitality industry, travel, e-banking, healthcare, entertainment industry, industrial production and maintenance, etc. ToD-oriented research has been recently catalysed by the growing ability and viability of deep learning techniques such as large-scale pretraining of language models (Ren et al., 2018; Wen et al., 2019; Henderson et al., 2020; Wu et al., 2020; Lin et al., 2020, *inter alia*). The momentum of development in this research area has, however, mainly targeted a very small proportion of potential beneficiaries: most existing ToD systems are predominantly built for English and a few other, major languages only (e.g., Chinese) (Lin et al., 2021; Ding et al., 2021). This limits the use, global reach, and transformative potential of ToD systems. Consequently, this deepens the chasm between speakers of dominant versus under-represented low-resource languages in their access to state-of-the-art language technology (Joshi et al., 2020; Blasi et al., 2021) and contributes to the digital language divide and inequality of information.[1]

Extending the reach of ToD technology is crucial for the democratisation and wide adoption of human–machine communication, with an inclusive long-term goal of bringing it to virtually *all* citizens of the world. Building on top of our recent comprehensive survey on the topic of *multilingual* ToD (Razumovskaia et al., 2021), in this tutorial our aim is to systematise the current research on multilingual ToD, and offer a fresh perspective to other researchers and NLP practitioners on the importance and challenges of developing multilingual ToD systems.

The tutorial will offer a comprehensive overview of multilingual ToD research and know-hows focused on the following central questions:[2]

(Q1) Why are multilingual ToD systems so hard to build? What are the main **roadblocks** and how can we facilitate their development? What are the main **design paradigms** of multilingual ToD?

(Q2) Which ToD **datasets** are currently available in one or more languages other than English? What are their strengths and weaknesses? How can we improve the current data design and collection efforts and protocols?

(Q3) What are the best **methods** and practices to incorporate language-specific information and perform target language adaptation for multilingual and cross-lingual ToD?

(Q4) How can multilingual ToD take inspiration from other **related fields** of NLP research to better tackle low-resource scenarios (e.g., cross-lingual transfer, injection of external knowledge into parameters of neural models)?

(Q5) How good are current (multilingual) ToD systems? Do automatic **evaluation** measures correlate with user satisfaction? What implication does multilinguality have on ToD evaluation?

(Q6) What are the **future challenges** faced when developing ToD systems in several different languages, especially with respect to voice-based and human-centered ToD?

2 Tutorial Overview and Structure

Part I: Introduction, Motivation, and ToD Preliminaries *(25 minutes)*

[1] http://labs.theguardian.com/digital-language-divide/

[2] All tutorial materials will be available at https://tinyurl.com/multilingualtod

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 44 - 50

Part I will cover the basics of modular and end-to-end dialogue systems (Young, 2010; Chen et al., 2017a; Wen et al., 2017; Ham et al., 2020), offering a brief overview of the full ToD system structure, and critical modules such as Speech-to-text/ASR, natural language understanding (NLU) for dialogue, dialogue state tracking (DST), dialogue management (DM), natural language generation (NLG), and text-to-speech (TTS), along with their functionality. We will analyse which components require language-specific processing and adaptation, and which modules are generally language-invariant. We will then proceed to define the detailed scope and schedule of the tutorial. Concretely, we plan to list and discuss all the current problems and challenges related to multilingual ToD development, and how we will introduce them in the subsequent tutorial parts. Topics overview:

- Main modules of ToD systems;
- Modular versus end-to-end ToD;
- Text-based (vs. other) ToD modules;
- Language-invariant vs. language-specific ToD modules;
- Why is development of multilingual ToD systems so difficult?

Part II: Methods and Resources for Multilingual *NLU* in ToD *(50 minutes)*

Part II will cover to-date work in multilingual NLU in ToD, including standard approaches and recent trends. We will provide a comprehenstive overview of methods for learning cross-lingual representation spaces in ToD (Liu et al., 2019; Siddhant et al., 2020; Liu et al., 2020; Moghe et al., 2021) and their applications in different setups (multilingual vs. cross-lingual, zero-shot vs. few-shot). Finally, we will list available resources: those created specifically for multilingual ToD NLU (Ding et al., 2021; Zuo et al., 2021; Hung et al., 2022, *inter alia*) as well as external resources useful for ToD NLU. Part II comprises the following topics:

- Joint versus separate training for NLU: intent detection, slot labeling, DST;
- Learning shared cross-lingual representation spaces; from cross-lingual word embeddings to multilingual text encoders – how to leverage them for NLU in ToD?
- Multilingual (pre)training versus cross-lingual transfer methods;
- Zero-shot and few-shot learning scenarios;
- Datasets and resources: (a) for in-task dialogue training and (b) external resources (e.g., parallel data, bilingual dictionaries, multilingual knowledge bases).

Part III: Methods and Resources for Multilingual *NLG* in ToD *(35 minutes)*

Part III will present the methods for multilingual natural language generation and their usage for cross-lingual transfer of ToD. First, we will discuss traditional, grammar-based methods for cross-lingual generation (Gatt and Krahmer, 2018; Vaudry and Lapalme, 2013) and their combination with statistical methods (García-Méndez et al., 2019) for more efficient learning. Secondly, we will discuss cross-lingual transfer of ToD using machine translation (MT) in two ways: a) translating the test data into English ('translate test', Wan et al., 2010); b) translating the training data into the target language ('translate train', Duan et al., 2019), and how improvements in MT and multilingual pretraining affect cross-lingual transfer of ToD. Next, we will analyse the choice between retrieval-based, generation-based and hybrid ToD systems through the prism of multilinguality. Finally, we will address the difficulties of corpora creation for multilingual ToD generation. Topics:

- Traditional NLG and its extension to multiple languages;
- Retrieval-based versus generation-based versus hybrid approaches: pros and cons in multilingual setups;
- Leveraging shared cross-lingual representation spaces for multilingual NLG; translation-based approaches;
- Zero-shot and few-shot learning scenarios and language-specific adaptations;
- Available resources and datasets for multilingual NLG (for ToD).

Part IV: Evaluation of Multilingual ToD Systems *(30 minutes)*

Part IV will focus on evaluation for (multilingual) ToD. We will cover both automatic metrics and human evaluation: automatic metrics allow for faster development cycles, but often do not correlate with user satisfaction with ToD systems (Liu et al., 2016; Novikova et al., 2017). We will discuss the shortcomings of automated ToD evaluation, but also the potential pitfalls of human evaluation (Clark et al., 2021). We will then analyse the difficulties that multilingual setups pose for both automatic metrics and human evaluation, including evaluation of generated responses in morpho-

logically rich languages and difficulty of finding qualified evaluators for rare languages. Topics:

- Current evaluation protocols in ToD;
- Automatic vs. human-centered evaluation in multilingual setups: pros and cons;
- How to evaluate language-specific phenomena and fluency;
- Difficulties in evaluation and current gaps in evaluation resources.

Part V: Open Challenges and Research Directions in Multilingual ToD *(40 minutes)*

In the concluding Part V, we will discuss the main open challenges impeding the development of ToD systems and reflect on the promising avenues for further progress. First, we will advocate for linguistically motivated design of multilingual ToD datasets focusing on linguistic diversity and idiomacity. To fulfill their role as gauges of model performance across languages (Hu et al., 2020; Liang et al., 2020), multilingual datasets should (i) maximise diversity along the dimensions of language family, geographic area, and typological features (Ponti et al., 2020). as well as (ii) adequately represent the linguistic and extra-linguistic (e.g., world knowledge, cultural references) properties of selected languages (rather than replicating dialogue structures, topics, and entities from a resource-rich source language). We will discuss first attempts at cultural adaptation for dialogue (Majewska et al., 2022). Second, we will outline how existing strategies for dealing with data scarcity can be borrowed from other NLP tasks to benefit multilingual and cross-lingual ToD NLU (Ponti et al., 2019; Hedderich et al., 2021). Third, we will emphasise the importance of user-centered evaluation as a way of assessing the fluency of generated responses and guiding improvements in ToD systems across different languages. Finally, we will discuss the significance of developments in multilingual ASR and TTS as keys to the ultimate success of multilingual ToD on a wide scale, and the potential of integrating speech-based and text-based modules in future research. Topics:

- Recommendations for creation of future multilingual ToD datasets: linguistic diversity and idiomacity, low-resource languages, expansion to new domains;
- Coping with low-resource scenarios: methods and lessons learned from other NLP tasks and applications; source selection for multi-source transfer and multilingual training;
- Fluency of generation, code switching;
- From text-based to voice-based multilingual ToD: promises and challenges;
- An overview of other related research areas that can benefit multilingual ToD;
- Listing key challenges, a short panel discussion and a QA session.

3 Tutorial Breadth and Diversity

According to the representative set of papers listed in the selected bibliography as well as in our recent survey paper (Razumovskaia et al., 2021), we anticipate that a total of 20%-25% of the tutorial concerns work which involves at least one of the five presenters. The rest of the tutorial will focus on providing a detailed comprehensive overview of the main topic by covering all the relevant work from other researchers: see again the wide bibliography and coverage in the survey paper.

Diversity and Inclusion. We consider the following aspects. First, our tutorial proposal focuses on multilingual NLP and promotes the ultimate long-term goal of NLP research: bringing (human-centered) language technology to minor and under-resourced languages, and acting as a vehicle of mitigating the *digital language divide* (see the footnote 1). As such, it is highly relevant to both special themes of ACL 2022 and NAACL-HLT 2022. Our tutorial will also expose prominent issues and gaps related to (lack of) diversity and inclusivity of current multilingual ToD models and datasets, and we hope to inspire research groups currently working separately on (i) ToD and (ii) low-resource languages and low-resource NLP to consider joining forces and research expertise in the future.

Concerning tutorial organization, we hope that our tutorial will connect researchers from different cultural backgrounds and research fields. We also note that two out of five tutorial presenters are female, and the pool of presenters offers a mix of more junior and experienced presenters.

4 Presenters

Evgeniia Razumovskaia is a PhD student in the Language Technology Lab at the University of Cambridge. She works on dialogue systems, focusing on efficient few-shot methods for multilingual dialogue systems. Web: `evgeniiaraz.github.io`

Goran Glavaš is a Full Professor (Chair for Natural Language Processing) and member of the

Center for Artificial Intelligence and Data Science (CAIDAS) at the University of Würzburg. His research focuses on multilingual representation learning and cross-lingual transfer (primarily for low-resource languages), fair and sustainable NLP, and NLP applications for social sciences and humanities. He has given tutorials at ACL 2019 and EMNLP 2019, organized workshops TextGraphs and SustainNLP, and served as reviewer and (senior) area chair for a number of *ACL events. He currently serves as an Editor-in-Chief for the ACL Rolling Review. Web: sites.google.com/view/goranglavas

Olga Majewska works at Amazon Alexa in Cambridge, UK, and an affiliated researcher at the Language Technology Lab, University of Cambridge, where she earned her PhD in computational linguistics in 2021. Her interests lie, among others, in multilingual expansion of conversational AI and development of efficient protocols for generation of task-oriented dialogue evaluation data for under-resourced languages. Web: om304.github.io

Edoardo Maria Ponti is a Visiting Postdoctoral Scholar at the University of Stanford and a Postdoctoral Fellow at MILA Montreal. He works on sample efficiency and modularity in neural networks, with applications to multilingual NLP. In 2020, he obtained a PhD in computational linguistics from the University of Cambridge, St John's College. Previously, he interned as an AI/ML researcher at Apple in Cupertino. His research earned him a Google Research Faculty Award and an ERC Proof of Concept grant. He received 2 Best Paper Awards at EMNLP 2021 and RepL4NLP 2019. Web: ducdauge.github.io

Ivan Vulić is a Senior Research Associate in the Language Technology Lab at the University of Cambridge, and a Senior Scientist at PolyAI. His research interests are in multilingual and multimodal representation learning, and transfer learning for low-resource languages and applications such as task-oriented dialogue systems. He has extensive experience giving invited and keynote talks, and co-organising tutorials (e.g., EMNLP 2017, NAACL-HLT 2018, ESSLLI 2018, ACL 2019, EMNLP 2019, AILC Lectures 2021) and workshops in areas relevant to this proposal (e.g., SIGTYP, DeeLIO, RepL4NLP, PC of *SEM 2021). For his contributions to NLP and IR, he obtained the 2021 Karen Spärck Jones award. Web: sites.google.com/site/ivanvulic

5 Prerequisites and Reading List

Math: no special requirements; *Linguistics*: basic knowledge of language typology and of morphology (recommended); *Machine Learning*: good grasp of core (supervised) machine learning concepts and familiarity with self-supervised pretraining of language models (required). Pre-tutorial reading list (examples):

- Wen, T. H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P. H., & Young, S. 2017. A Network-Based End-to-End Trainable Task-Oriented Dialogue System. EACL 2017 (pp. 438-449).
- Blasi, D., Anastasopoulos, A., & Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. arXiv preprint arXiv:2110.06733.
- Razumovskaia, E., Glavaš, G., Majewska, O., Korhonen, A., & Vulić, I. (2021). Crossing the Conversational Chasm: A Primer on Multilingual Task-Oriented Dialogue Systems. arXiv preprint arXiv:2104.08570.

6 Other Tutorial Information

Related Tutorials. Conversational AI and (components of) ToD systems have been taught in several tutorials in past years, where the focus has been put on diverse aspects such as: deep learning techniques for ToD (Chen et al., ACL 2017; Su et al., NAACL-HLT 2018; Gao et al., ACL 2018), data collection and end-to-end learning (Wen et al., EMNLP 2019), or NLG methods (Ji et al., EMNLP 2020) and their evaluation (Khapra and Sai, NAACL-HLT 2021). However, our tutorial is the first to focus on the crucial aspects of multilingualism and low-resource languages in relation to the design, development, evaluation, and application of (multilingual) ToD systems. Our tutorial offers a completely novel and unique perspective to ToD also through the optics of multilingual NLP.

Ethical Considerations. ToD systems can and should be used for greater good, but their use also comes with potential harmful implications. As part of the tutorial, we will therefore also point to guidelines and required ethical standards related to ToD-oriented data collection and (user-centered) evaluation, and also provide an overview of potential threats in current ToD-oriented models (e.g., gender, race or religion biases (Barikeri et al., 2021)). Furthermore, we will remind NLP researchers and practitioners to bear in mind potential data- and model-centered biases, and apply appropriate data filtering and debiasing techniques before deploying ToD systems in real-world settings.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. In *Proceedings of ACL 2021*, pages 1941–1955.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv preprint arXiv:2110.06733*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017a. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations*, 19(2):25–35.

Yun-Nung Chen, Asli Celikyilmaz, and Dilek Hakkani-Tür. 2017b. Deep learning for dialogue systems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 8–14.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of ACL-IJCNLP 2021*, pages 7282–7296.

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Aljunied Mahani, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. *CoRR*, abs/2110.07679.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of ACL 2019*, pages 3162–3172.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 2–7.

Silvia García-Méndez, Milagros Fernández-Gavilanes, Enrique Costa-Montenegro, Jonathan Juncal-Martínez, and F. Javier González-Castaño. 2019. A library for automatic natural language generation of Spanish texts. *Expert Systems with Applications*, 120:372–386.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of ACL 2020*, pages 583–592.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for Natural Language Processing in low-resource scenarios. In *Proceedings of NAACL-HLT 2021*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Proceedings of EMNLP 2020*, pages 2161–2174.

Matthew Henderson, Ivan Vulić, Inigo Casanueva, Paweł Budzianowski, Daniela Gerz, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Polyresponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking. In *Proceedings of EMNLP 2019*, pages 181–186.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML 2020*, pages 4411–4421.

Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. Multi^{2}WOZ: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page *to appear*.

Yangfeng Ji, Antoine Bosselut, Thomas Wolf, and Asli Celikyilmaz. 2020. The amazing world of neural language generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 37–42.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020*, pages 6282–6293.

Mitesh M. Khapra and Ananya B. Sai. 2021. A tutorial on evaluation metrics used in natural language generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 15–19.

S. Kim and R. E. Banchs. 2014. R-cube: A dialogue agent for restaurant recommendation and reservation. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–6.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of EMNLP 2020*, pages 4483–4499.

Xiujun Li, Yu Wang, Siqi Sun, Sarah Panda, Jingjing Liu, and Jianfeng Gao. 2018. Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. arXiv preprint arXiv:1807.11125.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of EMNLP 2020*, pages 6008–6018.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. MinTL: minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of EMNLP 2020*, pages 3391–3405.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling. *CoRR*, abs/2106.02787.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad 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. In *Proceedings of EMNLP 2016*, pages 2122–2132.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1297–1303.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of AAAI 2020*, pages 8433–8440.

Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2022. Cross-lingual dialogue dataset creation via outline-based generation. *arXiv preprint arXiv:2201.13405*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. DialoGLUE: A natural language understanding benchmark for task-oriented dialogue. arXiv preprint arXiv:2009.13570.

Nikita Moghe, Mark Steedman, and Alexandra Birch. 2021. Cross-lingual intermediate fine-tuning improves dialogue state tracking. In *Proceedings of EMNLP 2021*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of EMNLP 2021*, pages 2241–2252.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP 2020*, pages 2362–2376.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2206–2217.

Evgeniia Razumovskaia, Goran Glavas, Olga Majewska, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *CoRR*, abs/2104.08570.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of EMNLP 2018*, pages 2780–2786.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of AAAI 2020*, pages 8854–8861.

Pei-Hao Su, Nikola Mrkšić, Iñigo Casanueva, and Ivan Vulić. 2018. Deep learning for conversational AI. In *Proceedings of NAACL-HLT 2018: Tutorial Abstracts*, pages 27–32.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 183–187.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of ACL 2010*, pages 917–926.

Tsung-Hsien Wen, Pei-Hao Su, Paweł Budzianowski, Iñigo Casanueva, and Ivan Vulić. 2019. Data collection and end-to-end learning for conversational AI. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL 2017*, pages 438–449.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of EMNLP 2020*, pages 917–929.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of EMNLP 2020*, pages 5052–5063.

Steve Young. 2010. Still talking to machines (cognitively speaking). In *Proceedings of INTERSPEECH*, pages 1–10.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. Allwoz: Towards multilingual task-oriented dialog systems for all. *CoRR*, abs/2112.08333.