Crossing the Conversational Chasm: 
A Primer on Multilingual Task-Oriented Dialogue Systems

Evgeniia Razumovskaia♣ Goran Glavaš♦ Olga Majewska♣ Anna Korhonen♣ Ivan Vulić♣
♣Language Technology Lab, University of Cambridge, UK
♦Data and Web Science Group, University of Mannheim, Germany
PolyAI Limited, UK
{er563,om304,alk23,iv250}@cam.ac.uk
goran@informatik.uni-mannheim.de

Abstract
Despite the fact that natural language conversations with machines represent one of the central objectives of AI, and despite the massive increase of research and development efforts in conversational AI, task-oriented dialogue (ToD) – i.e., conversations with an artificial agent with the aim of completing a concrete task – is currently limited to a few narrow domains (e.g., food ordering, ticket booking) and a handful of major languages (e.g., English, Chinese). In this work, we provide an extensive overview of existing efforts in multilingual ToD and analyse the factors preventing the development of truly multilingual ToD systems. We identify two main challenges that combined hinder the faster progress in multilingual ToD: (1) current state-of-the-art ToD models based on large pretrained neural language models are data hungry; at the same time (2) data acquisition for ToD use cases is expensive and tedious. Most existing approaches to multilingual ToD thus rely on (zero- or few-shot) cross-lingual transfer from resource-rich languages (in ToD, this is basically only English), either by means of (i) machine translation or (ii) multilingual representation spaces. However, such approaches are currently not a viable solution for a large number of low-resource languages without parallel data and/or limited monolingual corpora. Finally, we discuss critical challenges and potential solutions by drawing parallels between ToD and other cross-lingual and multilingual NLP research.

1 Introduction and Motivation
Endowing machines with the ability to intelligently converse with humans has been one of the fundamental objectives in the pursuit of artificial intelligence. As compelling as it is challenging, developing dialogue systems capable of satisfying the end user on a par with human-human interaction remains an elusive target. Narrower in scope than general-purpose conversational assistants, task-oriented dialogue (ToD) systems (Gupta et al., 2005; Bohus and Rudnicky, 2009; Young et al., 2013; Muise et al., 2019) have attracted both scientific and business interest as a far more feasible application, with potential to help or altogether replace human operators in focused problems and areas such as restaurant booking (Kim and Banchs, 2014; Henderson et al., 2019a), banking (Hardy et al., 2004; Altinok, 2018), travel (Li et al., 2018; Zang et al., 2020), or healthcare (Laranjo et al., 2018; Denecke et al., 2019).

The accelerated pace at which new milestones are reached across natural language applications thanks to the growing viability of deep learning techniques has recently catalysed dialogue-oriented research (Ren et al., 2018; Wen et al., 2019; Henderson et al., 2020; Wu et al., 2020, inter alia). Coupled with the proliferation of affordable voice technology (e.g., Amazon Alexa, Google Assistant, Microsoft Cortana, Samsung Bixby), the so-far distant prospect of virtual assistants becoming part of everyday reality seems more attainable than ever. And yet, the momentum of developments in this area has mainly targeted a very small proportion of their potential beneficiaries, further deepening the chasm in accessibility of state-of-the-art language technology between speakers of dominant and under-represented languages. Extending the reach of conversational technology is crucial for democratisation of human-machine communication and requires focusing research efforts on developing approaches that generalise across diverse language varieties and linguistic phenomena, are robust to cross-cultural differences in dialogue.

1For example, Amazon Alexa, one of the most popular personal assistants, currently supports only eight resource-rich languages: English, French, German, Hindi, Italian, Japanese, Brazilian Portuguese, and Spanish.
behaviours, and efficiently capitalise on available training data, the scarcity of which continues to be one of the major obstacles on the way to truly multilingual conversational AI.

In this survey, we take stock of the work carried out to date on multilingual ToD, discuss the main open challenges and lay out possible avenues for future developments. In particular, we aim to systematise the current research and know-hows related to multilingual ToD, and shed new light on the following crucial topics:

(Q1) What methods have been applied to multilingual ToD to date; how can we incorporate language-specific information and conduct target-language adaptation into the current methods?

(Q2) What are the additional difficulties when developing ToD systems in a number of different target languages with their semantic and structural variation and differences?

(Q3) What ToD datasets, in languages other than English as well as multilingual, are available and what are their strengths and weaknesses?

(Q4) Which components of ToD systems rely on cross-lingual capabilities the most?

(Q5) What are the critical future challenges, and how can multilingual ToD borrow from other related fields of NLP research to better tackle them?

Despite recent positive trends and a slowly but steadily growing body of work on creating multilingual ToD data and methodology, our survey suggests that the pace of multilingual ToD research still lags behind other cross-lingual NLP work and niche NLP tasks (e.g., named entity recognition, dependency parsing, QA) when it comes to linguistic diversity, training and evaluation data availability, cross-lingual transfer methodology, joint multilingual modeling, etc. (Ponti et al., 2019a; Hedderich et al., 2021). We hope that this survey will inspire more work in this area, attempting at drawing direct links (including similarities and differences) between ToD sub-tasks and other cross-lingual NLP research, which could enable the use and adaptation of existing techniques for multilingual and cross-lingual ToD tasks, and 2) aiming to indicate the current lack of training and evaluation resources for a large number of languages and domains.

2 Task-Oriented Dialogue Systems

The purpose of ToD systems, prevalent in practical applications, is to allow users to complete a concrete task through conversational interaction with the system (Young, 2010; Chen et al., 2017; Su et al., 2018). The tasks are typically well-defined and commonly have a binary outcome, i.e., the task was either successfully completed through communication with the system or it was not. Common examples include booking use cases (restaurants, transportation, hotels), automation of customer support (e.g., in domains like banking or telecommunications), or retrieving and providing information (e.g., in healthcare or tourism). For completeness, we first provide a concise overview of the two existing approaches to task-oriented dialogue: (i) modular approach, in which ToD is broken down into a pipeline of subtasks and (ii) end-to-end ToD, where a single neural model is trained to generate responses based on the preceding context.

2.1 Modular Task-Oriented Dialogue

A modular approach to ToD addresses the complexity of the task by breaking it down into a sequence of subtasks. The solution, as depicted in Figure 1, is a pipeline of independently trained models (i.e., components): the discrete output of a preceding component in the pipeline serves as the input to the next. In this work, we focus our attention on dialogue systems that operate on text input and generate text output – such systems are then extendable to true conversational systems by prepping an automatic speech recognition (ASR, speech-to-text) component to the beginning of the pipeline and appending a speech synthesis (i.e., TTS, text-to-speech) component to its end. The three core text-based components of each modular ToD system are: natural language understanding

Figure 1: The typical architecture of a modular dialogue system. The gray rectangle spans the modules operating on text, which are in the focus of this survey.
(NLU), dialogue (policy) management (PM), and response generation (RG), outlined in what follows.

**Natural Language Understanding (NLU).** In the context of ToD systems,³ natural language understanding refers to the recognition of the crucial goals and information from the user’s utterances. It usually encompasses two subtasks, namely *intention classification* (also known as *dialogue act classification*) (Ravuri and Stolcke, 2015; Khanpour et al., 2016) and *slot filling* (also known as slot labelling or slot tagging) (Mesnil et al., 2014; Kurata et al., 2016). The former is a single-label or a multi-label classification task that assigns one or more intent labels to the whole user utterance, whereas the latter extracts values for specific informational slots expressed in the utterance. For example, the utterance “show flights from Boston to New York today” has the intent class *FindFlight* and specifies values for three informational slots – departure location: Boston; arrival location: New York; and time: today). Given that slot appearance depends on the utterance intent, the two tasks are often addressed jointly via multi-task learning (Xu and Sarikaya, 2013; Guo et al., 2014; Goo et al., 2018; Chen et al., 2019; Wu et al., 2020, *inter alia*).

Traditionally, ToD systems included a component for *dialogue state tracking (DST)*, considered to be in between NLU and dialogue management. The purpose of DST models (Henderson et al., 2014b; Mrkšić et al., 2017a; Perez and Liu, 2017; Zhong et al., 2018, *inter alia*) is to maintain the *dialogue belief state*, a discrete or probabilistic summary of the dialogue history, encompassing all user goals and slot values expressed by the user throughout the conversation. Input to DST at each user turn consists of the previous belief state and the output of intent classification and slot filling modules; the output is the new/updated belief state. More recently, however, attention-based Transformer models (Vaswani et al., 2017; Devlin et al., 2019), with their ability to encode long sequences and capture long-distance semantic dependencies, allowed to build latent representations of dialogue history (from scratch) at every turn. This removed the need for maintaining an explicit belief state, and consequently, eliminated DST from many recent ToD systems (Wolf et al., 2019; Budzianowski and Vulić, 2019). Despite its diminished importance in more recent Transformer-backed ToD systems, for completeness we still provide a brief overview of DST in multilingual ToD later in §3.

**Dialogue (Policy) Management (PM)** refers to a ToD component responsible for choosing the system actions based on the current dialogue state. Approaches to PM can be broadly categorized into rule-based, supervised, and those based on reinforcement learning (RL) (Su et al., 2018). RL-based PM has been the predominant paradigm in recent years – it is more flexible than rules and does not require utterance-level annotations like supervised learning. It does, however, require a large number of conversations with the final outcome label (e.g., successful or not successful) as reward/penalty for RL. This has directed the research efforts towards simulations of user interactions with the policy manager (El Asri et al., 2016; Cuayahuitl, 2017; Cao et al., 2020b). PM models are agnostic to the dialogue language – they receive an abstracted representations of the dialogue state from NLU and/or DST and produce an abstract action representation for the response generator; because of this, PM is not of particular interest in the context of multilingual ToD, that is, it inherits all the challenges and solutions directly from monolingual PM research.

**Response Generation (RG)** is a module in charge of producing the system utterances, i.e., responses to the user utterances, given a system action predicted by the policy manager. Much like early PM, early RG efforts relied on templates and rules hand-crafted by domain experts (Langkilde and Knight, 1998; Stent et al., 2004; Cheyer and Guzzoni, 2006; Mirkovic and Cavedon, 2011, *inter alia*). More recent data-driven approaches exploit ever-growing corpora of online human-human conversations (e.g., Reddit, Quora, Twitter) and produce system responses by either (1) generating natural language utterances (e.g., Sordoni et al., 2015; Li et al., 2016b; Wen et al., 2017; Zhang et al., 2018b; Zhu et al., 2019; Peng et al., 2020) or (2) retrieving the most suitable response from a predefined set of candidate replies, also referred to as response selection (e.g., Lowe et al., 2017a; Yang et al., 2018; Zhang et al., 2018c; Henderson et al., 2019b).

Retrieval methods, on the one hand, offer the ad-
vantages of fluency, grammatical correctness and high-quality of the replies; modern neural natural language generation generation (NLG), in contrast, often produce overly general, incoherent, and grammatically erroneous utterances (Li et al., 2016a; Gao et al., 2018; Serban et al., 2016b). On the other hand, reliance on fixed lists of candidate responses constrains the versatility of responses, making response quality of selection based approaches highly dependent on the size of the response inventory (i.e., corpus of human-human interactions). Hybrid methods combine the best of both worlds (Song et al.; Weston et al., 2018; Pandey et al., 2018; Yang et al., 2019): they first retrieve a set of response candidates and then provide them, together with the user utterance (or wider dialogue history), as input to a generative model, which then produce the final system response.

2.2 End-to-End Task-Oriented Dialogue (e2e)

Components of a modular ToD system are trained in isolation, i.e., the later pipeline components are not exposed to errors of earlier models at training time and, consequently, cannot compensate for those errors at inference. To remedy for this well-known error cascading issue of pipeline learning systems, end-to-end ToD relies on neural architectures (Wen et al., 2017; Liu et al., 2018; Qin et al., 2020). Some e2e models mirror the modules of the traditional pipeline (Wen et al., 2017), parameters of which are all jointly tuned in one training procedure. On the one hand, end-to-end training does address the component mismatch and error propagation issues of modular ToD. On the other hand, e2e models aim to capture complex interactions between intents, policies, and responses in a latent representation space: this typically requires a large number of model parameters, reliable estimation of which requires large amounts of conversations. Requiring large training data, E2E models have been much more successful in open-domain conversations (i.e., chat bots) (Serban et al., 2016a; Lowe et al., 2017b; Adiwardana et al., 2020; Zhang et al., 2020, inter alia) than in ToD.

2.3 Why is Developing Multilingual Dialogue Systems Difficult?

Subtasks of the modular ToD systems can be seen as specific instances of general NLP classes of problems, e.g., intent classification is a short-text classification task, whereas slot-filling can be seen as a sequence-labelling task, or even recast as a span extraction or a question answering task (cf.§4.2). The best performance in such tasks is obtained with supervised machine learning models. Truly addressing multilingualism, thus entails annotated data for most human languages, for each task of interest. The fact that collecting labeled data for most human languages is not feasible is the central bottleneck of multilingual NLP (Joshi et al., 2020); most existing datasets for higher-level language understanding and reasoning tasks (Conneau et al., 2018; Hu et al., 2020; Ponti et al., 2020) have training portions only in English. The fact that ToD most commonly entails a pipeline of supervised models makes the prospect of truly multilingual ToD several times more challenging: for optimal ToD for a given language, one would need to acquire language-specific annotations for each of the pipeline tasks (i.e., intent detection, slot filling, response selection and/or response generation).

Absence of language-specific annotations for most languages directed research efforts towards cross-lingual transfer: models trained on labeled data in a resource-rich language are used to make predictions for texts in resource-lean languages with few or no annotations. Successful cross-lingual transfer, however, requires abstracting over linguistic (i.e., typological) properties that vary across languages and is therefore generally easier to achieve between typologically and etymologically closer languages (Lin et al., 2019; Lauscher et al., 2020). Cross-lingual word embeddings (Ruder et al., 2019; Glavaš et al., 2019) and massively multilingual transformers (MMTs) (Devlin et al., 2019; Conneau et al., 2020b) have been the recent vehicles for cross-lingual transfer of NLP models. While MMTs have initially been particularly praised for their transfer capabilities (Pires et al., 2019; Wu and Dredze, 2019), recent work has shown that their effectiveness drastically drops in transfers to distant languages and/or languages represented with small-sized monolingual corpora in multilingual pretraining of these models (Lauscher et al., 2020).

While not feasible for low-resource languages, cross-lingual transfer with MMTs does seem to be effective for closely related languages with large monolingual corpora (Pires et al., 2019; Wu and Dredze, 2019). It could therefore represent a viable solution for task-oriented dialogue in major languages close to English (e.g., German, Italian, French, Spanish). Given the abundance of parallel
data between English and these major languages, another viable solution for ToD in those languages is addition of machine translation modules to the pipeline (i.e., from target language to English before NLU and from EN to target language after RG). Although conceptually feasible, there is only anecdotal evidence for effectiveness of these transfer approaches (Schuster et al., 2019a; Liu et al., 2019b), primarily due to the lack of multilingual ToD evaluation datasets. Creation of multilingual ToD evaluation datasets across diverse languages, such as the most recently published Multi-ATIS++ (Xu et al., 2020), and also across various domains, is thus necessary for a reliable estimate of feasibility of translation- and transfer-based approaches to multilingual ToD. With the current limitations, it also remains largely unknown how these different approaches compare against each other, and which method should be preferred in relation to particular classes of ToD-related tasks, languages, and domains one deals with.

3 Existing Efforts in Multilingual and Cross-Lingual ToD

We now provide an overview of existing efforts in multilingual ToD as well as cross-lingual transfer for ToD, focusing on each component of modular ToD (§3.1–3.3), and then on e2e ToD (§3.4).

3.1 Natural Language Understanding (NLU)

Joint versus Separate Training. NLU approaches can be divided into two groups depending on whether they tackle intent classification and slot filling (i) jointly, in multi-task training regimes (Schuster et al., 2019a; Liu et al., 2019b; Xu et al., 2020; Bunk et al., 2020, inter alia) or (ii) independently, addressing only one of the tasks or training an independent model for each of them (Ren and Xue, 2020; He et al., 2020; Arora et al., 2020, inter alia). Joint multi-task training, besides potentially reducing the number of parameters, is advantageous for NLU (Zhang et al., 2019b), as, the two tasks are clearly interdependent: intuitively, the slots for which the values may be provided in an utterance also depend on the intent of the utterance.

Transfers via MMTs. Given the absence of training-size data in other languages, the default approach to multilingual NLU is (zero-shot or few-shot) transfer of models trained on English datasets by means of pretrained massively multilingual transformers (Zhang et al., 2019b; Xu et al., 2020; Siddhant et al., 2020b; Krishnan et al., 2021). While most of the work relies on MMTs pretrained via language modeling objectives, e.g., mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020b), Siddhant et al. (2020b) show that a massively multilingual encoder trained via machine translation (MMTE) allows for a more effective zero-shot transfer for intent classification than mBERT. The most recent work of Liu et al. (2021) shows that an mBERT-based sequence labelling overfits to the word order of the source language and that regularizing for word order information (e.g., by removing positional embeddings or by shuffling tokens) leads to better transfer to languages with different word order for various sequence labelling tasks, including ToD slot filling.

Cross-Lingual Supervision. From the perspective of cross-lingual supervision, MMTs are unsupervised models, as their pretraining does not require any explicit alignment between the languages. Representation subspaces of individual languages can, however, be further/better aligned with explicit supervision in the form of word or sentence alignments (i.e., parallel data) (Cao et al., 2020b; Conneau et al., 2020c), leading to better cross-lingual transfer in downstream tasks. Aiming to improve for cross-lingual alignment of mBERT’s representations Kulshreshtha et al. (2020) systematically compare (a) projection-based vs. fine-tuning-based alignment methods driven by (b) cross-supervision in the form of word translations versus sentence translations. Their zero-shot transfer results on three ToD slot-filling datasets (Upadhyay et al., 2018; Schuster et al., 2019a; Bellomaria et al., 2019) and five target languages (Hindi, Turkish, Spanish, Thai, and Italian) indicate that fine-tuning based on word alignments is most consequential for zero-shot transfer.

Earlier work leveraged static cross-lingual word embedding spaces (CLWEs) (Mikolov et al., 2013; Smith et al., 2017; Artetxe et al., 2018; Joulin et al., 2018; Patra et al., 2019; Glavaš and Vulić, 2020, inter alia) as a mechanism for cross-lingual transfer of NLU models (Upadhyay et al., 2018; Chen et al., 2018; Schuster et al., 2019a). Reliable induction of useful CLWEs requires at least a few hundred word translation pairs (Vulić et al., 2019). Relying on limited in-domain word-level supervision, i.e., a small number (e.g., 10) of word alignments in

\[ \text{word translation pairs} \]
the actual ToD domain of interest, either for code-switching of the English training data (Liu et al., 2020b) or for refinement of the CLWE space (Liu et al., 2019b) can further improve the zero-shot transfer performance in NLU tasks.

Available Datasets. The scope of existing studies on multilingual and cross-lingual NLU has primarily been defined by the availability of multilingual datasets for model training and evaluation. While there are arguably more resources for multilingual NLU than for other tasks in modular ToD, the landscape of existing datasets is still very sparse. We provide an overview of monolingual NLU datasets in languages other than English in Table 1 and multilingual NLU datasets in Table 2. Existing NLU datasets in other languages have been obtained by translating original English datasets or some of their portions: Castellucci et al. (2019) translated the SNIPS dataset (Coucke et al., 2018) to Italian; Susanto and Lu (2017); Upadhyay et al. (2018); Xu et al. (2020) translated the ATIS dataset in 10 different languages – however, only the most recent effort of (Xu et al., 2020) introduced a diverse set of languages from different language families and with varying typological properties.

Truly general NLU models would need to generalize over both languages and domains. Most existing datasets, however, either cover multiple domains (Hakkani-Tür et al., 2016; Liu et al., 2019a) monolingually or the same domain across different languages (Xu et al., 2020), preventing the investigation of true generalizability of current cross-lingual transfer approaches for NLU.

3.2 Dialogue State Tracking (DST)

As discussed in §2.1, DST has recently lost much of its significance for modular ToD due to ability of Transformer-based models to capture long distance dependencies and model the entire dialogue history. For completeness, we briefly summarize the existing multilingual DST datasets and cross-lingual DST approaches, predominantly based on cross-lingual word embeddings.

Cross-Lingual Transfer Models. Neural Belief Tracker (NBT) (Mrkšić et al., 2017a; Mrkšić and Vulić, 2018) is a neural DST approach that estimates the user’s goal at every step of the dialogue. It learns representation of each slot-value pair and compares them with utterances in order to determine if a slot-value pair is mentioned. It was the first fully data-driven DST model which performed on-par with the models exploiting hand-crafted lexical rules. With the introduction of the multilingual WoZ dataset (see later), Mrkšić et al. (2017b) coupled NBT with cross-lingual word embeddings to enable zero-shot cross-lingual DST transfer. A body of subsequent work on specializing CLWEs for semantic similarity reported performance gains in cross-lingual transfer for DST, using NBT as the base model (Vulić et al., 2018; Glavaš and Vulić, 2018; Ponti et al., 2018b, 2019b). XL-NBT Chen et al. (2018) adapts NBT for target languages via multilingual knowledge distillation (Hinton et al., 2015): the DST knowledge of the English teacher is transferred to the target language student model by means of matching representations for parallel instances – word and sentence translations. The results of the most recent DSTC 9 challenge (Gunasekara et al., 2020), indicate, however, that training state-of-the-art monolingual DST models (Shan et al., 2020; Kim et al., 2020a) on machine translated training data in the target language outperforms the zero-shot and few-shot cross-lingual transfer of source language DST models. It is worth noting, that DSTC 9 includes only English and Chinese, major languages with huge monolingual corpora and abundance of parallel data between them. The translation-based approach to cross-lingual DST transfer would not be nearly as effective for low-resource languages.

Available Datasets and Benchmarks. Mrkšić et al. (2017b) translated the WoZ 2.0 DST dataset (Wen et al., 2017) to German and Italian. Within the dedicated Dialogue State Tracking Challenge (later renamed to Dialog System Technology Challenges), only 3 out of 9 editions to date included multilingual DST tracks. DSTC 5 (Kim et al., 2016) tested DST models in zero-shot cross-lingual transfer from English (training data) to Chinese (development and test data) on the data in the tourism domain. DSTC 6 (Hori et al., 2019) included a track on dialog breakdown detection in chat-oriented dialogues, design the test cross-lingual transfer abilities of breakdown detection models, from English to Japanese. Finally, as the first challenge to test the first to test cross-lingual DST systems on large scale datasets, DSTC 9 (Gunasekara et al., 2020) included a track testing the transfer between English and Chinese (in both directions), using MultiWOZ 2.1 (Eric et al., 2020) as the En-
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task(s)</th>
<th>Language(s)</th>
<th>Domains</th>
<th>Size</th>
<th># intents</th>
<th># slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-English monolingual datasets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDIA (Bonneau-Maynard et al., 2005)</td>
<td>slot extraction</td>
<td>fr</td>
<td>hotel reservations</td>
<td>15000</td>
<td>N/A</td>
<td>83</td>
</tr>
<tr>
<td>SLU-IT (Castellucci et al., 2019)</td>
<td>intent classification; slot extraction</td>
<td>it</td>
<td>7 domains, inter alia, music, weather, restaurant</td>
<td>7142</td>
<td>7</td>
<td>39</td>
</tr>
<tr>
<td>Almawave-SLU (Bellomaria et al., 2019)</td>
<td>intent classification; slot extraction</td>
<td>it</td>
<td>7 domains, inter alia, music, weather, restaurant</td>
<td>14484</td>
<td>7</td>
<td>39</td>
</tr>
<tr>
<td>(Zhang et al., 2017)</td>
<td>intent classification</td>
<td>zh</td>
<td>chat; task-oriented</td>
<td>4000</td>
<td>31</td>
<td>N/A</td>
</tr>
<tr>
<td>ECSA dataset (Gong et al., 2019)</td>
<td>slot extraction; named entity extraction</td>
<td>zh</td>
<td>online commerce</td>
<td>27615</td>
<td>N/A</td>
<td>N/A (sequence tags)</td>
</tr>
<tr>
<td>Chinese ATIS (He et al., 2013)</td>
<td>intent classification; slot extraction</td>
<td>zh</td>
<td>airline travels</td>
<td>5871</td>
<td>21</td>
<td>120</td>
</tr>
<tr>
<td>Vietnamese ATIS (Dao et al., 2021)</td>
<td>intent classification; slot extraction</td>
<td>vi</td>
<td>airline travels</td>
<td>5871</td>
<td>25</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 1: Monolingual NLU datasets. The table includes only non-English datasets. A non-exhaustive list of English datasets is provided in Table 7 in the Appendix.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task(s)</th>
<th>Languages</th>
<th>Domains</th>
<th>Size</th>
<th># intents</th>
<th># slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilingual TOP (Schuster et al., 2019a)</td>
<td>intent classification; slot extraction</td>
<td>en, es, th</td>
<td>Alarm, reminder, weather</td>
<td>43323 [en]</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>ATIS in Chinese and Indonesian (Susanto and Lu, 2017)</td>
<td>semantic parsing; slot extraction</td>
<td>en, zh, id</td>
<td>airline travels</td>
<td>5371</td>
<td>N/A</td>
<td>120 (166; λ-calculus)</td>
</tr>
<tr>
<td>Multilingual ATIS (Upadhyay et al., 2018)</td>
<td>intent classification; slot extraction</td>
<td>en, hi, tr</td>
<td>airline travels</td>
<td>1493 [hi]</td>
<td>21</td>
<td>120</td>
</tr>
<tr>
<td>MultiATIS++ (Xu et al., 2020)</td>
<td>intent classification; slot extraction</td>
<td>en, es, pt, de, fr, zh, ja, hi, tr</td>
<td>airline travels</td>
<td>5871 [en]</td>
<td>18; 17 [hi], 17 [tr]</td>
<td>84 consumers</td>
</tr>
<tr>
<td>MTOP (Li et al., 2021)</td>
<td>semantic parsing; intent classification; slot extraction</td>
<td>en, de, fr, es, hi, th</td>
<td>11 domains, inter alia, music, news, recipes</td>
<td>18788 [en]</td>
<td>117</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 2: Multilingual NLU datasets.

glish dataset and CrossWOZ (Zhu et al., 2020a) as the Chinese dataset. Table 3 summarizes the multilingual DST datasets.

### 3.3 Natural Language Generation (NLG)

In contrast to other tasks of modular ToD, multilingual response generation for ToD has not undergone the same rate of progress. We thus take a broader look at multilingual NLG in general.

**Traditional NLG.** General work aimed at producing natural language responses in languages other than English from non-linguistic data or meaning representations traditionally relied on pipeline NLG architectures (Reiter and Dale, 1997; Ehud and Robert, 2000). In such pipelines, a linguistic (surface) realiser serves as the last module responsible for outputting the final surface text based on language-specific morpho-syntactic and orthographic requirements (e.g., word order, inflectional morphology). Approaches to linguistic realisation include hand-crafted grammar-based systems (Gatt and Krahmer, 2018; Bateman, 1997; Elhadad and Robin, 1996), manually created templates (McRoy et al., 2003), and statistically driven methods (Filippova and Strube, 2007). To facilitate the general usage of grammar-based systems, characterised by high level of linguistic detail, simpler realisation engines that provide syntax and morphol-
Table 3: Multilingual DST datasets. Abbreviations: H2M – human-to-machine; H2H – human-to-human. A non-exhaustive list of English DST datasets is given in Table 8 in the Appendix.

Table 4: Multilingual datasets for end-to-end training.

Translation-Based Methods. Given the reliance of data-driven NLG models on the availability of training data and its scarcity in the vast majority of world languages, cross-lingual transfer methods have been leveraged to enable NLG in low-resource scenarios. To this end, machine translation (MT) has been employed to either (i) translate the target language input to English, feed it into an NLG system trained on English data, and subsequently translate the generated English text back to the target language (Wan et al., 2010), or (ii) translate
the English training data into the target language and train the NLG model in the target language (Shen et al., 2018; Duan et al., 2019).

**Multilingual Pretraining** of sequence to sequence models, popular in the most recent machine translation research (Liu et al., 2020a; Lin et al., 2020a; Kim et al., 2020b), has been successfully applied in cross-lingual transfer of NLG models in other applications as well. For example, Kumar et al. (2019) first pretrain a bilingual English-Hindi language model on monolingual corpora of both languages via denoising autoencoding and back-translation; they then fine-tune the model for question generation using a large English and small Hindi dataset.

Similarly, Chi et al. (2020) pretrain the multilingual Transformer-based encoder and decoder of the seq2seq model on monolingual corpora of multiple languages, using denoising autoencoding but additionally leverage parallel sentences for cross-lingual masked language modelling training. They then demonstrate the effectiveness of the pretrained multilingual seq2seq model in cross-lingual transfer for two downstream NLG tasks: question generation and abstractive summarization. Recent work has also explored the power of massively multilingual transformers to boost NLG performance across languages. Adopting the approach of Wang and Cho (2019), Rönqvist et al. (2019) evaluated mBERT (Devlin et al., 2019) on NLG tasks in English, German, Danish, Finnish, Norwegian (Bokmal and Nynorsk) and Swedish, and found that: (1) mBERT significantly underperforms monolingual counterparts for English and German and (2) it cannot handle the morphological complexity of Nordic languages, given that its subword vocabulary and representations are shared across 104 languages. In sum, while multilingual pretraining looks promising for NLG for resource-rich languages, it does not seem to be a viable solution for lower-resource languages with smaller amounts of unlabeled text.

**Available Datasets.** Training data for NLG in languages other than English is still very limited: there are small datasets in Korean (Chen et al., 2010), Spanish (García-Méndez et al., 2019), and Czech (Dušek and Jurčíček, 2019). There exist also structured data-to-text datasets for German and French (Nema et al., 2018) and image-to-description datasets in Chinese (Li et al., 2016c) and Dutch (van Miltenburg et al., 2017, 2018), as well as cross-lingual English-German data (Elliott et al., 2016).

### 3.4 End-to-End Dialogue

Lately, there has been increased interest in end-to-end dialogue modelling. Most of the algorithms utilise a sequence-to-sequence framework to generate system responses (Wen et al., 2017; Madotto et al., 2018; Ham et al., 2020). Unfortunately, training reliable seq2seq models requires large amounts of training data, making end-to-end dialogue systems data hungry. As a result, combined with the fact that collecting task-oriented dialogues is much more expensive than collecting open-domain conversations for training chatbots, there have been only a few monolingual end-to-end ToD efforts in languages other than English: Zhu et al. (2020a) and Quan et al. (2020) explore e2e ToD in Chinese and German. We list the available datasets for multilingual e2e ToD in Table 4. Although this survey focuses on ToD, for the sake of completeness, we additionally list available datasets for multilingual open-domain dialogue.

### 4 Challenges, Solutions, Outlook

#### 4.1 Linguistic Diversity

A long-term development of multilingual ToD systems for diverse languages will be driven by our ability to also evaluate on representative language samples: in turn, such language-representative evaluations would guarantee that developed systems can generalise to languages not present in the evaluation sets, but which come from similar language families, or display similar typological properties (Ponti et al., 2019a). The purpose of multilingual datasets is to assess the expected performance of a model across languages (Hu et al., 2020; Liang et al., 2020). If all the languages are similar, cross-lingual transfer is simplified and we can obtain overly optimistic performance (Ponti et al., 2019a). Thus, the languages within any multilingual dataset should ideally be linguistically diverse.

NLU is the only component which has multiple multilingual datasets. We assess the linguistic diversity of those datasets leveraging the language sample diversity metrics proposed by Ponti et al. (2020) to assess the typological, family and geographical diversity. The sample diversity scores

---

6To measure typological diversity, we calculate the mean of entropy of 103 binary URIEL (Litell et al., 2017) features for each language. The features are based on linguistic phenomena recorded in World Atlas of Language Structures (Dryer and Haspelmath, 2013). To measure family diversity, the number of distinct families in the dataset is divided by the total number of languages in the dataset. To measure ge-
are shown in Table 5.

For comparison, we include the most diverse datasets for other NLP tasks, e.g., natural language inference (XNLI; Conneau et al., 2018), causal commonsense reasoning (XCOPA; Ponti et al., 2020). It becomes apparent that the existing multilingual dialogue NLU datasets have several shortcomings. First, we observe that they cover only a single macroarea: Eurasia. Second, the languages within the multilingual datasets for dialogue NLU do not cover the full variety of linguistic phenomena. In particular, the languages are identical with respect to 23 out of 103 typological URIEL features.

Moreover, there is currently a crucial lack of multilingual and cross-lingual dialogue datasets beyond NLU, with full-fledged multi-turn conversations in diverse multiple languages (see Table 4). Such datasets, if they existed, would enable end-to-end training, truly multi-turn ToD systems in multiple languages which also need to (learn to) leverage dialogue history and a wider dialogue context. Finally, such datasets with comparable examples would also enable comparative analyses between different languages, widening our understanding of the critical ToD-related cross-linguistic similarities and differences. In conclusion, we see a potential future direction in dataset collection for multilingual ToD in languages which cover a much wider set of language families, macroareas and linguistic phenomena.

### 4.2 Coping with Low-Resource Scenarios

As discussed in §3, intent detection is a standard classification task, which can also be recast as a question answering task (Namazifar et al., 2020), while slot filling can be framed as standard sequence labeling (Louvan and Magnini, 2020) or a span extraction task (Coope et al., 2020; Hender-}

<table>
<thead>
<tr>
<th># languages</th>
<th>Multilingual TOP</th>
<th>Multilingual ATIS</th>
<th>MultiATIS++</th>
<th>MTOP</th>
<th>XCOPA</th>
<th>TyDi QA</th>
<th>XNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typology</td>
<td>0.196</td>
<td>0.257</td>
<td>0.343</td>
<td>0.267</td>
<td>0.41</td>
<td>0.41</td>
<td>0.39</td>
</tr>
<tr>
<td>Family</td>
<td>0.667</td>
<td>0.667</td>
<td>0.444</td>
<td>0.333</td>
<td>1.0</td>
<td>0.9</td>
<td>0.5</td>
</tr>
<tr>
<td>Macroareas</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.67</td>
<td>0.92</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 5: Assessment of typological and genealogical diversity of multilingual dialogue NLU datasets. Linguistically diverse datasets of several other NLP tasks shown for comparison: commonsense reasoning (XCOPA Ponti et al., 2020), natural language inference (XNLI; Conneau et al., 2018) and QA (TyDi QA; Clark et al., 2020). For the description of the three diversity measures, we refer the reader to (Ponti et al., 2020).
transfer performance has been achieved with multilingual Transformer-based language models such as multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020b), and mT5 (Xue et al., 2020), further applying (i) transfer techniques adapted to particular transfer directions (Schuster et al., 2019b; Conneau et al., 2020c; Cao et al., 2020a), truly low-resource languages (Pfeiffer et al., 2020a; Hederich et al., 2020; Üstün et al., 2020), even to those with unseen scripts (Pfeiffer et al., 2020b), (ii) few-shot learning with a small subset of target-language annotated examples (Lauscher et al., 2020; Bhattarchjee et al., 2020). However, current performance of cross-lingual transfer for low-resource languages (e.g., African languages, indigenous languages of the Americas) still cannot even remotely match transfer performance for high-resource target languages (Lauscher et al., 2020; Wu and Dredze, 2020; Zhao et al., 2020).

Pretrained language models can also be adaptively fine-tuned with unannotated in-domain data in both source and target language (Ponti et al., 2020) to pick up more domain-specific knowledge, which typically results in slight performance gains (Henderson et al., 2020; Gururangan et al., 2020). Along the same line, the entire research area focusing on domain adaptation in NLP (Kim et al., 2018; Ziser and Reichart, 2018; Rücklé et al., 2020; Ramponi and Plank, 2020) can also offer direct guidance on how to leverage high-resource ToD domains to boost performance in resource-lean ToD domains. Note that in multilingual setups, we might typically encounter extremely difficult “double-scarce” setups, simultaneously dealing with both low-resource domains and low-resource languages.

It might also be possible to make the multilingual ToD NLU models more robust in low-resource scenarios through data augmentation (Du et al., 2020; Xie et al., 2020): (i) at token level with synonymy-based substitutions generated automatically (Kobayashi, 2018; Gao et al., 2019) or taken from lexico-semantic resources (Raiman and Miller, 2017; Wei and Zou, 2019; Dai and Adel, 2020), or rule-based morphological inflection (Vulić et al., 2017; Vania et al., 2019), (ii) at sentence level with manipulating dependency trees (Ponti et al., 2018a; Şahin and Steedman, 2018), MT-based back-translation (Edunov et al., 2018), or generating synthetic adversarial examples (Garg and Ramakrishnan, 2020; Morris et al., 2020); (iii) at annotation level automatically labeling more sentences, filtering them, and using them as silver training data (Onoe and Durrett, 2019; Du et al., 2020). Slot tagging, as a sequence labeling task, might also profit from distant and weak supervision methods (Luo et al., 2017; Alt et al., 2019), often leveraged to boost structurally similar low-resource NER models (e.g., Cao et al., 2019; Mayhew et al., 2019; Lison et al., 2020). Similar principles can be applied to cross-lingual text classification with limited resources (Karamanolakis et al., 2020). Meta learning principles such as MAML (Finn et al., 2017) have also emerged recently as a means to deal with low-resource cross-lingual and cross-domain transfer for standard classification tasks (Nooralahzadeh et al., 2020; van der Heijden et al., 2021), but they are yet to find their true application in (multilingual) ToD systems.

**Source Selection.** Recently, there has been increased interest in zero-shot methods for multilingual NLU (Liu et al., 2020b; Xu et al., 2020; Krishnan et al., 2021). In the zero-shot scenario, a model trained on one or more languages (source languages) is tested on a target language different from the source. When porting a dialogue system to new languages, zero-shot transfer is an effective method to bypass costly data collection and annotation for every target language. However, we detect two crucial gaps which require more attention in future work: 1) similar to other research in cross-lingual NLP, it is highly likely that few-shot transfer will yield huge benefits over fully zero-shot transfer (Lauscher et al., 2020); 2) the actual source language(s) for zero-shot and few-shot cross-lingual transfer in low-resource scenarios may have a huge impact on the final task performance, as validated in other NLP areas (Zoph et al., 2016; Dabre et al., 2017; Lin et al., 2019), and more recently hinted for multilingual NLU (Krishnan et al., 2021). In other words, a standard go-to option of always transferring from English might be suboptimal for a large number of target languages.

In order to empirically verify and establish this conjecture, we conduct a simple empirical study validating that this holds for dialogue NLU, additionally showing that syntactic similarity is an especially strong predictor of zero-shot performance.

For the experiments, we rely on multilingual BERT fine-tuned for the two dialogue NLU tasks by adding classification layers on top. We fine-tune the model on every language available in Multi-
ATIS++ (see Table 2) and evaluate it on all languages excluding the language it was tuned on. In order to quantify language similarity, we use the cosine similarity between URIEL feature vectors (Littell et al., 2017), which capture typological (syntactic, phonological, phonetic), geographic and phylogenetic language properties. In order to measure the correlation between linguistic similarity and transfer performance, we compute Pearson’s r between language similarity and transfer performance on target languages.

The results on both tasks are summarised in Figure 2. We observe moderate and high correlation scores between source-target language similarity and task performance. In addition, an ablation study (see Table 6) looking at the correlation between zero-shot performance and similarity in different linguistic properties, reveals that syntactic similarity seems to play a particularly important role in choosing a suitable source. As a general finding, this small study suggests that: 1) carefully picking source languages, and 2) balancing

the annotation budget across and annotating more dialogue data for typologically diverse languages may steer and substantially improve dialogue NLU performance in the future.

4.3 Language Adaptation and Fluency

Besides coping with a wide semantic variability of user utterances in the NLU components, multilingual ToD systems also need to produce accurate and fluent responses fully adapted to the language at hand, and sounding native to the user. Given the dialogue context so far and constrained by the domain, the response generation module should output a response which is articulate and fits in the given context, without breaking the flow of the multi-turn conversation (Garbacea and Mei, 2020).

There are complexities which are common to Natural Language Generation tasks in general such as complex morphology (e.g., Slavic languages, agglutinative languages such Finnish or Turkish, polysynthetic languages). Generating fluent text in such morphologically rich languages is naturally much more complex than in morphologically poor ones (Kunchukuttan et al., 2014; Gerz et al., 2018). This is due to data scarcity stemming from the inability to hold all the possible word forms in the vocabulary or word “over-segmentation” with recent subword-based pretrained Transformers, which negatively affects word and sentence semantics (Rönnqvist et al., 2019; Rust et al., 2020). NLG in morphologically rich languages can benefit from dedicated language-specific tokenisers (Rust et al., 2020), from incorporating linguistic features (Klemen et al., 2020), or from multi-tasking (Passban et al., 2018), predicting the next word and morpho-

---

Table 6: Measuring correlation between zero-shot NLU transfer performance and source-target language similarity based on different groups of properties. Phonology and syntax features from URIEL (Littell et al., 2017) are based on the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). Phylogenetic features are based on Glottolog (Hammarström et al., 2017).

<table>
<thead>
<tr>
<th>Feature Group</th>
<th>Intent Classification</th>
<th>Slot Labelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonology</td>
<td>-0.2467</td>
<td>-0.2504</td>
</tr>
<tr>
<td>Geography</td>
<td>-0.2263</td>
<td>-0.3270</td>
</tr>
<tr>
<td>Phylogenetics</td>
<td>-0.4895</td>
<td>-0.6122</td>
</tr>
<tr>
<td>Syntax</td>
<td>-0.5131</td>
<td>-0.6919</td>
</tr>
</tbody>
</table>

Figure 2: Intent classification and slot labelling performance depending on linguistic source-target similarity.
logical information simultaneously.\footnote{Another problem in multilingual setups concerns different word orders, where NLG might be directly informed by typological information through structural adaptations (Ponti et al., 2018a), but currently there is little to no work at all coupling NLG and linguistic typology (O’Horan et al., 2016).}

Furthermore, there are linguistic phenomena specific for informal, conversational language, e.g., colloquialisms and code-switching. Code-switching is a phenomenon where interlocutors shift from one language to another during the conversation (Sankoff and Poplack, 1981). Previously, it was shown that code-switching might even improve task success of the system and its perceived friendliness (Ahn et al., 2020).

Banerjee and Khapra (2019) show that structure-aware generation is effective for code-switched data, even when dependency parsers are not available. With respect to modelling code-switched input, Khanuja et al. (2020) show that, to work in code-switching settings, cross-lingual models such as mBERT should be fine-tuned on code-switched data, as lexical distribution in code switched language is different from the union of two languages. Additionally, prior work shows that we obtain significant improvements on all dialogue tasks when large Transformer encoders are directly pretrained on conversational data (Henderson et al., 2020; Mehri et al., 2020). That means that in order to train ToD systems which can code switch we require large code-switching dialogue datasets which are not available yet.

Finally, language fluency and the more general user satisfaction, which concerns not only what the system responds, but also how it conveys information, cannot be entirely captured with fully automatic evaluation measures. This renders the need to conduct human-centered evaluations in order to really capture and trace any improvements in fluency and the user satisfaction with the general eloquence of ToD systems in different languages, leading us to the next challenge discussed in §4.4.

4.4 Evaluation of Multilingual ToD Systems

A crucial step in the development of ToD systems is evaluation (Deriu et al., 2021). For the modular ToD pipeline, there are standard automated metrics to evaluate each component, e.g., intent accuracy and slot $F_1$ for NLU or joint goal accuracy for DST. Recently, DialoGLUE (Mehri et al., 2020), a benchmark to evaluate ToD systems, has been proposed, but the benchmark is available only in English. Such benchmarks based on automated metrics are useful for evaluating progress on development of general ToD systems. We thus hope that, similar to recent benchmarks for general-purpose (i.e., non-ToD) cross-lingual NLU such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020), future work will look into building comprehensive and community-supported multilingual ToD benchmarks and services.

However, evaluation of ToD systems still adds another layer of difficulty, typically not present with other NLP tasks, such as the ones covered in XTREME or XGLUE. In short, previous work shows that strong performance on automated metrics does not always correlate with the overall user satisfaction with the system (Liu et al., 2016). This means that human evaluation is still the most reliable way to evaluate the ultimate dialogue system usability and usefulness.

Human evaluation for ToD systems aims to figure out whether the user was satisfied with the interaction (user satisfaction) or whether the system has completed the task (task success) (Deriu et al., 2021). Generally, human evaluations are costly and time-consuming. Multilingual ToD systems inherit the same costs but also pose new questions for commonly used human evaluation protocols. Firstly, hiring qualified users in many languages is even more expensive than in English. For lower-resource languages, the problem goes beyond the expenses: sometimes it is hard (or impossible) to find fluent speakers of those languages on commonly used platforms such as Amazon Mechanical Turk. Secondly, when hiring evaluators from different countries, one needs to consider whether there are cultural differences due to which user satisfaction could be altered.

4.5 Voice-Based Multilingual Dialogue?

This survey, following the current mainstream in monolingual and multilingual ToD modeling, has focused on text-based input. However, the assumption of working with clean, native input text is unrealistic: in fact, it underestimates the errors which can cascade from imperfect automatic speech recognition (ASR) to the subsequent text-based modules. While there is some monolingual English ToD work which pays attention to recovering from ASR errors and incorporating imperfect ASR output into subsequent text-based modules (Henderson et al., 2014a; Mrkšić et al., 2017a;
The crucial speech-to-text and text-to-speech bridges are typically overlooked in multilingual ToD research: this also means that we are currently overestimating the abilities of our voice-based ToD systems.

ASR and speech-to-text synthesis are wide research fields in their own right, also advocating an expansion towards multilinguality as a long-standing and crucial research goal (Le and Besacier, 2009; Ghoshal et al., 2013; Conneau et al., 2020a, inter alia). The current mainstream ASR paradigm also relies on transfer learning with large pretrained Transformer-based multilingual models (Conneau et al., 2020a; Pratap et al., 2020a). Similar to multilingual BERT or XLM-R in the text domain, a heavily parameterized multilingual ASR model is trained on a large multilingual speech corpus, and then fine-tuned with smaller amounts of speech data in particular target languages (Pratap et al., 2020b). Nonetheless, even this approach, termed 'massively multilingual' by Pratap et al. (2020a), spans only around 50 languages. A similar situation is observed with multilingual text-to-speech (TTS) research: despite recent efforts, multilingual TTS modules are available for a tiny fraction of languages (Zhang et al., 2019a; Nekvinda and Dusek, 2020), even smaller than what multilingual ASR currently supports.

This effectively means that voice-based ToD is still out of reach for the large majority of the world’s languages (Joshi et al., 2020). In the pursuit of wider-scale and democratized ToD technology, we advocate a tighter integration of speech-based and text-based modules in future work, as well as more realistic evaluation protocols which also include ASR and TTS error analyses. Any future developments of multilingual ToD are also tightly coupled with parallel developments in multilingual ASR and TTS as standalone research areas highly relevant to multilingual ToD.

4.6 Other Related Areas

We have attempted at covering multiple facets and research areas related to multilingual conversational AI, as an extremely wide multi-disciplinary and multi-layered field. However, we also acknowledge that there are other aspects associated with development and deployment of full-fledged and engaging multilingual ToD systems which remained out of our core focus. These other directions include (but are not limited to): 1) making ToD systems more adaptable and empathetic by relying on implicit conversational cues and (multilingual) emotion recognition (Pittermann et al., 2010; Heracleous et al., 2020; Meng et al., 2020); 2) incorporating the information from miscellaneous knowledge bases to improve the system’s commonsense reasoning and world knowledge capabilities (Eric et al., 2017; Madotto et al., 2018; Haihong et al., 2019); 3) grounding dialogue in perceptual (typically visual) contexts (de Vries et al., 2017; AlAmri et al., 2019; Shekhar et al., 2019; Agarwal et al., 2020).11 Stepping a bit further away, it is also quite intuitive that further developments in massively multilingual machine translation and MT for low-resource languages (Siddhant et al., 2020a,b; Garcia et al., 2020; Fan et al., 2020) will also (continue to) have substantial impact on the development of multilingual ToD.

5 Conclusion

Enabling machines to converse as humans is one of the central goals of AI, and achieving this in a multitude of the world’s languages is an even more complex challenge. In this work, we have presented an overview of the current efforts, which spans a survey on the existing methodology and available datasets, and future challenges concerning multilingual task-oriented dialogue (ToD) systems. We have also pointed at the main current limitations and gaps (e.g., a notoriously difficult and expensive dialogue data collection becomes even more difficult in multilingual scenarios), aiming to inspire more work in this important area. In the long run, we hope that our overview will fulfill its didactic purpose, as well as foster and guide future developments of multilingual ToD towards truly multilingual and inclusive conversational AI.

Finally, one by-product of this work, potentially useful to other researchers and practitioners interested in this emerging field, is an up-to-date overview of all the scoped monolingual and multilingual ToD datasets, which is available here.11

---

10On top of this, recent research has shown that ASR does not provide equitable service to native speakers of the same language from different backgrounds (Koennecke et al., 2020), which is yet another research problem in its own right.

11Besides providing the additional (situational) context to dialogues in general, multi-modal modeling might be even more useful in cross-lingual settings, since visual input (e.g., images, videos) also serve as naturally occurring interlingua (Kiela et al., 2015; Gella et al., 2017; Rotman et al., 2018; Sigurdsson et al., 2020).
Acknowledgments

Evgenia Razumovskaia is supported by a scholarship from Huawei. Ivan Vulić, Olga Majewska, and Anna Korhonen are supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no. 648909) and the ERC PoC Grant MultiConvAI: Enabling Multilingual Conversational AI (no. 957356). Goran Glavaš is supported by the Multi2ConvAI Grant (Mehrsprachige Grant MultiConvAI: Enabling Multilingual Domain-ubergreifende Conversational AI) of the Baden-Württemberg Ministry of Economy, Labor, and Housing.

References


Irene Langkilde and Kevin Knight. 1998. The practical value of n-grams is in generation. In *Natural Language Generation*.


Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2020. A closer look at few-shot crosslingual transfer:


## A English NLU datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Language</th>
<th>Domains</th>
<th>Size</th>
<th># intents</th>
<th># slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking-77</td>
<td>intent classification</td>
<td>en</td>
<td>banking</td>
<td>13083</td>
<td>77</td>
<td>N/A</td>
</tr>
<tr>
<td>(Casanueva et al., 2020)</td>
<td></td>
<td></td>
<td>10 domains, inter alia, banking, work, travel, small talk</td>
<td>23700</td>
<td>150</td>
<td>N/A</td>
</tr>
<tr>
<td>CLINC-150</td>
<td>intent classification</td>
<td>en</td>
<td>10 domains, inter alia, banking, work, travel, small talk</td>
<td>23700</td>
<td>150</td>
<td>N/A</td>
</tr>
<tr>
<td>(Larson et al., 2019)</td>
<td></td>
<td></td>
<td>21 domains, inter alia, music, news, calendar</td>
<td>25716</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td>HWU64</td>
<td>intent classification; entity extraction</td>
<td>en</td>
<td>21 domains, inter alia, music, news, calendar</td>
<td>25716</td>
<td>64</td>
<td>54</td>
</tr>
<tr>
<td>(Liu et al., 2019a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restaurants-8K</td>
<td>slot extraction</td>
<td>en</td>
<td>restaurant booking</td>
<td>11929</td>
<td>N/A</td>
<td>5</td>
</tr>
<tr>
<td>(Coope et al., 2020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snips</td>
<td>intent classification; slot extraction</td>
<td>en</td>
<td>7 domains, inter alia, music, weather, restaurant</td>
<td>14484</td>
<td>7</td>
<td>39</td>
</tr>
<tr>
<td>(Coucke et al., 2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATIS</td>
<td>intent classification; slot extraction</td>
<td>en</td>
<td>airline travels</td>
<td>5871</td>
<td>21</td>
<td>120</td>
</tr>
<tr>
<td>(Price, 1990)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: English NLU datasets. This list is non exhaustive.
## B English DST datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Language(s)</th>
<th>Domain</th>
<th>Size (dialogues)</th>
<th>H2H / H2M</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSTC1</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>bus information</td>
<td>15886</td>
<td>H2M</td>
</tr>
<tr>
<td>(Raux et al., 2005; Williams et al., 2013)</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>restaurant booking</td>
<td>3000</td>
<td>H2M</td>
</tr>
<tr>
<td>DSTC2</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>restaurant booking</td>
<td>1200</td>
<td>H2H</td>
</tr>
<tr>
<td>(Henderson et al., 2014a)</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>restaurant booking</td>
<td>1200</td>
<td>H2H</td>
</tr>
<tr>
<td>WOZ2.0</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>restaurant booking</td>
<td>1200</td>
<td>H2H</td>
</tr>
<tr>
<td>(Wen et al., 2017; Mrkšić et al., 2017a)</td>
<td>dialogue state tracking</td>
<td>en</td>
<td>restaurant booking</td>
<td>1200</td>
<td>H2H</td>
</tr>
</tbody>
</table>

Table 8: English DST datasets. This list is non exhaustive. Abbreviations: H2M – human-to-machine; H2H – human-to-human.
### English end-to-end datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>Language(s)</th>
<th>Domain</th>
<th>Size (dialogues)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiWOZ</td>
<td>end-to-end; dialogue state tracking; slot extraction;</td>
<td>en</td>
<td>7 domains, including restaurant, taxi</td>
<td>10438</td>
<td>H2H;</td>
</tr>
<tr>
<td>(Badzianowski et al., 2018)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taskmaster-1</td>
<td>end-to-end</td>
<td>en</td>
<td>6 domains, including ordering pizza, movie tickets</td>
<td>7708</td>
<td>Self-dialogues;</td>
</tr>
<tr>
<td>(Byrne et al., 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MultiDoGo</td>
<td>end-to-end; intent classification; slot extraction; dialogue acts classification;</td>
<td>en</td>
<td>6 domains, including airline, software</td>
<td>40576</td>
<td>H2H</td>
</tr>
<tr>
<td>(Peskov et al., 2019)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ConvAI2</td>
<td>end-to-end</td>
<td>en</td>
<td>chit chat, not goal oriented</td>
<td>19893</td>
<td>H2H; derived from Persona-Chat (Zhang et al., 2018a)</td>
</tr>
<tr>
<td>(Dinan et al., 2020)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: English datasets for end-to-end training.