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 so-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.[1] 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

[1]For 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, **1)** 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 con-

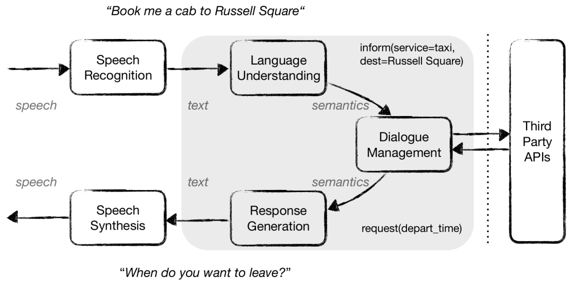


Figure 1: The typical architecture of a modular dialogue system. The gray rectangle spans the modules operating on text, which are in the focus of this survey.

crete 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 prepending 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.[2] The three core text-based components of each modular ToD system are: natural language understanding

[2]This extension, as discussed later in §4.5, comes with its own set of research challenges, but a comprehensive overview of (multilingual) ASR and TTS approaches falls way beyond the scope of this overview.

(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 *intent 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 `Find_flight` 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 nor 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; Cuayáhuitl, 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-

[3]It is important to note that in the wider NLP context, NLU has a different meaning: it refers to the set of difficult NLP tasks, solving of which is presumed to require human-level language understanding competencies and successful modeling of semantic compositionality in natural language. Representative NLU tasks include natural language inference (Williams et al., 2018), reading comprehension (Rajpurkar et al., 2016), and commonsense reasoning (Sap et al., 2020).

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 MultiATIS++ (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[4] is most consequential for zero-shot transfer.

Earlier work leveraged static cross-lingual word emebedding 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

[4]The word alignments, however, were obtained automatically from parallel sentences via FastAlign (Dyer et al., 2013).

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.[5] 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 transfered 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-

[5] For completeness, we also provide a subset of the most established English-only NLU resources in the Appendix.

Dataset	Task	Language	Domains	Size	# intents	# slots
Non-English monolingual datasets						
MEDIA (Bonneau-Maynard et al., 2005)	slot extraction	fr	hotel reservations	15000	N/A	83
SLU-IT (Castellucci et al., 2019)	intent classification; slot extraction	it	7 domains, inter alia, music, weather, restaurant	7142	7	39
Almawave-SLU (Bellomaria et al., 2019)	intent classification; slot extraction	it	7 domains, inter alia, music, weather, restaurant	14484	7	39
(Zhang et al., 2017)	intent classification	zh	chit chat; task-oriented	4000	31	N/A
ECSA dataset (Gong et al., 2019)	slot extraction; named entity extraction	zh	online commerce	27615	N/A	N/A (sequence tags)
Chinese ATIS (He et al., 2013)	intent classification; slot extraction	zh	airline travels	5871	21	120
Vietnamese ATIS (Dao et al., 2021)	intent classification; slot extraction	vi	airline travels	5871	25	120

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.

Dataset	Task(s)	Languages	Domains	Size	# intents	# slots
Multilingual TOP (Schuster et al., 2019a)	intent classification; slot extraction	en, es, th	Alarm, reminder, weather	43323 [en] 8643 [es] 5082 [th]	12	11
ATIS in Chinese and Indonesian (Susanto and Lu, 2017)	semantic parsing; slot extraction	en, zh, id	airline travels	5371	N/A	120 (166; λ-calculus)
Multilingual ATIS (Upadhyay et al., 2018)	intent classification; slot extraction	en, hi, tr	airline travels	1493 [hi] 1315 [tr]	21	120
MultiATIS++ (Xu et al., 2020)	intent classification; slot extraction	en, es, pt, de, fr, zh, ja, hi, tr	airline travels	5871 [en] 5871 [es] 5871 [pt] 5871 [de] 5871 [fr] 5871 [zh] 5871 [ja] 2493 [hi] 1353 [tr]	18; 17 [hi], 17 [tr]	84 75 [hi] 71 [tr]
MTOP (Li et al., 2021)	semantic parsing; intent classification; slot extraction	en, de, fr, es, hi, th	11 domains, inter alia, music, news, recipes	18788 [en] 16585 [de] 15459 [fr] 16182 [es] 15195 [hi] 18788 [th]	117	78

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-

Dataset	Task	Language(s)	Domain	Size (# dialogues)	H2H / H2M
DSTC 5 (Kim et al., 2016)	dialogue state tracking	en, zh	tourist information	35 [en] 12 [zh]	H2H
Multilingual WOZ 2.0 (Mrkšić et al., 2017b)	dialogue state tracking	en, de, it	restaurant booking	1200	H2H (translated)
DSTC 6 (Hori et al., 2019)	dialogue breakdown detection	en, ja	dialogues from existing datasets and those collected for the challenge	615 [en] 1696 [ja]	H2M
DSTC 9 (Gunasekara et al., 2020)	dialogue state tracking	en, zh	7 domains in [en] 5 domains in [zh]	10438 [en] 6012 [zh]	H2M

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.

Dataset	Task	Language(s)	Domain	Size (dialogues)	Comments
Non-English monolingual datasets					
CrossWOZ (Zhu et al., 2020a)	dialogue state tracking; end-to-end;	zh	5 domains, including attraction, hotel, taxi	6012	H2H;
RiSAWOZ (Zhu et al., 2020a)	dialogue state tracking; end-to-end;	zh	12 domains: including education, car, hospitality	11200	H2H;
DuConv (Wu et al., 2017)	end-to-end	zh	chit-chat	1060000 (context-response pairs)	H2H; web scraped from social network;
KdConv (Zhou et al., 2020)	end-to-end	zh	chit-chat about films. music, travel	4500	H2H; using an external knowledge base;
WMT 2020 Chat	end-to-end	de	6 domains, including ordering pizza, movie tickets	692	H2H; translated from Byrne et al. (2019); part of challenge on conversational data translation
Multilingual datasets					
XPersona (Lin et al., 2020b)	end-to-end	en, it, fr, id, zh, ko, ja	chit-chat (persona chats)	19893 [en] 17158 [it] 17375 [fr] 17846 [id] 17322 [zh] 17477 [ko] 17428 [ja]	H2H; automatically translated from Dinan et al. (2020);

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

ogy APIs have been developed (Gatt and Reiter, 2009) and subsequently adapted to a number of languages, including Spanish (Ramos-Soto et al., 2017), Galician (Cascallar-Fuentes et al., 2018), Italian (Mazzei et al., 2016), German (Bollmann, 2011), Brazilian Portuguese (De Oliveira and Sripada, 2014), as well as a bilingual English-French realiser (Vaudry and Lapalme, 2013). A hybrid approach coupling linguistic knowledge (i.e., a grammar and a lexicon) with statistical methods was recently proposed by García-Méndez et al. (2019).

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), Rönnqvist 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.[6] The sample diversity scores

[6]To measure typological diversity, we calculate the mean of entropy of 103 binary URIEL (Littell 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-

	Multilingual TOP	Multilingual ATIS	MultiATIS++	MTOP	XCOPA	TyDi QA	XNLI
# languages	3	3	9	6	11	11	15
Typology	0.196	0.257	0.343	0.267	0.41	0.41	0.39
Family	0.667	0.667	0.444	0.333	1.0	0.9	0.5
Macroareas	0	0	0	0	1.67	0.92	0.37

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).

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; Henderson and Vulić, 2021). Along the same line, DST is sometimes formulated as a standard semantic parsing task in monolingual multi-domain settings Cheng et al. (2020): dialogue states are represented as hierarchical semantic structures which include the information about domain, actions and slots which were filled or requested.[7]

This effectively means that the standard methodological 'machinery', currently used for structurally similar NLP task to deal with low-resource languages and domains with scarce data, can be directly applied to guide multilingual modeling and cross-lingual transfer for NLU in multilingual TOD (Ponti et al., 2019a; Hedderich et al., 2021). In what follows, we provide a a very brief and non-exhaustive overview of promising cutting-edge techniques that might also be applied in low-resource TOD.[8] The reader should, however, still bear in mind the core deficiencies of the current methodology in relation to multilingual TOD, as also discussed in §2.3.

Low-resource languages should benefit from annotated resources in higher-resource languages. Besides direct MT transfer (Upadhyay et al., 2018; Schuster et al., 2019a; Hu et al., 2020), annotations can be propagated source-to-target using parallel data and word alignments (Ni et al., 2017; Jain et al., 2019; Xu et al., 2020). Annotation and model transfer can also be realised via cross-lingual word embeddings (Glavaš et al., 2019; Ruder et al., 2019). More recently, unmatched

ographical diversity, the entropy of macroareas to which the languages in the dataset belong to is calculated.

[7]Formulating DST as semantic parsing opens up several paths for future research. First, structured representations naturally allow for semantic compositionality and cross-domain knowledge sharing. Similarly, they could allow for cross-lingual knowledge sharing. Secondly, structured representations facilitate the use of external knowledge. Tables are widely used in semantic parsing (Zhu et al., 2020b; Sun et al., 2019) and could be applied for efficient search and information extraction. This could be efficient for cross-lingual slot labelling (e.g., if a table includes names of a city in different languages).

[8]For a comprehensive survey of methods for low-resource NLP, we refer the reader to (Hedderich et al., 2021).

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; Hedderich 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; Bhattacharjee 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 achieved for high-resource target languages (Lauscher et al., 2020; Wu and Dredze, 2020; Zhao et al., 2020).

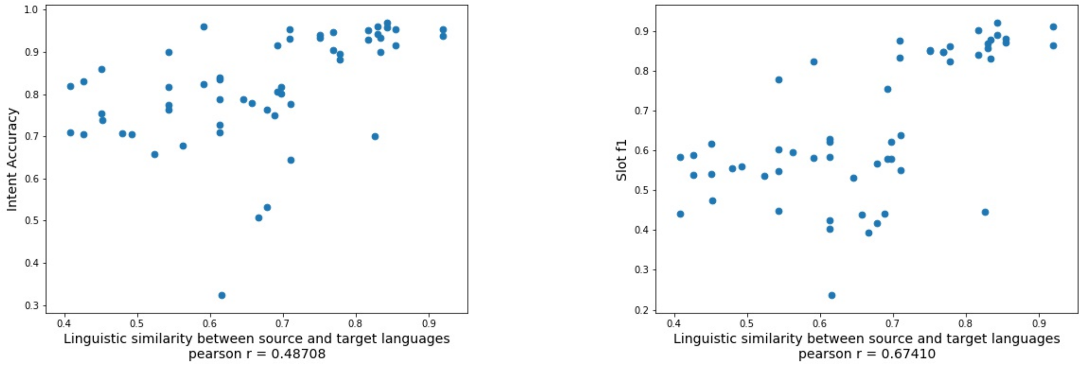
Pretrained language models can also be adaptively fine-tuned with unannotated in-domain data in both source and target language (Ponti et al., 2020) to pick up more domain-specific knowledge, which typically results in slight performance gains (Henderson et al., 2020; Gururangan et al., 2020). Along 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-



(a) Intent classification (b) Slot labelling

Figure 2: Intent classification and slot labelling performance depending on linguistic source-target similarity.

Feature Group	Intent Classification	Slot Labelling
Phonology	-0.2467	-0.2504
Geography	-0.2263	-0.3270
Phylogenetics	-0.4895	-0.6122
Syntax	-0.5131	-0.6919

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).

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 morphological information simultaneously.[9]

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;

[9] 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).

Ohmura and Eskénazi, 2018, *inter alia*), 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 its 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.[10] 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 of majority of the world's languages (Joshi et al., 2020). In the pursuit of wider-scale and democratised 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 fulfil 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`.

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

[11] Besides 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

Evgeniia 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 und domänenübergreifende Conversational AI) of the Baden-Württemberg Ministry of Economy, Labor, and Housing.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of ACL 2020*, pages 8182–8197.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What code-switching strategies are effective in dialog systems? In *Proceedings of SCIL 2020*, pages 254–264.

Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio visual scene-aware dialog. In *Proceedings of CVPR 2019*, pages 7558–7567.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of ACL 2019*, pages 1388–1398.

Duygu Altinok. 2018. An ontology-based dialogue management system for banking and finance dialogue systems. In *Proceedings of the First Financial Narrative Processing Workshop (FNP)*.

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020. Cross-lingual transfer learning for intent detection of Covid-19 utterances.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pages 789–798.

Suman Banerjee and Mitesh M Khapra. 2019. Graph convolutional network with sequential attention for goal-oriented dialogue systems. *TACL*, 7:485–500.

John A Bateman. 1997. Enabling technology for multilingual natural language generation: the kpml development environment. *Natural Language Engineering*, 3(1):15–55.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-SLU: A new dataset for slu in italian. *arXiv preprint arXiv:1907.07526*.

Kasturi Bhattacharjee, Miguel Ballesteros, Rishita Anubhai, Smaranda Muresan, Jie Ma, Faisal Ladhak, and Yaser Al-Onaizan. 2020. To BERT or not to BERT: Comparing task-specific and task-agnostic semi-supervised approaches for sequence tagging. In *Proceedings of EMNLP 2020*, pages 7927–7934.

Dan Bohus and Alexander I Rudnicky. 2009. The ravenclaw dialog management framework: Architecture and systems. *Computer Speech & Language*, 23(3):332–361.

Marcel Bollmann. 2011. Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138.

Hélene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. 2005. Semantic annotation of the French media dialog corpus. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2-how can i help you? Towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ-A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP 2018*, pages 5016–5026.

Tanja Bunk, Daksh Varshneya, Vladimir Vlasov, and Alan Nichol. 2020. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of EMNLP-IJCNLP 2019*, pages 4506–4517.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020a. Multilingual alignment of contextual word representations. In *Proceedings of ICLR 2020*.

Yan Cao, Keting Lu, Xiaoping Chen, and Shiqi Zhang. 2020b. Adaptive dialog policy learning with hindsight and user modeling. In *Proceedings of SIGDIAL 2020*, pages 329–338.

Yixin Cao, Zikun Hu, Tat-seng Chua, Zhiyuan Liu, and Heng Ji. 2019. Low-resource name tagging learned with weakly labeled data. In *Proceedings of EMNLP-IJCNLP 2019*, pages 261–270.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online.

Andrea Cascallar-Fuentes, Alejandro Ramos-Soto, and Alberto Bugarín Diz. 2018. Adapting SimpleNLG to Galician language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 67–72, Tilburg University, The Netherlands. Association for Computational Linguistics.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint BERT-based model. *arXiv preprint arXiv:1907.02884*.

David L Chen, Joohyun Kim, and Raymond J Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37:397–435.

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

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*.

Wenhu Chen, Jianshu Chen, Yu Su, Xin Wang, Dong Yu, Xifeng Yan, and William Yang Wang. 2018. Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of EMNLP 2018*, pages 414–424.

Jianpeng Cheng, Devang Agrawal, Héctor Martínez Alonso, Shruti Bhargava, Joris Driesen, Federico Flego, Dain Kaplan, Dimitri Kartsaklis, Lin Li, Dhivya Piraviperumal, Jason D. Williams, Hong Yu, Diarmuid Ó Séaghdha, and Anders Johannsen. 2020. Conversational semantic parsing for dialog state tracking. In *Proceedings of EMNLP 2020*, pages 8107–8117.

Adam Cheyer and Didier Guzzoni. 2006. Method and apparatus for building an intelligent automated assistant. Technical report.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7570–7577.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*, 8:454–470.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020a. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020c. Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL 2020*, pages 6022–6034.

Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. 2020. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of ACL 2020*, pages 107–121.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Heriberto Cuayáhuitl. 2017. Simpleds: A simple deep reinforcement learning dialogue system. In *Dialogues with social robots*, pages 109–118. Springer.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of COLING 2020*, pages 3861–3867.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent detection and slot filling for Vietnamese. *arXiv preprint arXiv:2104.02021*.

Rodrigo De Oliveira and Somayajulu Sripada. 2014. Adapting simplenlg for Brazilian Portuguese realisation. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 93–94.

Kerstin Denecke, Mauro Tschanz, Tim Lucas Dorner, and Richard May. 2019. Intelligent conversational agents in healthcare: hype or hope. *Stud Health Technol Inform*, 259:77–84.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, volume 1, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). *The NeurIPS'18 Competition*, pages 187–208.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *CoRR*, abs/2010.02194.

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.

Ondřej Dušek and Filip Jurčíček. 2019. Neural generation for Czech: Data and baselines. *The 12th International Conference on Natural Language Generation*, pages 563–574.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL-HLT 2013*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of EMNLP 2018*, pages 489–500.

Reiter Ehud and Dale Robert. 2000. Building natural language generation systems.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *Proceedings of INTERSPEECH 2016*, pages 1151–1155.

Michael Elhadad and Jacques Robin. 1996. An overview of surge: a reusable comprehensive syntactic realization component. *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Kumar Goyal, Peter Ku, and Dilek Hakkani-Tür. 2020. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of LREC 2020*, pages 422–428.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of SIGDIAL 2017*, pages 37–49.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Katja Filippova and Michael Strube. 2007. Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 320–327, Prague, Czech Republic. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML 2017*, pages 1126–1135.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. In *Proceedings of ACL 2019*, pages 5539–5544.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Cristina Garbacea and Qiaozhu Mei. 2020. Neural language generation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2007.15780*.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur P. Parikh. 2020. Harnessing multilinguality in unsupervised machine translation for rare languages. *CoRR*, abs/2009.11201.

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.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of EMNLP 2020*, pages 6174–6181.

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.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of EMNLP 2017*, pages 2839–2845.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP 2018*, pages 316–327.

Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. 2013. Multilingual training of deep neural networks. In *Proceedings of ICASSP 2013*, pages 7319–7323.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL 2019*, pages 710–721.

Goran Glavaš and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of ACL 2018)*, pages 34–45.

Goran Glavaš and Ivan Vulić. 2020. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL 2020*, pages 7548–7555.

Yu Gong, Xusheng Luo, Yu Zhu, Wenwu Ou, Zhao Li, Muhua Zhu, Kenny Q Zhu, Lu Duan, and Xi Chen. 2019. Deep cascade multi-task learning for slot filling in online shopping assistant. In *Proceedings of AAAI 2019*, volume 33, pages 6465–6472.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of NAACL-HLT 2018*, volume 2, pages 753–757.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the 9th Dialog System Technology Challenge: DSTC9. *arXiv preprint arXiv:2011.06486*.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 554–559. IEEE.

Narendra Gupta, Gokhan Tur, Dilek Hakkani-Tur, Srinivas Bangalore, Giuseppe Riccardi, and Mazin Gilbert. 2005. The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pages 8342–8360.

E Haihong, Wenjing Zhang, and Meina Song. 2019. Kb-transformer: Incorporating knowledge into end-to-end task-oriented dialog systems. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 44–48. IEEE.

Dilek Hakkani-Tür, Gökhan Tür, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of Interspeech 2016*, pages 715–719.

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.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.0. *Max Planck Institute for the Science of Human History*.

Hilda Hardy, Tomek Strzalkowski, Min Wu, Cristian Ursu, Nick Webb, Alan Biermann, R. Bryce Inouye, and Ashley McKenzie. 2004. Data-driven strategies for an automated dialogue system. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 71–78, Barcelona, Spain.

Keqing He, Weiran Xu, and Yuanmeng Yan. 2020. Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, 8:29407–29416.

Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *Proceedings of ICASSP 2013*, pages 8342–8346. IEEE.

Michael A. Hedderich, David Ifeoluwa Adelani, Dawei Zhu, Jesujoba O. Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual Transformer models: A study on african languages. In *Proceedings of EMNLP 2020*, pages 2580–2591.

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*.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of EACL 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, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of SIGDIAL 2014*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014b. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of SIGDIAL 2014*, pages 292–299.

Matthew Henderson and Ivan Vulić. 2021. ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of NAACL-HLT 2021*.

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

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Panikos Heracleous, Yasser Mohammad, and Akio Yoneyama. 2020. Integrating language and emotion features for multilingual speech emotion recognition. In *International Conference on Human-Computer Interaction*, pages 187–196.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *Proceddings of the Deep Learning Workshop at NeurIPS*.

Chiori Hori, Julien Perez, Ryuichiro Higashinaka, Takaaki Hori, Y-Lan Boureau, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, Koichiro Yoshino, and Seokhwan Kim. 2019. Overview of the sixth dialog system technology challenge: Dstc6. *Computer Speech & Language*, 55:1–25.

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.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1083–1092.

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.

Armand Joulin, Piotr Bojanowski, Tomáš Mikolov, Hervé Jégou, and Édouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP 2018*, pages 2979–2984.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2020. Cross-lingual text classification with minimal resources by transferring a sparse teacher. In *Proceedings of EMNLP 2020*, pages 3604–3622.

Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016*, pages 2012–2021.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of ACL 2020*, pages 3575–3585.

Douwe Kiela, Ivan Vulić, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of EMNLP 2015*, pages 148–158.

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.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, Matthew Henderson, and Koichiro Yoshino. 2016. The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020a. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of ACL 2020*, pages 567–582.

Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, and Ruhi Sarikaya. 2018. Efficient large-scale neural domain classification with personalized attention. In *Proceedings of ACL 2018*, pages 2214–2224.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020b. When and why is unsupervised neural machine translation useless? In *Proceedings of EAMT 2020*, pages 35–44.

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2020. Enhancing deep neural networks with morphological information. *arXiv preprint arXiv:2011.12432*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of NAACL-HLT 2018*, pages 452–457.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *PNAS*, 117(14):7684–7689.

Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. 2021. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. *arXiv preprint arXiv:2103.07792*.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching Yun Chang. 2020. Cross-lingual alignment methods for multilingual bert: A comparative study. In *Proceedings of EMNLP 2020*, pages 933–942.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, and Pushpak Bhattacharyya. 2014. Shata-anuvadak: Tackling multiway translation of Indian languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1781–1787, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder lstm for semantic slot filling. In *Proceedings of EMNLP 2016*, pages 2077–2083.

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

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1311–1316.

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.

Viet Bac Le and Laurent Besacier. 2009. Automatic speech recognition for under-resourced languages: Application to Vietnamese Language. *IEEE Transactions on Speech Audio Processing*, 17(8):1471–1482.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of EACL 2021*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016c. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275.

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.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL 2019*, pages 3125–3135.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020a. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of EMNLP 2020*, pages 2649–2663.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020b. XPersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Pierre Lison, Jeremy Barnes, Aliaksandr Hubin, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of ACL 2020*, pages 1518–1533.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL 2017*, volume 2, pages 8–14.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of NAACL-HLT 2018*, pages 2060–2069.

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.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021. On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of of AAAI 2021*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. 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. 2020b. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of AAAI 2020*, pages 8433–8440.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of COLING 2020*, pages 480–496.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017a. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017b. Training end-to-end dialogue systems with the Ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1):31–65.

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of ACL 2017*, pages 430–439.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of ACL 2018*, pages 1468–1478.

Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. Named entity recognition with partially annotated training data. In *Proceedings of CoNLL 2019*, pages 645–655.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th International Natural Language Generation conference*, pages 184–192.

Susan W McRoy, Songsak Channarukul, and Syed S Ali. 2003. An augmented template-based approach to text realization. *Natural Language Engineering*, 9(4):381.

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

Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and Maarten de Rijke. 2020. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of SIGIR 2020*, pages 1151–1160.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2014. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR, abs/1309.4168*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions.

Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. 2018. DIDEC: The Dutch image description and eye-tracking corpus. In *Proceedings of COLING 2018*, pages 3658–3669.

Danilo Mirkovic and Lawrence Cavedon. 2011. Dialogue management using scripts. US Patent 8,041,570.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of EMNLP 2020: System Demonstrations*, pages 119–126.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017a. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL 2017*, volume 1, pages 1777–1788.

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *Proceedings of ACL 2018*, pages 108–113.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017b. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *TACL*, 5:309–324.

Christian Muise, Tathagata Chakraborti, Shubham Agarwal, Ondrej Bajgar, Arunima Chaudhary, Luis A Lastras-Montano, Josef Ondrej, Miroslav Vodolan, and Charlie Wiecha. 2019. Planning for goal-oriented dialogue systems. *arXiv preprint arXiv:1910.08137*.

Mahdi Namazifar, Alexandros Papangelis, Gökhan Tür, and Dilek Hakkani-Tür. 2020. Language model is all you need: Natural language understanding as question answering. *CoRR*, abs/2011.03023.

Tomás Nekvinda and Ondrej Dusek. 2020. One model, many languages: Meta-learning for multilingual text-to-speech. In *Proceedings of INTERSPEECH 2020*, pages 2972–2976.

Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, and Mitesh M. Khapra. 2018. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1539–1550, New Orleans, Louisiana. Association for Computational Linguistics.

Jian Ni, Georgiana Dinu, and Radu Florian. 2017. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of ACL 2017*, pages 1470–1480.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of EMNLP 2020*, pages 4547–4562.

Junki Ohmura and Maxine Eskénazi. 2018. Context-aware dialog re-ranking for task-oriented dialog systems. In *Proceedings of SLT 2018*, pages 846–853.

Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in Natural Language Processing. In *Proceedings of COLING 2016*, pages 1297–1308.

Yasumasa Onoe and Greg Durrett. 2019. Learning to denoise distantly-labeled data for entity typing. In *Proceedings of NAACL-HLT 2019*, pages 2407–2417.

Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. 2018. Exemplar encoder-decoder for neural conversation generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1329–1338, Melbourne, Australia. Association for Computational Linguistics.

Peyman Passban, Qun Liu, and Andy Way. 2018. Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of NAACL-HLT 2018*, pages 58–68.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of EMNLP 2020*, pages 172–182.

Julien Perez and Fei Liu. 2017. Dialog state tracking, a machine reading approach using memory network. In *Proceedings of EACL 2017*, volume 1, pages 305–314.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of EMNLP-IJCNLP 2019*, pages 4518–4528.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020a. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of EMNLP 2020*, pages 7654–7673.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. UNKs Everywhere: Adapting multilingual language models to new scripts. *CoRR*, abs/2012.15562.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL 2019*, pages 4996–5001.

Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. 2010. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 13(1):49–60.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019a. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018a. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of ACL 2018*, pages 1531–1542.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018b. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP 2018*, pages 282–293.

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

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020a. Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. In *Proceedings of INTERSPEECH 2020*, pages 4751–4755.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020b. MLS: A large-scale multilingual dataset for speech research. In *Proceedings of INTERSPEECH 2020*, pages 2757–2761.

Patti Price. 1990. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 2020. Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the ACL 2020*, pages 6344–6354.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of EMNLP 2020*, pages 930–940, Online. Association for Computational Linguistics.

Jonathan Raiman and John Miller. 2017. Globally normalized reader. In *Proceedings of EMNLP 2017*, pages 1059–1069.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pages 2383–2392.

Alejandro Ramos-Soto, Julio Janeiro-Gallardo, and Alberto Bugarín Diz. 2017. Adapting simplenlg to spanish. In *Proceedings of the 10th international conference on natural language generation*, pages 144–148.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of COLING 2020*, pages 6838–6855.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let's go public! taking a spoken dialog system to the real world. In *Proceedings of INTERSPEECH 2005*.

Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Fuji Ren and Siyuan Xue. 2020. Intention detection based on siamese neural network with triplet loss. *IEEE Access*, 8:82242–82254.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786, Brussels, Belgium. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36.

Guy Rotman, Ivan Vulić, and Roi Reichart. 2018. Bridging languages through images with Deep Partial Canonical Correlation Analysis. In *Proceedings of ACL 2018*, pages 910–921.

Andreas Rücklé, Jonas Pfeiffer, and Iryna Gurevych. 2020. MultiCQA: Zero-shot transfer of self-supervised text matching models on a massive scale. In *Proceedings of EMNLP 2020*, pages 2471–2486.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. 2020. How good is your tokenizer? On the monolingual performance of multilingual language models. *CoRR*, abs/2012.15613.

Gözde Gül Şahin and Mark Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of EMNLP 2018*, pages 5004–5009.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of ACL 2020: Tutorial Abstracts*, pages 27–33.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019a. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT 2019*, volume 1, pages 3795–3805.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019b. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT 2019*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016a. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI 2016*, volume 30.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016b. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of ACL 2020*, pages 6322–6333.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of NAACL-HLT 2019*, pages 2578–2587.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Reddy Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020a. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of ACL 2020*, pages 2827–2835.

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

Gunnar A. Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Proceedings of CVPR 2020*, pages 10847–10856.

Samuel L. Smith, David H.P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR 2017*.

Yiping Song, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, Dongyan Zhao, and Rui Yan. In *Proceedings of IJCAI 2018*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings* *of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86, Barcelona, Spain.

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

Yibo Sun, Duyu Tang, Jingjing Xu, Nan Duan, Xiaocheng Feng, Bing Qin, Ting Liu, and Ming Zhou. 2019. Knowledge-aware conversational semantic parsing over web tables. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 827–839. Springer.

Raymond Hendy Susanto and Wei Lu. 2017. Neural architectures for multilingual semantic parsing. In *Proceedings of ACL 2017*, volume 2, pages 38–44.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of EMNLP 2020*, pages 2302–2315.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1105–1116.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 6000–6010.

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.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! visual object discovery through multi-modal dialogue. In *Proceedings of CVPR 2017*, pages 4466–4475.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT 2018*, pages 516–527.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of EMNLP-IJCNLP 2019*, pages 4398–4409.

Ivan Vulić, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL 2017*, pages 56–68.

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.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP-IJCNLP 2019*, pages 6382–6388.

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 EMNLP-IJCNLP 2019: 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*, volume 1, pages 438–449.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT 2018*, volume 1, pages 1112–1122.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

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.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP 2019*, pages 833–844.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of NeurIPS 2020*.

Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1341–1350.

Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of SIGIR 2018*, pages 245–254.

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

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

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, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of ACL 2018*, pages 2204–2213.

Wei-Nan Zhang, Zhigang Chen, Wanxiang Che, Guoping Hu, and Ting Liu. 2017. The first evaluation of chinese human-computer dialogue technology. *arXiv preprint arXiv:1709.10217*.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018b. Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of NeurIPS 2018*, pages 1815–1825.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020*, pages 270–278.

Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R. J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019a. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In *Proceedings of INTERSPEECH 2019*, pages 2080–2084.

Zhichang Zhang, Zhenwen Zhang, Haoyuan Chen, and Zhiman Zhang. 2019b. A joint learning framework with bert for spoken language understanding. *IEEE Access*, 7:168849–168858.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018c. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Variance, benchmarks and baselines. *CoRR*, abs/2012.15682.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of ACL 2018*, volume 1, pages 1458–1467.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of ACL 2020*, pages 7098–7108.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2019. Multi-task learning for natural language generation in task-oriented dialogue. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1261–1266.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020a. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *TACL*, 8:281–295.

Shuguang Zhu, Xiang Cheng, and Sen Su. 2020b. Conversational semantic parsing over tables by decoupling and grouping actions. *Knowledge-Based Systems*, 204:106237.

Yftah Ziser and Roi Reichart. 2018. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of EMNLP 2018*, pages 238–249.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP 2016*, pages 1568–1575.

A English NLU datasets

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

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

B English DST datasets

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

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

C English end-to-end datasets

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

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