Crossing the Conversational Chasm:
A Primer on Natural Language Processing for
Multilingual Task-Oriented Dialogue Systems

Evgeniia Razumovskaia   ER563@CAM.AC.UK
Language Technology Lab, University of Cambridge, UK

Goran Glavaš   GORAN.GLAVAS@UNI-WUERZBURG.DE
Center for AI and Data Science (CAIDAS),
University of Würzburg, Germany

Olga Majewska   OM304@CAM.AC.UK
Language Technology Lab, University of Cambridge, UK

Edoardo M. Ponti   EDOARDO-MARIA.PONTI@MILA.QUEBEC
Mila - Quebec AI Institute and McGill University, Canada
Language Technology Lab, University of Cambridge, UK

Anna Korhonen   ALK23@CAM.AC.UK
Language Technology Lab, University of Cambridge, UK

Ivan Vulić   IV250@CAM.AC.UK
Language Technology Lab, University of Cambridge, UK
PolyAI Limited, UK

Abstract

In task-oriented dialogue (ToD), a user holds a conversation with an artificial agent with the aim of completing a concrete task. Although this technology represents one of the central objectives of AI and has been the focus of ever more intense research and development efforts, it is currently limited to a few narrow domains (e.g., food ordering, ticket booking) and a handful of languages (e.g., English, Chinese). This work provides an extensive overview of existing methods and resources in multilingual ToD as an entry point to this exciting and emerging field. We find that the most critical factor preventing the creation of truly multilingual ToD systems is the lack of datasets in most languages for both training and evaluation. In fact, acquiring annotations or human feedback for each component of modular systems or for data-hungry end-to-end systems is expensive and tedious. Hence, state-of-the-art approaches to multilingual ToD mostly rely on (zero- or few-shot) cross-lingual transfer from resource-rich languages (almost exclusively English), either by means of (i) machine translation or (ii) multilingual representations. These approaches are currently viable only for typologically similar languages and languages with parallel / monolingual corpora available. On the other hand, their effectiveness beyond these boundaries is doubtful or hard to assess due to the lack of linguistically diverse benchmarks (especially for natural language generation and end-to-end evaluation). To overcome this limitation, we draw parallels between components of the ToD pipeline and other NLP tasks, which can inspire solutions for learning in low-resource scenarios. Finally, we list additional challenges that multilinguality poses for related areas (such as speech, fluency in generated text, and human-centred evaluation), and indicate future directions that hold promise to further expand language coverage and dialogue capabilities of current ToD systems.

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[1] (Gupta et al., 2005; Bohus & Rudnicky, 2009; Young et al., 2013; Muise et al., 2019) have attracted interest from the scientific community as well as businesses as a so-far more feasible application. In fact, this technology has the potential to help or altogether replace human operators in focused problems and areas such as restaurant booking (Kim & Banchs, 2014; Henderson et al., 2019a), banking (Hardy et al., 2004; Altinok, 2018), travel (Li et al., 2018; Zang et al., 2020), or (non-emergency) healthcare (Laranjo et al., 2018; Denecke et al., 2019).
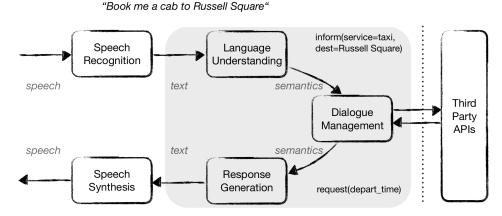
The accelerated pace at which new milestones are reached across natural language applications thanks to the growing viability of deep learning techniques has recently catalysed dialogue-oriented research (Ren et al., 2018; Wen et al., 2019; Henderson et al., 2020; Wu et al., 2020a, *inter alia*). Coupled 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 between speakers of dominant and under-represented languages in their access to state-of-the-art language technology.[2] Extending the reach of conversational technology is crucial for the democratisation of human–machine communication. This endeavour requires developing approaches that generalise across diverse language varieties and linguistic phenomena, are robust to cross-cultural differences in dialogue behaviours, and efficiently capitalise on available training data, whose scarcity continues to hold back 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-how related to multilingual ToD, and shed new light on the following topics:

(Q1) Which ToD **datasets** are available in one or more languages other than English? What are their strengths and weaknesses?

(Q2) What are the best **methods** and practices to incorporate language-specific information and conduct target language adaptation for multilingual and cross-lingual ToD?

(Q3) How can multilingual ToD take inspiration from other **related fields** of NLP research to better tackle low-resource scenarios?

(Q4) What are the **future challenges** faced when developing ToD systems in several different languages, especially with respect to voice-based and human-centred dialogue?

---

1. Note that throughout the paper the word 'task' can refer to either 1) the goal of a dialogue, or 2) a discriminative or generative problem, since both usages are standard in the fields of dialogue and NLP, respectively. The intended use of the term should always be easy to disambiguate from the context.

2. For example, Amazon Alexa, one of the most popular virtual personal assistants, currently supports only eight resource-rich languages: English, French, German, Hindi, Italian, Japanese, Brazilian Portuguese, and Spanish.

"Book me a cab to Russell Square"

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.

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 other NLP tasks and applications (e.g., named entity recognition, dependency parsing, QA) with respect to linguistic diversity, training and evaluation data availability, cross-lingual transfer methodology, and joint multilingual modelling (Ponti et al., 2019; Hedderich et al., 2021). We hope that this survey will inspire more work in these areas. First, by drawing direct links (including similarities and differences) between ToD components and other cross-lingual NLP tasks, we advocate for the use and adaptation of existing techniques in support of multilingual ToD. Secondly, by surveying existing (multilingual) ToD resources and, consequently, exposing the current lack of training and evaluation resources for a large number of languages and domains, we emphasise the need for their creation.

2. Task-Oriented Dialogue Systems

The purpose of ToD systems, geared towards practical applications, is to assist users in completing a concrete task through conversational interaction (Young, 2010; Chen et al., 2017; Su et al., 2018). Typically, the tasks are well-defined and the communication with the system has a binary outcome: either the task is successfully completed or not. Common examples of tasks include booking (restaurants, transportation, hotels), customer support (e.g., in banking or telecommunications), or retrieving and providing information (e.g., in healthcare or tourism).

For completeness, we first give a concise overview of the two existing approaches to task-oriented dialogue: (i) the *modular* approach, in which ToD is broken down into a pipeline of sub-tasks and (ii) *end-to-end* ToD, where a single neural model is trained to generate responses based on the context of the preceding interactions.

2.1 Modular Task-Oriented Dialogue

The modular approach to ToD addresses the complexity of goal-oriented dialogue by breaking it down into a sequence of sub-tasks. The solution, as depicted in Figure 1, is a pipeline of independently trained and executed models (components): the output of each serves as the input to the next.[3] In this work, we focus our attention on dialogue systems that read text input and generate text output—which can be further extended to voice-based interactions by prepending an automatic speech recognition (ASR, speech-to-text) component to the start of the pipeline and appending a speech synthesis (TTS, text-to-speech) component to its end.[4] The three core text-based components of each modular ToD system are: natural language understanding (NLU), dialogue (policy) management (PM), and response generation (RG), outlined in what follows.

Natural Language Understanding (NLU). In the context of ToD systems,[5] natural language understanding consists of identifying the main goals and information expressed in the user's utterances. It usually encompasses two sub-tasks, namely *intent classification* (also known as *dialogue act classification*; Ravuri & 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 multi-label classification task where one or more intent labels must be assigned to the whole utterance of a user; the latter, instead, requires to fill in predefined slots of information by extracting values from the content of the utterance. For example, the utterance "*Show me the flights from Boston to New York today*" corresponds to the intent `find_flight` and specifies values for three slots of information: `Boston` as departure location, `New York` as arrival location, and `today` as time. Given that the available slots depend on the utterance intent, the two tasks are often addressed jointly via multi-task learning (Xu & Sarikaya, 2013; Guo et al., 2014; Goo et al., 2018; Chen et al., 2019; Wu et al., 2020a, *inter alia*).

Traditionally, ToD systems included a component for *dialogue state tracking* (DST), which falls in between NLU and dialogue management. The purpose of DST models (Henderson et al., 2014b; Mrkšić et al., 2017; Perez & Liu, 2017; Zhong et al., 2018, *inter alia*) is to maintain a *dialogue belief state*, the summary of the dialogue history. This includes all the goals and slot values expressed by the user throughout the conversation. The input to DST at each user turn consists of the previous belief state and the outputs of the intent classification and slot filling modules; the output of DST is the new or updated belief state. With the advent of attention-based Transformer models (Vaswani et al., 2017; Devlin et al., 2019) and their ability to encode long sequences and capture long-distance dependencies, however, it has become possible to build latent representations of the dialogue history (from scratch) at every turn. Since maintaining an explicit belief state was no longer a strict requirement, this diminished the importance of DST in several Transformer-powered ToD systems (Wolf

---

3. These outputs are often in a discrete and/or structured form that makes them non-differentiable with respect to the model parameters.

4. 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.

5. Note that, in the wider NLP context, NLU refers instead to the set of NLP tasks whose solution is presumed to require human-level competence in understanding natural language and composing meaning. Representative NLU tasks include natural language inference (Williams et al., 2018), reading comprehension (Rajpurkar et al., 2016), and commonsense reasoning (Sap et al., 2020), among others.

et al., 2019; Budzianowski & Vulić, 2019). Nonetheless, more recent work demonstrates that explicit dialogue state tracking may still be beneficial even for these models (Lee, 2021; Lee et al., 2021). Hence, we provide a brief overview of DST in multilingual ToD later in §3.

Dialogue (Policy) Management (PM) refers to the ToD component responsible for choosing the system actions based on the current dialogue state. Approaches to PM can be broadly categorised into those based on rules, supervised learning, or reinforcement learning (RL; Su et al., 2018). RL-based PM has been the predominant paradigm in recent years, because it is more flexible than rules and does not require utterance-level annotations like supervised learning. Nevertheless, a large number of conversations along with the final outcome label (i.e., successful or not successful) are still needed as a reward/penalty for RL. This has directed the research efforts towards simulating user interactions with the policy manager (El Asri et al., 2016; Cuayáhuitl, 2017; Cao et al., 2020b). PM models are independent from the dialogue language: they receive an abstract representation 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, as it inherits all the challenges and corresponding solutions directly from the research on monolingual PM.

Response Generation (RG) is the module in charge of producing the system utterances in response to the user utterances, given the system action predicted by the policy manager. Much like PM, early work on RG relied on templates and rules hand-crafted by domain experts (Langkilde & Knight, 1998; Stent et al., 2004; Cheyer & Guzzoni, 2006; Mirkovic & 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., 2018; Zhu et al., 2019; Peng et al., 2020) or (2) *retrieving* the most suitable response from a predefined set of candidate replies, also known as *response selection* (e.g., Lowe et al., 2017a; Yang et al., 2018; Zhang et al., 2018; Henderson et al., 2019b).

Retrieval methods, on the one hand, offer the advantages of fluency, grammatical correctness, and high quality of the replies; modern neural natural language generation (NLG), in contrast, often produces 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 system versatility, making response quality highly dependent on the size of the response inventory (based on a corpus of human–human interactions). Hybrid methods combine the best of both worlds (Song et al., 2018; 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 produces the final system response.

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

The components of a modular ToD system are trained and executed in isolation. As a consequence, the later pipeline components inherit the errors of earlier components, but cannot provide feedback to correct them. To remedy this well-known issue of pipeline learning systems, ToD can instead rely on end-to-end 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), the parameters of which are all jointly tuned in a single training procedure; however, this is by no means indispensable.

On the one hand, end-to-end training addresses the main issue of modular ToD, namely component isolation and error cascading. On the other hand, e2e models aim to capture complex interactions among intents, policies, and responses in a latent representation space: this typically requires a large number of model parameters, whose reliable estimation in turn requires a large number of examples. Because of this exigency, e2e models have been much more successful in open-domain conversations (i.e., chatbots; 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?

Sub-tasks of modular ToD systems can be interpreted as specific instances of more general classes of NLP problems. For instance, intent classification is a short-text classification task, whereas slot-filling can be cast as a sequence labelling, span extraction, or even a question answering task (cf.§4.2). Top performance in such tasks is obtained with supervised learning. Adopting this strategy for multilingual ToD, however, postulates the availability of labelled data for most natural languages, for each task of interest. The impossibility of fulfilling this precondition, due to the cost- and time-intensive nature of the annotation process, is the principal bottleneck of multilingual NLP (Joshi et al., 2020): most existing datasets for general 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 this prospect even more remote: for optimal ToD in a given language, one would need to collect language-specific annotations for each sub-task (intent detection, slot filling, response selection and/or response generation).

The absence of language-specific annotations for most languages steered research efforts towards cross-lingual transfer, where knowledge from resource-rich languages is harnessed to facilitate zero-shot or few-shot predictions in resource-lean languages. Cross-lingual transfer, however, requires bridging between languages whose properties may vary considerably. Therefore, it is more likely to succeed when the source and target languages are close in terms of typology, family, and/or area (Lin et al., 2019; Lauscher et al., 2020). Recently, cross-lingual word embeddings (Ruder et al., 2019; Glavaš et al., 2019) and massively multilingual pre-trained encoders (MEs) based on Transformers (Devlin et al., 2019; Conneau et al., 2020) have been the most popular vehicles for cross-lingual transfer of NLP models. While the transfer capabilities of MEs were initially showcased as remarkable (Pires et al., 2019; Wu & Dredze, 2019), Lauscher et al. (2020) and Wu and Dredze (2020) have recently demonstrated that their effectiveness is drastically diminished in distant target languages whose monolingual corpora are small-sized. Likewise, MEs might represent a viable solution for task-oriented dialogue only in well-documented languages similar to English.

An alternative paradigm for cross-lingual transfer in NLP is contingent on (neural) machine translation (MT; Banea et al., 2008; Durrett et al., 2012; Conneau et al., 2018) and comes in two flavours: 'translate test' maps the evaluation data onto the source language, whereas 'translate train' maps the training data onto the target language. In both cases, a *monolingual* model (in the source or target language, respectively) can be subsequently deployed on the

task. As a consequence, this approach is not prone to the 'curse of multilinguality' that plagues massively multilingual encoders (Arivazhagan et al., 2019; Conneau et al., 2020). On the other hand, the coverage of translation-based transfer is constrained in terms of languages, as the development of MT models hinges upon the availability of parallel corpora, and in terms of tasks, as only sentence-level (but not token-level) properties can be preserved in translation. In modular ToD, this approach is therefore compatible with a subset of NLU tasks (intent classification and DST, but not slot filling) and RG. In end-to-end ToD, it could instead be operationalised by integrating MT modules into the pipeline before NLU and after RG. Foreshadowing §3.4, however, this idea has never been given concrete form thus far.

Although cross-lingual transfer in ToD is conceptually sound and practically feasible, there is only anecdotal evidence for its wide-range effectiveness (Schuster et al., 2019a; Liu et al., 2019), primarily due to the scarcity of multilingual evaluation benchmarks. For the same reason, it also remains unclear how different transfer approaches (e.g., machine translation and multilingual encoders) compare against each other, and which one should be preferred in relation to particular sets of ToD-related tasks, languages, and domains.

3. Existing Methods and Resources for Multilingual ToD

We now provide an overview of existing methods and resources for multilingual and cross-lingual ToD. Fig. 2 displays the taxonomy of these approaches. We group the approaches to multilingual and cross-lingual ToD in two main categories: (a) those that modify (i.e., adjust) the data, either training or evaluation data, in order to achieve better multilingual or cross-lingual transfer performance; and (b) those that introduce models (or adjustments to existing models) better tailored for multilingual ToD or cross-lingual ToD transfer in general, or a particular target setup (particular language(s) and/or task(s)). We next cover the multilingual and cross-lingual work focusing on each component of modular ToD (§3.1–3.3), and then follow with the overview of the multilingual and cross-lingual efforts in end-to-end 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., 2019; Xu et al., 2020; Bunk et al., 2020, inter alia) or (ii) independently, addressing each of the tasks with separately trained models (Ren & 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 that may be filled with values given an utterance also depend on the intent of that utterance.

Cross-Lingual Transfer Methods. Given the absence of sufficiently large training data in many languages, the default approach to multilingual NLU is (zero-shot or few-shot) transfer of models trained on English datasets by means of massively multilingual Transformer-based encoders (Zhang et al., 2019b; Xu et al., 2020; Siddhant et al., 2020b; Krishnan et al., 2021). While most of the work relies on MEs pretrained via masked language modelling,
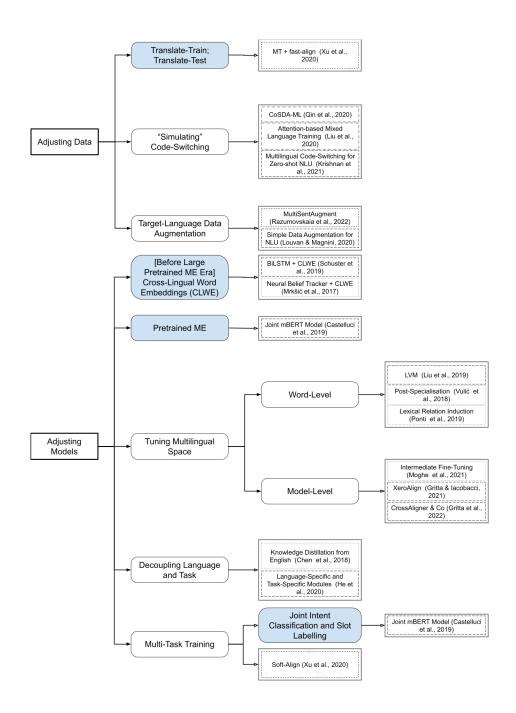
Figure 2: Taxonomy of approaches to multilingual dialogue systems, with prominent examples for established methods. Abbreviations: ME – multilingual encoders. The shaded nodes indicate well established methods, commonly used as baselines; for these methods we provide example papers where they were used. Nodes with dashed edges denote methods that are commonly applied to NLU tasks; nodes with dotted edges indicate methods specifically designed for DST; The nodes with mixed dashes-and-dots edges correspond to methods that can be effectively used for both tasks.

Dataset	Task	Language	Domains	Size	# Intents	# Slots
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 (e.g., music, weather, restaurant)	7142	7	39
Almawave-SLU (Bellomaria et al., 2019)	intent classification; slot extraction	it	7 domains (e.g., 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 in non-English languages. Shaded rows correspond to datasets created from scratch, white rows to datasets translated from English. A non-exhaustive list of English NLU datasets is provided in the Appendix.

such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), Siddhant et al. (2020b) show that MEs pretrained via machine translation leads to more effective zero-shot transfer for intent classification. Moreover, Liu et al. (2021) illustrated how mBERT-based sequence labelling overfits to the word order of the source language. Regularising for word order information (e.g., by removing positional embeddings or by shuffling tokens) benefits cross-lingual transfer in various sequence labelling tasks, including ToD slot filling. He et al. (2020) demonstrated the efficiency of dividing slot tagging models into language-specific modules and modules shared across languages, and complementing the input data with language discriminators.

Furthermore, representation sub-spaces of individual languages in MEs can be better aligned with the explicit supervision of word or sentence alignments (Cao et al., 2020b; Conneau et al., 2020). Kulshreshtha et al. (2020) found that these yielded superior zero-shot performance 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) compared to vanilla MEs. Recent work (Gritta & Iacobacci, 2021; Gritta et al., 2022) shows that contrastive learning can significantly improve the cross-lingual alignment of text representations for ToD slot-filling and intent classification, even if starting from pretrained massively multilingual encoders, known to suffer from the "curse of multilinguality" (Conneau et al., 2020; Pfeiffer et al., 2022).

Earlier work leveraged static cross-lingual word embedding spaces (CLWEs; Mikolov et al., 2013; Smith et al., 2017; Artetxe et al., 2018; Joulin et al., 2018; Patra et al., 2019; Glavaš & 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). In fact, even a small number of word alignments in the actual ToD domain of interest is sufficient to refine the CLWE space (Liu et al., 2019) or simulate code-switching data (Liu et al., 2020).

Available Datasets. While multilingual NLU is better resourced than other tasks in modular ToD, the landscape of existing datasets is still sparsely populated. Table 1 lists

Dataset	Task(s)	Languages	Domain(s)	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 & 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, es, pt, de, fr, ja zh] 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 (e.g., music, news, recipes)	18788 [en] 16585 [de] 15459 [fr] 16182 [es] 15195 [hi] 18788 [th]	117	78
xSID♠ (van der Goot et al., 2021)	intent classification; slot extraction	ar, da, de, de-st, en, id, it, kk, nl, sr, tr, zh	6 domains, combining Multilingual TOP and SNIPS	800 [ar, da, de, de-st, en, id, it, kk, nl, sr, tr, zh] 400 [ja]	16	33
MASSIVE (Bastianelli et al., 2020)	intent classification; slot extraction	51 languages, incl. am, fi, mn, my, nb, alb, tl, ur	18 domains (e.g., alarm, calendar, music, email)	19521[per lang]	60	55

Table 2: Multilingual NLU datasets. ♠xSID is an evaluation-only dataset, whereas all the other datasets listed here also contain target-language training sets.

available monolingual NLU datasets in languages other than English and Table 2 lists multilingual NLU datasets.[6] Most of these datasets have been obtained by translating pre-existing or newly created examples from English: SNIPS (Coucke et al., 2018) was translated by Castellucci et al. (2019), Bellomaria et al. (2019) into Italian; ATIS was translated by He et al. (2013) into Chinese, by Dao et al. (2021) into Vietnamese, and by Susanto and Lu (2017), Upadhyay et al. (2018), Xu et al. (2020) into 10 typologically diverse languages. However, a small group of monolingual datasets (Bonneau-Maynard et al., 2005; Zhang et al., 2017; Gong et al., 2019) were built from scratch based on original conversations in the target language.

Ideally, NLU models should be able to generalise both over languages and over domains. Most existing datasets, however, either cover multiple domains in a single language (Hakkani-Tür et al., 2016; Liu et al., 2019) or the same domain across different languages (Xu et al., 2020). Fortunately, the most recent generation of NLU datasets (Li et al., 2021; van der Goot et al., 2021; Majewska et al., 2022) is both multi-lingual and multi-domain, thus opening up the possibility to assess the true generality of current cross-lingual transfer approaches.

---

6. For completeness, we also provide a (non-exhaustive) collection of English-only NLU resources in the Appendix.

Dataset	Task	Languages	Domains	Size (# dialogues)	Type
DSTC 5 (Kim et al., 2016)	dialogue state tracking	en, zh	tourist information	35 [en] 12 [zh]	H2H
Multilingual WOZ 2.0 (Mrkšić & Vulić, 2018)	dialogue state tracking	en, de, it	restaurant booking	1200	H2H (translated)
DSTC 6 (Hori et al., 2019)	dialogue breakdown detection	en, ja	chit chat	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
GlobalWoZ (Ding et al., 2022)	dialogue state tracking	es, id, zh	8 domains (e.g., restaurants, taxi)	1500	H2H
AllWOZ (Zuo et al., 2021)	dialogue state tracking	fr, vi, pt, ko, zh, hi, th	5 domains (e.g., attraction, hotel, restaurant)	700	H2H (translated)
Multi²WOZ (Hung et al., 2022)	dialogue state tracking	ar, de, ru, zh	5 domains (e.g., attraction, hotel, restaurant)	8,000	H2H (translated)

Table 3: Multilingual DST datasets. Shaded rows correspond to datasets created from scratch, white rows to datasets translated from English. Abbreviations: human-to-machine (H2M) and human-to-human (H2H). A non-exhaustive list of English DST datasets is given in Table 10 in the Appendix.

3.2 Dialogue State Tracking (DST)

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

Cross-Lingual Transfer Methods. The Neural Belief Tracker (NBT; Mrkšić et al., 2017; Mrkšić & Vulić, 2018) is a DST model that updates its internal representation of the state of a conversation at every dialogue turn in a fully data-driven fashion. As such, NBT was the first neural approach based on word embeddings performing on a par with the models exploiting hand-crafted lexical rules. With the introduction of the multilingual WoZ dataset, Mrkšić et al. (2017) coupled NBT with cross-lingual word embeddings to enable zero-shot cross-lingual DST transfer. A body of subsequent work on retrofitting CLWEs for pure semantic similarity reported performance gains in cross-lingual transfer for DST, using NBT as the base model (Vulić et al., 2018; Glavaš & Vulić, 2018; Ponti et al., 2018b, 2019). XL-NBT (Chen et al., 2018) resorts instead to multilingual knowledge distillation (Hinton et al., 2015): the DST knowledge of the English teacher model is transferred to the target language student model by matching their latent representations of parallel words and sentences. A similar technique has been applied more recently for general sentence-level representation learning (Reimers & Gurevych, 2020).

The leaderboard of the recent *DSTC 9* challenge (Gunasekara et al., 2020), however, indicates that machine translation of the training data coupled with state-of-the-art mono-lingual DST models in the target language (Shan et al., 2020; Kim et al., 2020), known as 'translate train', outperforms zero-shot and few-shot cross-lingual transfer based on multilin-

Dataset	Task	Language(s)	Domains	Size (dialogues)
CrossWOZ (Zhu et al., 2020a)	dialogue state tracking; end-to-end;	zh	5 domains (e.g., attraction, hotel, taxi)	6012
RiSAWOZ (Quan et al., 2020)	dialogue state tracking; end-to-end;	zh	12 domains (e.g., education, car, hospitality)	11200
WMT 2020 Chat (Farajian et al., 2020)	end-to-end	de	6 domains (e.g., ordering pizza, movie tickets)	692
ViWOZ (Van et al., 2022)	NLU; dialogue state tracking; end-to-end;	vi	7 domains (e.g., attraction, bas, hospital, hotel)	5000
BiTOD (Lin et al., 2021)	end-to-end	en, zh	5 domains (e.g., attractions, hotel, restaurant, weather)	7,232
COD (Majewska et al., 2022)	NLU; dialogue state tracking; end-to-end;	en, ar, id, ru, sw	11 domains (e.g., flights, homes, music, payments)	368

Table 4: Multilingual datasets for end-to-end training. The table also includes monolingual non-English datasets. Shaded rows correspond to datasets created from scratch, white rows to datasets translated from English. A non-exhaustive list of English datasets is provided in the Appendix.

gual representations. On the other hand, DSTC 9 includes only English and Chinese, both endowed with abundant monolingual and parallel data. Hence, this finding may not hold true for other languages.

Available Datasets. Mrkšić and Vulić (2018) translated the WoZ 2.0 DST dataset (Wen et al., 2017) to German and Italian. Within the dedicated Dialogue State Tracking Challenge (later renamed Dialogue System Technology Challenges), only 3 out of 9 editions to date included multilingual DST tracks. DSTC 5 (Kim et al., 2016) evaluated DST models on zero-shot cross-lingual transfer from English (training data) to Chinese (development and test data) in the tourism domain. DSTC 6 (Hori et al., 2019) included a track on dialogue breakdown detection in chat-oriented dialogues, which required transferring knowledge from English to Japanese. Finally, as the first challenge to benchmark cross-lingual DST systems on large scale datasets, DSTC 9 (Gunasekara et al., 2020) included a track focused on transfer between English and Chinese (in both directions), using MultiWOZ 2.1 (Eric et al., 2020) as the English dataset and CrossWOZ (Zhu et al., 2020a) as the Chinese dataset. Table 3 summarises the key properties of the above-mentioned multilingual DST datasets.

3.3 Natural Language Generation (NLG)

In contrast to other sub-tasks of modular ToD, multilingual response generation for ToD has received limited attention. We thus take a broader look at multilingual NLG in general.

Traditional NLG. Traditionally, the production of sentences in languages other than English, conditioned on structured or continuous representations of the desired meaning, relied on NLG pipelines (Reiter & Dale, 1997, 2000). The last module of such pipelines is a linguistic

(surface) realiser, responsible for outputting the final surface text based on language-specific morpho-syntactic and orthographic constraints (e.g., word order, inflectional morphology). These constraints can be enforced through hand-crafted grammar-based systems (Gatt & Krahmer, 2018; Bateman, 1997; Elhadad & Robin, 1996), manually created templates (McRoy et al., 2003), and statistical methods (Filippova & Strube, 2007). To facilitate the general usage of grammar-based systems, characterised by a high level of linguistic detail, simpler realisation engines that provide syntax and morphology APIs have been developed (Gatt & 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 & Sripada, 2014), and French (Vaudry & 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 Cross-Lingual Transfer. Given the reliance of data-driven NLG models on the availability of training data and their scarcity in the vast majority of the world's languages, cross-lingual transfer methods have been leveraged to enable NLG in low-resource scenarios. They have been employed in two main ways (cf. §2.3): i) in a 'translate test' setting, an NLG system is trained on English. At evaluation time, the input text in a target language is translated into English before feeding it to the system, and the English text generated in output is translated back to the target language (Wan et al., 2010). Alternatively, ii) in a 'translate train' setting, the English training data is translated into the target language, which provides supervision to learn an NLG system directly in the target language (Shen et al., 2018; Duan et al., 2019).

Cross-Lingual Transfer via Multilingual Representations. Pretraining of multilingual sequence-to-sequence models, initially devised for neural machine translation (Liu et al., 2020; Lin et al., 2020a; Kim et al., 2020), has been successfully applied to facilitate cross-lingual transfer in other NLG applications as well. For example, Kumar et al. (2019) first pretrain a bilingual English-Hindi model through unsupervised MT (via denoising autoencoding and back-translation); they then fine-tune the model for question generation on both (large) English and (small) Hindi data. Similarly, Chi et al. (2020) pretrain a Transformer-based encoder-decoder on several languages, with denoising auto-encoding and cross-lingual masked language modelling objectives, and fine-tune it for question generation and abstractive summarisation. However, recent work also exposed a series of limitations of massively multilingual Transformer-based encoders. 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. They found that massively multilingual models (1) significantly under-perform their monolingual counterparts and (2) struggle in handling complex morphology. This somehow undermines the viability of MEs for large-scale multilingual NLG, although more recent work suggests that the general task performance of the MEs can be partially recovered through language-specific adaptation (Rust et al., 2021).

Available Datasets. Training data for NLG in languages other than English remain scarce: small datasets are available in Korean (Chen et al., 2010), Spanish (García-Méndez et al., 2019), and Czech (Dušek & 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, end-to-end dialogue modelling has emerged as a promising alternative to modular ToD. Most of the algorithms fall within the sequence-to-sequence (seq2seq) framework, reading user utterances in input and generating system responses in output (Wen et al., 2017; Madotto et al., 2018; Ham et al., 2020). Unfortunately, training reliable seq2seq models demands extensive supervision, making end-to-end dialogue systems data-hungry. However, collecting task-oriented dialogues is much more expensive and laborious than collecting open-domain conversations for training chatbots. As a result, only a few monolingual end-to-end ToD system have been developed in languages other than English, such as Chinese (Zhu et al., 2020a; Quan et al., 2020). Future work should also look into different flavours of E2E modelling in multilingual context, e.g., distinguishing between (i) *strict E2E* models, where the system leverages only natural dialogue texts, and (ii) *weak E2E* models, where the system is additionally allowed to rely on intermediate dialogue annotations (e.g., dialogue states and dialogue acts).

Finally, we list the available datasets for non-English e2e ToD in Table 4. Although this survey concentrates on ToD, we additionally list available datasets for multilingual open-domain dialogue in the Appendix.

4. Open Challenges and Future Directions

In light of the current state of multilingual ToD research, covered in §3, we now identify the main open challenges, hint to some of their possible solutions, and conjecture future directions of research. In particular, in §4.1 we analyse the linguistic diversity, idiomacity, and cultural adaptation of current datasets for multilingual ToD, the aspects which, we argue, should be prioritised in future resource creation endeavours. In §4.2, we argue that cross-lingual transfer in ToD NLU can adopt solutions devised for related NLP tasks, especially to deal with low-resource scenarios. Furthermore, we consider the compounded difficulties that multilinguality adds for fluency in text generation (especially with respect to complex morphology and code switching) in §4.3 and for collecting human evaluation in §4.4. Finally, in §4.5 we broaden the scope of our survey to multilingual speech for voice-based ToD.

4.1 Outlook for Multilingual ToD Datasets

Linguistic Diversity. Evaluating on representative language samples is key for long-term development of multilingual ToD systems. In fact, 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 in the evaluation sample are similar, cross-lingual transfer is simplified and the resulting estimates could be overly optimistic (Ponti et al., 2020). Instead, the sample of languages should ideally be diverse in terms of language family, area, and typological features (Ponti et al., 2019).

	M. TOP	M. ATIS	MultiATIS++	MTOP	xSID	XCOPA	TyDi QA	XNLI
# languages	3	3	9	6	13	11	11	15
Typology	0.20	0.29	0.33	0.29	0.37	0.41	0.41	0.39
Family	0.67	0.67	0.44	0.33	0.50	1.0	0.9	0.5
Macroareas	0	0	0	0	0	1.67	0.92	0.37

Table 5: Diversity indices of multilingual dialogue NLU datasets in terms of typology, family, and macroarea. 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).

NLU is the only component of modular ToD whose datasets cover an extensive sample of languages. We quantify the linguistic diversity of these datasets through the metrics proposed by Ponti et al. (2020), which are based on typological, family and geographical properties.[7] The sample diversity scores are shown in Table 5. For comparison, we include the most diverse datasets for other NLP tasks, such as natural language inference (XNLI; Conneau et al., 2018), question answering (TyDi QA; Clark et al., 2020), and causal commonsense reasoning (XCOPA; Ponti et al., 2020).

From Table 5, several shortcomings of the language samples within existing multilingual dialogue NLU datasets become apparent. First, they mostly originate from a single macroarea, namely Eurasia. The only exception is xSID, which contains Indonesian from Papunesia. Second, they fall short of representing a significant variety of linguistic phenomena. In particular, the languages are identical with respect to 23 out of 103 typological URIEL features. Third, only MultiATIS++ and xSID cover a number of languages comparable to non-dialogue NLP datasets, but even so, the majority of them belong to the same family (Indo-European). Datasets for other NLP tasks can therefore serve as a polestar for the selection of more diverse language samples in dialogue NLU.

Worse yet, beyond NLU, there is currently a complete lack of large dialogue datasets with full-fledged multi-turn conversations in multiple typologically diverse languages (see Table 4). If such a dataset existed, it would enable end-to-end training of multi-turn ToD systems in multiple languages and widen our understanding of similarities and differences across languages at the level of entire dialogues. In conclusion, the collection of (possibly end-to-end) datasets covering a *much wider* set of families, macroareas and typological properties could define the path to future milestones in multilingual ToD.

Idiomacity and Cultural Adaptation. In addition to linguistic diversity, another limitation of current multilingual ToD datasets stems from their creation process. In fact, training and evaluation instances are often directly translated from English. However, this might cause unwanted ramifications. First, 'translationese', the language variety of translated texts, is different from that of natural and spontaneous texts. This is due to both univer-

---

7. To measure typological diversity, we calculate the entropy of 103 binary typological features from URIEL (Littell et al., 2017) across languages, and then average across features. To measure family diversity, the number of distinct families represented in the dataset is divided by the total number of languages in the dataset. To measure geographical diversity, we calculate the entropy of the distribution across geographical macroareas of the dataset languages.

sal patterns in translation (simplification, normalisation, and explicitation) and linguistic interference: the source language spills its lexical and structural properties over the target language (Lembersky et al., 2012; Volansky et al., 2015). These artefacts, introduced by the translation procedure, could make the dataset not representative of real-life dialogue and cultural context of the target language (Hershcovich, Frank, Lent, de Lhoneux, Abdou, Brandl, Bugliarello, Piqueras, Chalkidis, Cui, et al., 2022) and instead give an edge to translation-based cross-lingual transfer. Hence, the evaluation performance becomes unreliable and excessively optimistic (Artetxe et al., 2020). Koppel and Ordan (2011) studied the differences between translated-into-English and original English texts. They demonstrate that there is a significant difference in lexical characteristics of the texts: e.g., there are some stark differences in the frequency of usage of functional words and pronouns. Recent work by Majewska et al. (2022) presents a qualitative analysis in the context of dataset creation for multilingual ToD, comparing dialogue data obtained via translation and free-form generation by native speakers of the target language. The paper presents multiple examples of the bias from English on both lexical and structural syntactic level. We refer the reader to the paper for some concrete examples.

Secondly, the information and the topics touched upon in a conversation (e.g., in the domain of airline travels, names of destinations or flight companies) may vary across cultures and locales. However, translation-based approaches reflect the perspective of the English-speaking culture, its 'presupposed' factual knowledge, and the worldview of its community of speakers (Clark et al., 2020). For these reasons, ToD benchmarks should be ideally based on original utterances grounded in the appropriate locale and culture.

Recognising this need, there have been some very recent developments in the direction of creating localised and culturally adapted ToD datasets. Namely, Ding et al. (2022) undertake an automatic approach to localisation in which the English slots are substituted by their local counterparts, obtained via Web-crawling suitable values. In contrast, Majewska et al. (2022) ask the native speakers to generate utterances based on templatic dialogue outline. The dialogue creators are encouraged to substitute English slot values with their target-language counterparts. While the former approach presents a more controlled automatic localisation procedure, the latter is more human-driven, providing a closer contact with the local community speaking the target language. Furthermore, Fitzgerald et al. (2022) localise their multilingual NLU dataset in two stages. In the first stage, native speakers were asked to translate and localise slot values; in the second stage, another group of human subjects translated or localised the entire phrase using the slot task output provided by the first worker. Going beyond research in ToD, Hershcovich et al. (2022) suggest that collecting multilingual data within (large) local communities results in culturally richer data and avoids imposing English-driven use cases. Additionally, it can reduce the per-person manual effort (by dividing the work between more people) which is often a bottleneck in the data collection process. Other potential approaches to reducing manual data creation and curation effort are briefly discussed below.

Reducing Human Effort in ToD Data Curation. Collecting dialogue datasets is notoriously hard, time-consuming and requires a lot of manual labour. In multilingual ToD, the problem scales proportionally with the number of languages. It is thus natural to seek for methods to reduce manual human effort without compromising data quality.

Firstly, several automated approaches to reduce human effort were proposed, predominantly focusing on data augmentation. These approaches, previously applied to multilingual dialogue, include (i) word or span substitution, creating code-switched data between source and target languages (Liu et al., 2020; Krishnan et al., 2021; Qin et al., 2021) or synonymous span substitution in a target language (Louvan & Magnini, 2020b)2020; and (ii) semi-supervised training with target language sentence retrieval (Razumovskaia et al., 2022).

Secondly, since direct translation still dominates multilingual ToD data collection, there have been several approaches to lower human effort in the translation procedure. In most cases translators would simultaneously annotate the datasets with slot labels and/or dialogue states, depending on the tasks the dataset covers (Mrkšić et al., 2017; Xu et al., 2020; van der Goot et al., 2021). One approach simplifies the translation process itself, which typically proceeds in two stages: (i) machine translation into the target language; (ii) manual post-editing by native speakers of the language (Zuo et al., 2021; Hung et al., 2022). One can also reduce the amount of manual annotation work by projecting labels from the source language to the target language. While label projection for sentence classification tasks is trivial, it is harder but still attainable for the word- and context-level tasks. Some of the existing methods are unsupervised and rely on parallel corpora (Dyer, Chahuneau, & Smith, 2013; Dou & Neubig, 2021). Other methods complement machine translation with a word alignment step (Jain, Paranjape, & Lipton, 2019). Once the labels have been automatically projected, one could hire annotators, native speakers of the target language, only for editing the projected slots.

4.2 Coping with Low-Resource Scenarios in NLU

Parallels with Other NLP Tasks. 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). On the other hand, slot filling can be framed as a sequence labelling (Louvan & Magnini, 2020a) or a span extraction task (Coope et al., 2020; Henderson & Vulić, 2021). Along the same lines, DST is sometimes formulated as a semantic parsing task in monolingual multi-domain settings (Cheng et al., 2020): dialogue states are represented as a hierarchical semantic structure which includes information about the domain, past actions, and the slots filled or requested.[8]

This effectively means that the standard methodological 'machinery' forged to handle low-resource languages and domains can be directly applied to joint multilingual modelling and cross-lingual transfer in ToD NLU (Ponti et al., 2019; Hedderich et al., 2021). In what follows, we provide a very brief and non-exhaustive overview of promising cutting-edge techniques that might also benefit low-resource ToD.[9] The reader should, however, still bear

---

8. 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. In multilingual ToD, they could similarly allow for cross-lingual knowledge sharing. Secondly, structured representations facilitate the integration of external knowledge. For instance, tables like those widely adopted in semantic parsing (Zhu et al., 2020b; Sun et al., 2019) may improve cross-lingual slot labelling by mapping expressions for named entities (e.g., city names) across languages.

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

in mind the core deficiencies of the current methodology in relation to multilingual ToD, as previously discussed in §2.3.

Low-resource languages should profit from annotated resources in higher-resource languages. Besides translation-based 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). Recently, unmatched performance in cross-lingual transfer has been achieved by multilingual Transformer-based encoders such as multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and mT5 (Xue et al., 2021), or cross-lingual alignment of their monolingual counterparts (Schuster et al., 2019b; Conneau et al., 2020; Cao et al., 2020a). These models can be additionally (i) extended to cover radically low-resource languages (Pfeiffer et al., 2020; Ponti et al., 2020; Hedderich et al., 2020; Üstün et al., 2020) and even languages with unseen scripts (Pfeiffer et al., 2021), or (ii) refined via few-shot learning on a small subset of target-language annotated examples (Lauscher et al., 2020; Bhattacharjee et al., 2020). However, the current performance of cross-lingual transfer to low-resource target languages (e.g., African languages, indigenous languages of the Americas) still lags dramatically behind that of transfer to high-resource target languages (Lauscher et al., 2020; Wu & Dredze, 2020; Zhao et al., 2020).

Pretrained language models can also be adaptively fine-tuned with unannotated in-domain data in both the source and the target language to pick up more task-specific knowledge, which typically results in slight performance gains (Henderson et al., 2020; Gururangan et al., 2020). Along the same lines, the entire research area focusing on domain adaptation in NLP (Kim et al., 2018; Ziser & Reichart, 2018; Rücklé et al., 2020; Ramponi & 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 ToD, we typically encounter formidably difficult "double-scarce" setups, simultaneously dealing with both low-resource domains *and* low-resource languages.

Another plausible solution to data paucity is making the multilingual ToD NLU models more robust in low-resource scenarios through data augmentation (Du et al., 2021; Xie et al., 2020): (i) at the token level with synonymy-based substitutions generated automatically (Kobayashi, 2018; Gao et al., 2019) or taken from lexico-semantic resources (Raiman & Miller, 2017; Wei & Zou, 2019; Dai & Adel, 2020), or rule-based morphological inflection (Vulić et al., 2017; Vania et al., 2019), (ii) at the sentence level by manipulating dependency trees (Ponti et al., 2018a; Şahin & Steedman, 2018), back-translating (Edunov et al., 2018), or generating synthetic adversarial examples (Garg & Ramakrishnan, 2020; Morris et al., 2020); (iii) at the annotation level by automatically labelling more sentences, filtering them, and using them as silver training data (Onoe & Durrett, 2019; Du et al., 2021).

Distant and weak supervision methods (Luo et al., 2017; Alt et al., 2019), often leveraged to share knowledge between structurally similar low-resource NER models (e.g., Cao et al., 2019; Mayhew et al., 2019; Lison et al., 2020), might also prove beneficial for slot filling, as a sequence labelling task. Meta-learning frameworks such as MAML (Finn et al., 2017) have also emerged recently as strategies to tackle low-resource cross-lingual and cross-domain transfer for several NLP classification tasks (Nooralahzadeh et al., 2020; van der Heijden et al., 2021); however, they are yet to find their application in (multilingual) ToD systems.
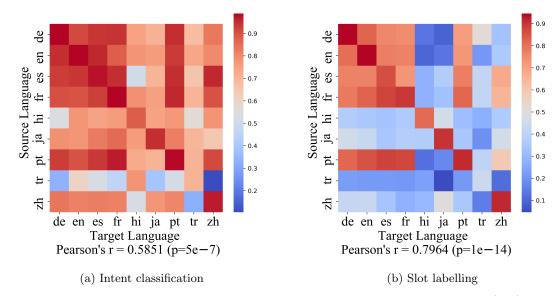
(a) Intent classification

(b) Slot labelling

Figure 3: Pair-wise cross-lingual transfer performances on intent classification (left) and slot labelling (right) between each source language (y axis) and each target language (x axis).

Feature Group	Intent Classification		Slot Labelling	
	Pearson's r	p-value $\leq$	Pearson's r	p-value $\leq$
Phonology	-0.2467	5×10^{-2}	-0.2504	5×10^{-2}
Geography	-0.2263	5×10^{-1}	-0.3270	5×10^{-2}
Phylogenetics	-0.4895	5×10^{-4}	-0.6122	5×10^{-7}
Syntax	-0.5131	5×10^{-5}	-0.6919	5×10^{-11}

Table 6: Correlation between single-source zero-shot transfer performance in NLU and source–target language distance according to different groups of properties. Phonology and syntax features come from URIEL (Littell et al., 2017), phylogenetic features from Glottolog (Hammarström et al., 2017).

While taking the inspiration from related NLP tasks is a natural approach to speed up the progress in multilingual ToD, it is important to also take into account the aspects where the development of ToD systems differs from working with related NLP tasks. Firstly, the models need to operate with and integrate (multi-speaker) dialogue history as conversational context, which implies processing long sequences. Most pretrained Transformer-based models cannot handle such long input, as they are limited by the self-attention operation which scales quadratically with the length of the input sequence. Models such as LongFormer (Beltagy et al., 2020) and BigBIRD (Zaheer et al., 2020) propose modifying the self-attention with localised attention and sparse attention, respectively, for long sequence processing. These efficient Transformer-based models are yet to find their application in the context of multilingual ToD. Further, such heuristic solutions might still be suboptimal for languages with word order and manner of speaking which is radically different to English and other major languages (Ruder, Vulić, & Søgaard, 2022).

The inherent conversational nature of ToD tasks poses additional challenges to direct transfer of methods developed for other NLP tasks to multilingual dialogue. For instance, models pretrained with response selection or retrieval tasks on conversational data are more suited for modelling dialogue than their counterparts pretrained with language modelling task (Mehri et al., 2020; Coope et al., 2020). This indicates that we might witness substantial improvements in multilingual dialogue if standard multilingual encoders are replaced with ones pretrained or fine-tuned on conversational data with dialogue-oriented objectives (Henderson et al., 2020; Hung et al., 2022). There are also linguistic phenomena (e.g., code switching) which are more widespread in the colloquial speech than in the written format. The methods developed for multilingual ToD must also be able to deal with utterances containing such phenomena (see also later §4.3).

Source Selection for Cross-Lingual Transfer. 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, based on prior work in general-purpose cross-lingual NLP, we detect three crucial gaps which require more attention in future work and that may play an instrumental role in the final task performance: **1**) the choice of source language(s) (Zoph et al., 2016; Dabre et al., 2017; Lin et al., 2019), as recently hinted at for multilingual ToD NLU by Krishnan et al. (2021); **2**) harnessing multiple source languages rather than a single one (Zoph et al., 2016; Pan et al., 2020; Wu et al., 2020b; Ansell et al., 2021); **3**) few-shot transfer with a small number of target-language examples, as opposed to fully zero-shot transfer (Lauscher et al., 2020). In other words, the go-to option of always transferring from English in a zero-shot fashion might be sub-optimal for a large number of target languages.

In order to adduce evidence in support of these two conjectures, we conduct a series of preliminary empirical studies. For our experiments, we focus on the two core dialogue NLU tasks, intent classification and slot filling, as documented in MultiATIS++ (see Table 2). As a neural architecture, we adopt the *de facto* standard of a pretrained encoder (multilingual BERT) and a classifier head (see 3.1).[10]

First, we investigate, for every target language in MultiATIS++, its most compatible source language for cross-lingual transfer. We fine-tune our model on a single source at a time and evaluate it in a zero-shot setting on all targets. The resulting performances are pictured in the heat maps of Figure 3. It emerges that English (en) is not always the best source language (e.g., it is surpassed by es for transfer to fr in intent classification). Moreover, the set of suitable sources is task-dependent. In particular, the knowledge needed for intent classification is more amenable to be transferred to languages outside a language family (ja, tr, zh are not Indo-European) and with different scripts (hi, ja, zh are not written with the Latin alphabet) than slot filling, which is more language-sensitive. As a consequence, scores in intent classification have generally a smaller variance. Finally, as expected, the training dataset size plays a significant role: Turkish and Hindi may count on less annotated examples, which partially explains their lower figures.

---

10. As the baseline model, uncased multilingual version of BERT-base (Devlin et al., 2019) is used. We use Adam optimizer (Kingma & Ba, 2015) with learning rate 5e-5 and warm-up ratio of 0.1. The models are trained with the batch size of 32 for 20 epochs. The models are implemented using gluonNLP package (Guo, He, He, Lausen, Li, Lin, Shi, Wang, Xie, Zha, et al., 2020).

	de	en	es	fr	hi	ja	pt	tr	zh	AVG
					ZERO-SHOT					
typ-closest	95.30	91.60	93.84	96.87	78.11	77.72	95.30	67.80	64.39	75.91
multi-source	97.20	93.95	97.20	89.59	84.61	86.11	92.16	80/57	82.31	89.30
ensemble	92.50	90.26	96.64	95.18	77.88	77.04	95.30	75.04	84.99	87.20
					FEW-SHOT					
typ-closest	94.66	95.07	94.92	95.19	88.86	91.08	94.29	87.61	92.35	92.67
multi-source	95.78	96.38	94.92	95.41	91.56	92.98	93.95	89.13	93.43	93.73

Table 7: Intent classification results (accuracy) on MultiATIS++ (Xu et al., 2020) for 3 transfer methods. Few-shot results are averaged across 3 runs.

	de	en	es	fr	hi	ja	pt	tr	zh	AVG
					ZERO-SHOT					
typ-closest	80.87	82.50	88.20	87.27	22.96	50.11	77.48	51.11	45.80	65.14
multi-source	87.45	91.05	86.39	82.61	68.25	74.92	87.25	60.09	69.03	78.56
ensemble	82.75	85.05	77.56	76.19	14.14	9.44	74.00	45.63	37.29	55.78
					FEW-SHOT					
typ-closest	91.29	92.67	84.93	86.56	81.99	86.45	88.43	73.44	86.57	85.81
multi-source	92.51	93.76	86.25	87.43	85.46	87.47	89.08	81.34	88.29	87.95

Table 8: Slot labelling results (F-1) on MultiATIS++ (Xu et al., 2020) based on 3 transfer methods. Few-shot results are averaged across 3 runs.

To shed further light onto source selection, we use the cosine similarity between URIEL feature vectors (Littell et al., 2017), which capture syntactic, phonological, geographic and phylogenetic language properties, as a proxy for the similarity between languages. Based on this metric, we measure the Pearson's correlation between linguistic similarity and the cross-lingual transfer performances reported in Figure 3. The results of this analysis are summarised in Table 6. We discover moderate and strong correlations between language similarity and performance in intent classification and slot labelling, respectively. The highest correlation coefficients are found for phylogenetic and syntactic properties, which can be therefore considered reliable indicators of compatibility for language pairs.

Secondly, we run an additional set of experiments on intent classification and slot filling in MultiATIS++, in order to compare alternative cross-lingual transfer methods and data paucity regimes. In particular, for every target language we compare three transfer methods: i) *typ-closest*, where we select the model fine-tuned on the typologically closest source language, as determined by the cosine similarity of URIEL (Littell et al., 2017) typological features;[11] ii) *multi-source*, where the model is fine-tuned on all languages but the target; iii) *ensemble*, where the predictions are determined by 8 independent models, each separately trained on a single source language, through majority voting. As for different data paucity regimes, in addition to zero-shot transfer, we also consider few-shot adaptation in the spirit

---

11. A table of the typologically closest language for each target language is provided in the Appendix.

of Lauscher et al. (2020): for every target language, we continue fine-tuning its *typ-closest* or *multi-source* models on 500 additional examples sampled from its training split.

The results for intent prediction and slot labelling are summarised in Tables 7–8. Comparing different transfer methods, *multi-source* has a significant edge over both *typ-closest* (+13.39 points in accuracy for intent classification, +13.42 in F_1 score for slot filling) and *ensemble* (+2.10 for intent classification, +22.78 for slot filling). Note that the performance drop of *ensemble* model in slot labelling proves again its sensitivity to the source languages, as the majority of the voters is unreliable. However, jointly training over all sources not only remedies to this, but further cements the cross-lingual synergy by letting languages borrow statistical strength from one another. This corroborates previous findings (Xu et al., 2020) that multi-source fine-tuning is beneficial even when full training sets in the target language is available. Furthermore, this is evidence that multilingual datasets are crucial not only for the evaluation of multilingual dialogue systems, but also for better training them.

Comparing different data scarcity regimes, the results in Tables 7–8 show that similar patterns emerge in few-shot performance on both tasks. However, in this case *typ-closest* almost completely bridges the gap with *multi-source*. This hints at the fact that even (larger) data from typologically diverse sources is no match for (smaller) in-domain supervision. Even more importantly, however, the performance of both transfer methods dramatically increases over their zero-shot counterparts.

The results of the study provide an interesting insight on cross-lingual NLU with single-source and multi-source training. In line with previous work (Ansell et al., 2021), they show that multi-source finetuning yields better performance than single-source, even when the single source is a language linguistically close to the target language. We assume that the reason is that larger linguistic diversity of training data leads to more generalised representations enabling the model to perform better on an unseen language. From a more practical perspective, with the limited annotation resources one should consider splitting it between more languages rather than annotating a large amount of data in one language only.

As a general finding, this small study suggests that: **1)** carefully picking source languages; **2)** transferring knowledge from multiple sources; and **3**) adapting models to few target examples may steer and substantially improve dialogue NLU performance in the future. All these aspects, in turn, ultimately depend on the ability to wisely balance the annotation budget across typologically diverse languages when creating multilingual datasets.

4.3 Fluency of Generated Language: Rich Morphology and Code Switching

Besides coping with a wide cross-lingual variability of user utterances in the NLU components, multilingual ToD systems also need to produce accurate *and* fluent responses suited for the target language. Given the dialogue history and constrained by the domain, the NLG module should output a response which is articulate, sounds native to the user, and fits in the given context, without breaking the flow of the multi-turn conversation (Garbacea & Mei, 2020).

There are challenges which are common to both ToD and Natural Language Generation tasks in general, such as complex morphology (e.g., fusional languages like the Slavic genus, agglutinative languages such Finnish or Turkish, polysynthetic languages like Inuktitut). Generating fluent and grammatical text in such morphologically rich languages is naturally much more arduous than in isolating languages without inflectional and derivative morphology

(e.g., Vietnamese; Kunchukuttan et al., 2014; Gerz et al., 2018). This stems from the inability to hold all the possible word forms in the vocabulary with traditional word embeddings or from word "over-segmentation" with recent subword-based pretrained Transformers (Rönnqvist et al., 2019; Rust et al., 2021). NLG in morphologically rich languages can benefit from dedicated language-specific tokenizers (Rust et al., 2021), the incorporation of linguistic features (Klemen et al., 2020), or through multi-tasking, predicting word and morphological information simultaneously (Passban et al., 2018).[12]

Furthermore, there are linguistic phenomena specific to informal, colloquial language. For instance, code-switching is a phenomenon where interlocutors shift from one language to another during the conversation (Sankoff & Poplack, 1981). It has been shown that code-switching might even improve the chances of successful task completion and the system's perceived friendliness (Ahn et al., 2020). Banerjee and Khapra (2019) show that structure-aware generation is effective for code-switched texts, even when dependency parsers are not available. Furthermore, Khanuja et al. (2020) maintain that, to tailor cross-lingual models such as mBERT for code-switching settings, they should be *fine-tuned* on code-switched data, since the lexical distribution in code-switched language is different from the union of two languages. Additionally, prior work demonstrate that *pretraining* Transformer-based encoders on conversational data leads to significant improvements on all dialogue tasks (Henderson et al., 2020; Mehri et al., 2020). Unfortunately, however, large code-switching dialogue corpora are not available yet.

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 justifies the need to conduct human-centred evaluations in order to reliably trace any real improvements in the perceived user-friendliness and the general eloquence of ToD systems in different languages. This leads us to the next challenge, discussed in §4.4.

4.4 Human-Centred Evaluation of Multilingual ToD Systems

A crucial compass to chart 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: accuracy and/or F_1 score for intent classification, F_1 score for slot labelling, or joint goal accuracy for DST. Recently, DialoGLUE (Mehri et al., 2020), a benchmark for ToD systems based on automated metrics, has been proposed. Unfortunately, DialoGLUE is available only in English. We thus hope that, similar to recent benchmarks for general-purpose (i.e., non-dialogue) cross-lingual NLU such as XTREME(-R) (Hu et al., 2020; Ruder et al., 2021) and XGLUE (Liang et al., 2020), future work will strive to building comprehensive and community-supported multilingual ToD benchmarks. An additional aspect to consider when automatically evaluating multilingual ToD systems is the confusion between low cross- and multi-lingual performance and cultural variance. In other words, one needs to differentiate between the models performing badly due to their inability to support cross-lingual transfer

---

12. Another problem in cross-lingual setups concerns the adaptation to different word orders, where NLG might be directly informed by typological information through structural interventions (Ponti et al., 2018a) or re-ordering (Daiber et al., 2016).

linguistically versus their inability to transfer between different cultural contexts and realities (Hershcovich et al., 2022).

Evaluation of ToD systems still adds another layer of difficulty, typically not encountered in the above-mentioned benchmarks for general-purpose NLP tasks. In short, received wisdom cautions that strong performance on automated metrics does not always positively correlate with the overall user satisfaction with the system (Liu et al., 2016; Smith et al., 2022). This means that human evaluation remains the most faithful 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, collecting human feedback is money- and time-consuming. Multilingual ToD systems inherit the same costs but also pose new questions for widespread human evaluation protocols. Firstly, the expenses for hiring qualified users scales linearly with the number of languages. Secondly, especially for lower-resource languages, sometimes it is hard (or impossible) to find fluent speakers on commonly used platforms such as Amazon Mechanical Turk. Thirdly, when hiring evaluators from different countries, one needs to consider whether cultural differences subsist which may alter the way user satisfaction is perceived or reported.

Another issue is the current lack of standardisation of human-based evaluation in multilingual ToD. Howcroft et al. (2020) show that even the definitions of the terms 'accuracy' and 'fluency' for NLG differ between researchers, which makes the comparison between two evaluation runs virtually impossible. To this end, GENIE (Khashabi, Stanovsky, Bragg, Lourie, Kasai, Choi, Smith, & Weld, 2021) and GEM (Gehrmann, Adewumi, Aggarwal, Ammanamanchi, Aremu, Bosselut, Chandu, Clinciu, Das, Dhole, et al., 2021) were proposed, with the central aim of standardising protocols for more reproducible human evaluation. While creating such standardised protocols is feasible and more straightforward in a monolingual context, defining a unified protocol is still an open question in multilingual contexts. User satisfaction with a dialogue system is highly dependent on what the user sees as good, useful, appropriate versus bad, incorrect, pointless. These notions are highly culture-dependent (Kendall, Linden, & Murray, 2005). It means that an appropriate response of a dialogue system in the Anglosphere can be wrong, impolite and even aggressive when ported without any adaptation to a different culture (Hershcovich et al., 2022). Thus, accounting for ethics, cross-cultural differences and social norms is crucial when evaluating multilingual NLP in general (Solaiman & Dennison, 2021) and dialogue systems in particular. A potential approach to make the evaluation protocols more informed of the culture is to involve social scientists into multilingual ToD research. They might provide guidance on how to ground the protocol in the target language culture, including taboo topics and sensitive discussion points that should be avoided by the systems.

4.5 Voice-Based Multilingual Dialogue?

This survey, following the current mainstream in monolingual and multilingual ToD modelling, has hitherto focused on text-based input. However, the assumption of working with clean, complete, and fluent input text may be naïve: in fact, it underestimates the errors cascading from imperfect automatic speech recognition (ASR) onto the subsequent text-based modules.

While previous work on English ToD paid attention to recovering from ASR errors and incorporating imperfect ASR output into the text-based pipeline (Henderson et al., 2014a; Mrkšić et al., 2017; Ohmura & 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 the true abilities of voice-based ToD systems are likely misconstrued.

ASR and speech-to-text synthesis are wide research fields in their own right, also seeking an expansion towards multilinguality as a long-standing and crucial research goal (Le & Besacier, 2009; Ghoshal et al., 2013; Conneau et al., 2021, *inter alia*). The current mainstream ASR paradigm for cross-lingual transfer also relies on large pretrained Transformer-based multilingual models (Conneau et al., 2021; Pratap et al., 2020a). Similar to multilingual BERT or XLM-R in the text domain, a heavily parameterised ASR model is trained on a large multilingual corpus of pure speech, and then fine-tuned with smaller amounts of speech-to-text transcriptions 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.[13] Yet, cutting-edge fully unsupervised models, such as wav2vec-U (Baevski et al., 2021), hold promise to extend ASR even to languages without transcribed speech available. In particular, after the standard pre-training on unlabelled speech, wav2vec-U representations are fed to a Generative Adversarial Network, which tries to discriminate between real and generated phoneme sequences. The encouraging early results of this method on low-resource languages will need to be validated in future work, and its coupling with multilingual ToD is yet to come.

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 & Dusek, 2020), even smaller than what multilingual ASR currently supports. However, breakthroughs in unsupervised TTS may be decisive to widen language coverage (Zhang & Lin, 2020).

This effectively means that voice-based ToD is currently out of reach for the overwhelming majority of the world's languages (Joshi et al., 2020). More generally, in the pursuit of wider-scale and democratised ToD technology, we advocate an even 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. In fact, any future developments of multilingual ToD are also tightly coupled with parallel developments in multilingual ASR and TTS as standalone research areas.

4.6 Other Related Areas

We have attempted to cover multiple threads of research connected to multilingual conversational AI, as an extremely wide multi-disciplinary and multi-layered field. However, we also acknowledge that there are other areas related to the development and deployment of full-fledged and engaging multilingual ToD systems which remained out of our main focus. These other directions include (but are not limited to): **1)** making ToD systems more flexible and empathetic by relying on implicit conversational cues and (multilingual)

---

13. 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) or under different recording circumstances (Salesky et al., 2021).

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).[14] Stepping a bit further away, it is also quite reasonable to assume that further advances in machine translation for massively multilingual and low-resource settings (Siddhant et al., 2020a, 2020b; Garcia et al., 2020; Fan et al., 2021) will also (continue to) have substantial impact on multilingual ToD by enhancing cross-lingual transfer capability.

5. Conclusion

Enabling machines to converse similarly to humans is one of the central goals of AI. Achieving this in a multitude of the world's languages is an even more ambitious challenge. In this work, we have presented an overview of the current challenges and efforts, including the state-of-the-art methods and available datasets, and future directions concerning multilingual task-oriented dialogue (ToD) systems. In light of this survey, we can now attempt to answer the questions first posed in §1 as follows:

(**A1**) Despite gaining much more traction recently, the availability of multilingual ToD **datasets** is still scarce, especially for NLG and end-to-end dialogue. While NLU is better documented, the datasets dedicated to this task often lack typological diversity in their language sample and idiomacity in their conversations (i.e., by avoiding 'translationese' and cultural biases).

(**A2**) Given the paucity of multilingual ToD data, the most established **methods** include model transfer based on massively multilingual encoders and translation-based transfer (especially for NLG). Moreover, based on our experimental findings in ToD NLU, we identify multi-source transfer and few-shot fine-tuning as two promising solutions to substantially boost performance.

(**A3**) While it should remain aware of major (dis)similarities, multilingual ToD can and should borrow from other **related fields** of NLP the well-oiled 'machinery' tried and tested to alleviate resource-poor settings. In particular, strategies to extend the language coverage of massively multilingual encoders, unsupervised domain adaptation, data augmentation, meta learning, and conditional generation may prove essential to extend the reach of multilingual ToD systems.

(**A4**) The **future challenges** of multilingual ToD include (but are not limited to): i) the generation of fluent text in morphologically complex languages and code-switching scenarios; ii) the refinement of strategies to collect expensive annotated data and enable human-centred evaluation protocols; iii) a tighter integration of speech-based and text-based modules, as recent unsupervised learning techniques hold promise to enable ASR and TTS for a plethora of (resource-poor) languages.

---

14. Besides providing additional (situational) context to dialogues in general, multi-modal modelling might also be advantageous to multilingual settings. Indeed, visual input (e.g., images, videos) can also serve as a naturally occurring interlingua for cross-lingual alignment (Kiela et al., 2015; Gella et al., 2017; Rotman et al., 2018; Sigurdsson et al., 2020).

With these findings, we aim to inspire more work in these important areas. In the long run, we hope that our overview will contribute to foster and guide future developments in ToD towards truly multilingual and inclusive conversational AI.

Finally, an additional contribution of this work, potentially useful to other researchers and practitioners interested in this emerging field, is an up-to-date online repository of all the datasets falling within the scope of multilingual ToD, which is available at `https://github.com/evgeniiaraz/datasets_multiling_dialogue`.

Acknowledgments

Appendix A. English NLU Datasets

Dataset	Task	Language	Domains	Size	# intents	# slots
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., 2019)	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 9: English NLU datasets. This list is non exhaustive.

Appendix B. English DST Datasets

Dataset	Task	Language	Domain	Size (dialogues)	H2H / H2M
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., 2017)	dialogue state tracking	en	restaurant booking	1200	H2H

Table 10: Englsh DST datasets. This list is non exhaustive. Abbreviations: H2M – human-to-machine; H2H – human-to-human.

Appendix C. English End-to-End Datasets

Dataset	Task	Language	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., 2018)

Table 11: English datasets for end-to-end training. H2H means human-to-human dialogues.

Appendix D. Multilingual Chit-Chat Datasets

Dataset	Task	Language(s)	Domain	Size (dialogues)	Comments
MONOLINGUAL DATASETS					
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;
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 12: Multilingual chit-chat datasets for end-to-end training. H2H means human-to-human dialogues.

Appendix E. Typologically Closest Source Language for Each Target Language in MultiATIS++

Target Language	de	en	es	fr	hi	ja	pt	tr	zh
Linguistically Closest Source Language	en	de	pt	pt	zh	zh	es	de	ja

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. arXiv preprint arXiv:2001.09977.

Agarwal, S., Bui, T., Lee, J.-Y., Konstas, I., & Rieser, V. (2020). History for visual dialog: Do we really need it?. In *Proceedings of ACL 2020*, pp. 8182–8197.

Ahn, E., Jimenez, C., Tsvetkov, Y., & Black, A. W. (2020). What code-switching strategies are effective in dialog systems?. In *Proceedings of SCIL 2020*, pp. 254–264.

AlAmri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T. K., Hori, C., Anderson, P., Lee, S., & Parikh, D. (2019). Audio visual scene-aware dialog. In *Proceedings of CVPR 2019*, pp. 7558–7567.

Alt, C., Hübner, M., & Hennig, L. (2019). Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of ACL 2019*, pp. 1388–1398.

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

Ansell, A., Ponti, E. M., Pfeiffer, J., Ruder, S., Glavaš, G., Vulić, I., & Korhonen, A. (2021). MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of EMNLP 2021*, pp. 4762–4781.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G. F., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019). Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, *abs/1907.05019*.

Arora, A., Shrivastava, A., Mohit, M., Lecanda, L. S.-M., & Aly, A. (2020). Cross-lingual transfer learning for intent detection of Covid-19 utterances. OpenReview.

Artetxe, M., Labaka, G., & Agirre, E. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, pp. 789–798.

Artetxe, M., Labaka, G., & Agirre, E. (2020). Translation artifacts in cross-lingual transfer learning. In *Proceedings of the EMNLP 2020*, pp. 7674–7684.

Baevski, A., Hsu, W.-N., Conneau, A., & Auli, M. (2021). Unsupervised speech recognition. In *Advances in Neural Information Processing Systems 2021*, Vol. 34.

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 127–135, Honolulu, Hawaii. Association for Computational Linguistics.

Banerjee, S., & Khapra, M. M. (2019). Graph convolutional network with sequential attention for goal-oriented dialogue systems. *TACL*, *7*, 485–500.

Bastianelli, E., Vanzo, A., Swietojanski, P., & Rieser, V. (2020). SLURP: A spoken language understanding resource package. In *Proceedings of EMNLP 2020*, pp. 7252–7262.

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

Bellomaria, V., Castellucci, G., Favalli, A., & Romagnoli, R. (2019). Almawave-SLU: A new dataset for SLU in Italian. *CoRR*, *abs/1907.07526*.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *CORR*, *2004.05150*.

Bhattacharjee, K., Ballesteros, M., Anubhai, R., Muresan, S., Ma, J., Ladhak, F., & Al-Onaizan, Y. (2020). To BERT or not to BERT: Comparing task-specific and task-agnostic semi-supervised approaches for sequence tagging. In *Proceedings of EMNLP 2020*, pp. 7927–7934.

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

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

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., & Mostefa, D. (2005). Semantic annotation of the French media dialog corpus. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

Budzianowski, P., & Vulić, I. (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*, pp. 15–22.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gasic, M. (2018). MultiWOZ-A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP 2018*, pp. 5016–5026.

Bunk, T., Varshneya, D., Vlasov, V., & Nichol, A. (2020). Diet: Lightweight language understanding for dialogue systems. *CORR*, *2004.09936*.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., & Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 4506–4517.

Cao, S., Kitaev, N., & Klein, D. (2020a). Multilingual alignment of contextual word representations. In *Proceedings of ICLR 2020*.

Cao, Y., Lu, K., Chen, X., & Zhang, S. (2020b). Adaptive dialog policy learning with hindsight and user modeling. In *Proceedings of SIGDIAL 2020*, pp. 329–338.

Cao, Y., Hu, Z., Chua, T.-s., Liu, Z., & Ji, H. (2019). Low-resource name tagging learned with weakly labeled data. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 261–270.

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, Online.

Cascallar-Fuentes, A., Ramos-Soto, A., & Bugarín Diz, A. (2018). Adapting SimpleNLG to Galician language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 67–72, Tilburg University, The Netherlands. Association for Computational Linguistics.

Castellucci, G., Bellomaria, V., Favalli, A., & Romagnoli, R. (2019). Multi-lingual intent detection and slot filling in a joint BERT-based model. arXiv preprint arXiv:1907.02884.

Chen, D. L., Kim, J., & Mooney, R. J. (2010). Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, *37*, 397–435.

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations*, *19*(2), 25–35.

Chen, Q., Zhuo, Z., & Wang, W. (2019). BERT for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909.

Chen, W., Chen, J., Su, Y., Wang, X., Yu, D., Yan, X., & Wang, W. Y. (2018). Xl-nbt: A cross-lingual neural belief tracking framework. In *Proceedings of EMNLP 2018*, pp. 414–424.

Cheng, J., Agrawal, D., Martínez Alonso, H., Bhargava, S., Driesen, J., Flego, F., Kaplan, D., Kartsaklis, D., Li, L., Piraviperumal, D., Williams, J. D., Yu, H., Ó Séaghdha, D., & Johannsen, A. (2020). Conversational semantic parsing for dialog state tracking. In *Proceedings of EMNLP 2020*, pp. 8107–8117.

Cheyer, A., & Guzzoni, D. (2006). Method and apparatus for building an intelligent automated assistant. Tech. rep., EPFL.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X.-L., & Huang, H. (2020). Cross-lingual natural language generation via pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7570–7577.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*, *8*, 454–470.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of INTERSPEECH 2021*, pp. 2426–2430.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020*, pp. 8440–8451.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pp. 2475–2485.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., & Stoyanov, V. (2020). Emerging cross-lingual structure in pretrained language models. In *Proceedings of ACL 2020*, pp. 6022–6034.

Coope, S., Farghly, T., Gerz, D., Vulić, I., & Henderson, M. (2020). Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of EMNLP 2020*, pp. 107–121.

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., et al. (2018). SNIPS voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR, abs/1907.07526*.

Cuayáhuitl, H. (2017). Simpleds: A simple deep reinforcement learning dialogue system. In *Dialogues with social robots*, pp. 109–118. Springer.

Dabre, R., Nakagawa, T., & Kazawa, H. (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*, pp. 282–286.

Dai, X., & Adel, H. (2020). An analysis of simple data augmentation for named entity recognition. In *Proceedings of COLING 2020*, pp. 3861–3867.

Daiber, J., Stanojević, M., & Sima'an, K. (2016). Universal reordering via linguistic typology. In *Proceedings of COLING 2016*, pp. 3167–3176.

Dao, M. H., Truong, T. H., & Nguyen, D. Q. (2021). Intent detection and slot filling for Vietnamese. In *Proceedings of INTERSPEECH 2011*, pp. 4698–4702.

De Oliveira, R., & Sripada, S. (2014). Adapting simplenlg for Brazilian Portuguese realisation. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pp. 93–94.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. C. (2017). GuessWhat?! visual object discovery through multi-modal dialogue. In *Proceedings of CVPR 2017*, pp. 4466–4475.

Denecke, K., Tschanz, M., Dorner, T. L., & May, R. (2019). Intelligent conversational agents in healthcare: hype or hope. *Stud Health Technol Inform, 259*, 77–84.

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review, 54*(1), 755–810.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, Vol. 1, pp. 4171–4186.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., Williams, J., Pineau, J., Burtsev, M., & Weston, J. (2020). The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pp. 187–208.

Ding, B., Hu, J., Bing, L., Aljunied, S. M., Joty, S., Si, L., & Miao, C. (2022). GlobalWoZ: Globalizing MultiWoZ to Develop Multilingual Task-Oriented Dialogue Systems. In *Proceedings of ACL 2022*, pp. 1639–1657.

Dou, Z.-Y., & Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of EACL 2021*, pp. 2112–2128.

Du, J., Grave, E., Gunel, B., Chaudhary, V., Celebi, O., Auli, M., Stoyanov, V., & Conneau, A. (2021). Self-training Improves Pre-training for Natural Language Understanding.. pp. 5408–5418.

Duan, X., Yin, M., Zhang, M., Chen, B., & Luo, W. (2019). Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of ACL 2019*, pp. 3162–3172.

Durrett, G., Pauls, A., & Klein, D. (2012). Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1–11, Jeju Island, Korea. Association for Computational Linguistics.

Dušek, O., & Jurčíček, F. (2019). Neural generation for Czech: Data and baselines. In *The 12th International Conference on Natural Language Generation*, pp. 563–574.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL-HLT 2013*, pp. 644–648.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of EMNLP 2018*, pp. 489–500.

El Asri, L., He, J., & Suleman, K. (2016). A sequence-to-sequence model for user simulation in spoken dialogue systems. In *Proceedings of INTERSPEECH 2016*, pp. 1151–1155.

Elhadad, M., & Robin, J. (1996). An overview of SURGE: A reusable comprehensive syntactic realization component. In *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp. 70–74.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A. K., Ku, P., & Hakkani-Tür, D. (2020). Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of LREC 2020*, pp. 422–428.

Eric, M., Krishnan, L., Charette, F., & Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. In *Proceedings of SIGDIAL 2017*, pp. 37–49.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*, 1–48.

Farajian, M. A., Lopes, A. V., Martins, A. F. T., Maruf, S., & Haffari, G. (2020). Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 65–75, Online. Association for Computational Linguistics.

Filippova, K., & Strube, M. (2007). Generating constituent order in German clauses. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 320–327, Prague, Czech Republic. Association for Computational Linguistics.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML 2017*, pp. 1126–1135.

Fitzgerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., & Natarajan, P. (2022). MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. *CORR*, *2204.08582*.

Gao, F., Zhu, J., Wu, L., Xia, Y., Qin, T., Cheng, X., Zhou, W., & Liu, T.-Y. (2019). Soft contextual data augmentation for neural machine translation. In *Proceedings of ACL 2019*, pp. 5539–5544.

Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 607–613, Brussels, Belgium. Association for Computational Linguistics.

Garbacea, C., & Mei, Q. (2020). Neural language generation: Formulation, methods, and evaluation. arXiv preprint arXiv:2007.15780.

Garcia, X., Siddhant, A., Firat, O., & Parikh, A. P. (2020). Harnessing multilinguality in unsupervised machine translation for rare languages. *CoRR*, *abs/2009.11201*.

García-Méndez, S., Fernández-Gavilanes, M., Costa-Montenegro, E., Juncal-Martínez, J., & Javier González-Castaño, F. (2019). A library for automatic natural language generation of spanish texts. *Expert Systems with Applications*, *120*, 372–386.

Garg, S., & Ramakrishnan, G. (2020). BAE: BERT-based adversarial examples for text classification. In *Proceedings of EMNLP 2020*, pp. 6174–6181.

Gatt, A., & Krahmer, E. (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.

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

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., et al. (2021). The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pp. 96–120.

Gella, S., Sennrich, R., Keller, F., & Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. In *Proceedings of EMNLP 2017*, pp. 2839–2845.

Gerz, D., Vulić, I., Ponti, E. M., Reichart, R., & Korhonen, A. (2018). On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of EMNLP 2018*, pp. 316–327.

Ghoshal, A., Swietojanski, P., & Renals, S. (2013). Multilingual training of deep neural networks. In *Proceedings of ICASSP 2013*, pp. 7319–7323.

Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019). How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL 2019*, pp. 710–721.

Glavaš, G., & Vulić, I. (2018). Explicit retrofitting of distributional word vectors. In *Proceedings of ACL 2018)*, pp. 34–45.

Glavaš, G., & Vulić, I. (2020). Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of ACL 2020*, pp. 7548–7555.

Gong, Y., Luo, X., Zhu, Y., Ou, W., Li, Z., Zhu, M., Zhu, K. Q., Duan, L., & Chen, X. (2019). Deep cascade multi-task learning for slot filling in online shopping assistant. In *Proceedings of AAAI 2019*, Vol. 33, pp. 6465–6472.

Goo, C.-W., Gao, G., Hsu, Y.-K., Huo, C.-L., Chen, T.-C., Hsu, K.-W., & Chen, Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of NAACL-HLT 2018*, Vol. 2, pp. 753–757.

Gritta, M., Hu, R., & Iacobacci, I. (2022). CrossAligner & Co: Zero-Shot Transfer Methods for Task-Oriented Cross-lingual Natural Language Understanding. In *Findings of ACL 2022*, pp. 4048–4061.

Gritta, M., & Iacobacci, I. (2021). XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of ACL-IJCNLP 2021*, pp. 371–381. Association for Computational Linguistics.

Gunasekara, C., Kim, S., D'Haro, L. F., Rastogi, A., Chen, Y.-N., Eric, M., Hedayatnia, B., Gopalakrishnan, K., Liu, Y., Huang, C.-W., et al. (2020). Overview of the 9th Dialog System Technology Challenge: DSTC9. arXiv preprint arXiv:2011.06486.

Guo, D., Tur, G., Yih, W.-t., & Zweig, G. (2014). Joint semantic utterance classification and slot filling with recursive neural networks. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 554–559. IEEE.

Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S., et al. (2020). Gluoncv and gluonnlp: deep learning in computer vision and natural language processing.. *J. Mach. Learn. Res.*, *21*(23), 1–7.

Gupta, N., Tur, G., Hakkani-Tur, D., Bangalore, S., Riccardi, G., & Gilbert, M. (2005). The at&t spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 213–222.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL 2020*, pp. 8342–8360.

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

Hakkani-Tür, D., Tür, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L., & Wang, Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of Interspeech 2016*, pp. 715–719.

Ham, D., Lee, J.-G., Jang, Y., & Kim, K.-E. (2020). End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of ACL 2020*, pp. 583–592.

Hammarström, H., Forkel, R., & Haspelmath, M. (2017). Glottolog 3.0. Tech. rep., Max Planck Institute for the Science of Human History.

Hardy, H., Strzalkowski, T., Wu, M., Ursu, C., Webb, N., Biermann, A., Inouye, R. B., & McKenzie, A. (2004). Data-driven strategies for an automated dialogue system. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 71–78, Barcelona, Spain.

He, K., Xu, W., & Yan, Y. (2020). Multi-level cross-lingual transfer learning with language shared and specific knowledge for spoken language understanding. *IEEE Access*, *8*, 29407–29416.

He, X., Deng, L., Hakkani-Tur, D., & Tur, G. (2013). Multi-style adaptive training for robust cross-lingual spoken language understanding. In *Proceedings of ICASSP 2013*, pp. 8342–8346. IEEE.

Hedderich, M. A., Adelani, D. I., Zhu, D., Alabi, J. O., Markus, U., & Klakow, D. (2020). Transfer learning and distant supervision for multilingual Transformer models: A study on african languages. In *Proceedings of EMNLP 2020*, pp. 2580–2591.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for Natural Language Processing in low-resource scenarios. In *Proceedings of NAACL-HLT 2021*.

Henderson, M., Casanueva, I., Mrkšić, N., Su, P.-H., Wen, T.-H., & Vulić, I. (2020). ConveRT: Efficient and accurate conversational representations from transformers. In *Proceedings of EMNLP 2020*, pp. 2161–2174.

Henderson, M., Thomson, B., & Williams, J. D. (2014a). The second dialog state tracking challenge. In *Proceedings of SIGDIAL 2014*, pp. 263–272.

Henderson, M., Thomson, B., & Young, S. (2014b). Word-based dialog state tracking with recurrent neural networks. In *Proceedings of SIGDIAL 2014*, pp. 292–299.

Henderson, M., & Vulić, I. (2021). ConVEx: Data-efficient and few-shot slot labeling. In *Proceedings of NAACL-HLT 2021*.

Henderson, M., Vulić, I., Casanueva, I., Budzianowski, P., Gerz, D., Coope, S., Spithourakis, G., Wen, T.-H., Mrkšić, N., & Su, P.-H. (2019a). Polyresponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking. In *Proceedings of EMNLP 2019*, pp. 181–186.

Henderson, M., Vulić, I., Gerz, D., Casanueva, I., Budzianowski, P., Coope, S., Spithourakis, G., Wen, T.-H., Mrkšić, N., & Su, P.-H. (2019b). Training neural response selection

for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5392–5404, Florence, Italy. Association for Computational Linguistics.

Heracleous, P., Mohammad, Y., & Yoneyama, A. (2020). Integrating language and emotion features for multilingual speech emotion recognition. In *International Conference on Human-Computer Interaction*, pp. 187–196.

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Piqueras, L. C., Chalkidis, I., Cui, R., et al. (2022). Challenges and strategies in cross-cultural nlp..

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. In *Proceddings of the Deep Learning Workshop at NeurIPS*.

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.-L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., & Kim, S. (2019). Overview of the sixth dialog system technology challenge: Dstc6. *Computer Speech & Language*, *55*, 1–25.

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of ICNLG 2020*, pp. 169–182.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML 2020*, pp. 4411–4421.

Hung, C.-C., Lauscher, A., Vulić, I., Ponzetto, S. P., & Glavaš, G. (2022). Multi2woz: A robust multilingual dataset and conversational pretraining for task-oriented dialog. In *Proceedings of NAACL 2022*.

Jain, A., Paranjape, B., & Lipton, Z. C. (2019). Entity projection via machine translation for cross-lingual NER. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 1083–1092.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL 2020*, pp. 6282–6293.

Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., & Grave, É. (2018). Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of EMNLP 2018*, pp. 2979–2984.

Kendall, D. E., Linden, R., & Murray, J. L. (2005). *Sociology in our times*. Thomson Wadsworth Belmont, CA.

Khanpour, H., Guntakandla, N., & Nielsen, R. (2016). Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016*, pp. 2012–2021.

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., & Choudhury, M. (2020). GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of ACL 2020*, pp. 3575–3585.

Khashabi, D., Stanovsky, G., Bragg, J., Lourie, N., Kasai, J., Choi, Y., Smith, N. A., & Weld, D. S. (2021). Genie: A leaderboard for human-in-the-loop evaluation of text generation..

Kiela, D., Vulić, I., & Clark, S. (2015). Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of EMNLP 2015*, pp. 148–158.

Kim, S., & Banchs, R. E. (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*, pp. 1–6.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., & Yoshino, K. (2016). The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 511–517. IEEE.

Kim, S., Yang, S., Kim, G., & Lee, S.-W. (2020). Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of ACL 2020*, pp. 567–582.

Kim, Y.-B., Kim, D., Kumar, A., & Sarikaya, R. (2018). Efficient large-scale neural domain classification with personalized attention. In *Proceedings of ACL 2018*, pp. 2214–2224.

Kim, Y., Graça, M., & Ney, H. (2020). When and why is unsupervised neural machine translation useless?. In *Proceedings of EAMT 2020*, pp. 35–44.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of ICLR (Poster)*.

Klemen, M., Krsnik, L., & Robnik-Šikonja, M. (2020). Enhancing deep neural networks with morphological information..

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of NAACL-HLT 2018*, pp. 452–457.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *PNAS*, *117*(14), 7684–7689.

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. In *Proceedings of ACL-HLT 2011*, pp. 1318–1326.

Krishnan, J., Anastasopoulos, A., Purohit, H., & Rangwala, H. (2021). Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling..

Kulshreshtha, S., Garcia, J. L. R., & Chang, C. Y. (2020). Cross-lingual alignment methods for multilingual bert: A comparative study. In *Proceedings of EMNLP 2020*, pp. 933–942.

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G., & Jyothi, P. (2019). Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4863–4872, Florence, Italy. Association for Computational Linguistics.

Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R., & Bhattacharyya, P. (2014). Shata-anuvadak: Tackling multiway translation of Indian languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 1781–1787, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Kurata, G., Xiang, B., Zhou, B., & Yu, M. (2016). Leveraging sentence-level information with encoder lstm for semantic slot filling. In *Proceedings of EMNLP 2016*, pp. 2077–2083.

Langkilde, I., & Knight, K. (1998). The practical value of n-grams is in generation. In *Natural Language Generation*.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y., et al. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., et al. (2019). An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 1311–1316.

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of EMNLP 2020*, pp. 4483–4499.

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

Lee, H., Jo, S., Kim, H., Jung, S., & Kim, T.-Y. (2021). Sumbt+ larl: Effective multi-domain end-to-end neural task-oriented dialog system.. Vol. 9, pp. 116133–116146.

Lee, Y. (2021). Improving end-to-end task-oriented dialog system with a simple auxiliary task. In *Findings of EMNLP 2021*, pp. 1296–1303.

Lembersky, G., Ordan, N., & Wintner, S. (2012). Language models for machine translation: Original vs. translated texts. *Computational Linguistics*, *38*(4), 799–825.

Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., & Mehdad, Y. (2021). MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of EACL 2021*, pp. 2950–2962.

Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, B. (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*, pp. 110–119, San Diego, California. Association for Computational Linguistics.

Li, J., Monroe, W., Ritter, A., Jurafsky, D., Galley, M., & Gao, J. (2016b). Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Austin, Texas. Association for Computational Linguistics.

Li, X., Lan, W., Dong, J., & Liu, H. (2016c). Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 271–275.

Li, X., Wang, Y., Sun, S., Panda, S., Liu, J., & Gao, J. (2018). Microsoft dialogue challenge: Building end-to-end task-completion dialogue systems. arXiv preprint arXiv:1807.11125.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., & Zhou, M. (2020). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of EMNLP 2020*, pp. 6008–6018.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., et al. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL 2019*, pp. 3125–3135.

Lin, Z., Pan, X., Wang, M., Qiu, X., Feng, J., Zhou, H., & Li, L. (2020a). Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of EMNLP 2020*, pp. 2649–2663.

Lin, Z., Liu, Z., Winata, G. I., Cahyawijaya, S., Madotto, A., Bang, Y., Ishii, E., & Fung, P. (2020b). XPersona: Evaluating multilingual personalized chatbot. arXiv preprint arXiv:2003.07568.

Lin, Z., Madotto, A., Winata, G. I., Xu, P., Jiang, F., Hu, Y., Shi, C., & Fung, P. (2021). Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Lison, P., Barnes, J., Hubin, A., & Touileb, S. (2020). Named entity recognition without labelled data: A weak supervision approach. In *Proceedings of ACL 2020*, pp. 1518–1533.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL 2017*, Vol. 2, pp. 8–14.

Liu, B., Tür, G., Hakkani-Tür, D., Shah, P., & Heck, L. (2018). Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of NAACL-HLT 2018*, pp. 2060–2069.

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP 2016*, pp. 2122–2132.

Liu, X., Eshghi, A., Swietojanski, P., & Rieser, V. (2019). Benchmarking natural language understanding services for building conversational agents. arXiv preprint arXiv:1903.05566.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *TACL*, *8*, 726–742.

Liu, Z., Indra Winata, G., Cahyawijaya, S., Madotto, A., Lin, Z., & Fung, P. (2021). On the importance of word order information in cross-lingual sequence labeling. In *Proceedings of of AAAI 2021*.

Liu, Z., Shin, J., Xu, Y., Winata, G. I., Xu, P., Madotto, A., & Fung, P. (2019). Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 1297–1303.

Liu, Z., Winata, G. I., Lin, Z., Xu, P., & Fung, P. (2020). Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of AAAI 2020*, pp. 8433–8440.

Louvan, S., & Magnini, B. (2020a). Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of COLING 2020*, pp. 480–496.

Louvan, S., & Magnini, B. (2020b). Simple data augmentation for multilingual nlu in task oriented dialogue systems.. p. 250.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (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)*, pp. 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Lowe, R., Pow, N., Serban, I. V., Charlin, L., Liu, C.-W., & Pineau, J. (2017b). Training end-to-end dialogue systems with the Ubuntu dialogue corpus. *Dialogue & Discourse*, *8*(1), 31–65.

Luo, B., Feng, Y., Wang, Z., Zhu, Z., Huang, S., Yan, R., & Zhao, D. (2017). Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of ACL 2017*, pp. 430–439.

Madotto, A., Wu, C.-S., & Fung, P. (2018). Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of ACL 2018*, pp. 1468–1478.

Majewska, O., Razumovskaia, E., Ponti, E. M., Vulić, I., & Korhonen, A. (2022). Cross-lingual dialogue dataset creation via outline-based generation. In *CORR:2201.13405*.

Mayhew, S., Chaturvedi, S., Tsai, C.-T., & Roth, D. (2019). Named entity recognition with partially annotated training data. In *Proceedings of CoNLL 2019*, pp. 645–655.

Mazzei, A., Battaglino, C., & Bosco, C. (2016). Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th International Natural Language Generation conference*, pp. 184–192.

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

Mehri, S., Eric, M., & Hakkani-Tur, D. (2020). DialoGLUE: A natural language understanding benchmark for task-oriented dialogue. arXiv preprint arXiv:2009.13570.

Meng, C., Ren, P., Chen, Z., Sun, W., Ren, Z., Tu, Z., & Rijke, M. d. (2020). Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. In *Proceedings of SIGIR 2020*, pp. 1151–1160.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., 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.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Mirkovic, D., & Cavedon, L. (2011). Dialogue management using scripts. Google Patents. US Patent 8,041,570.

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of EMNLP 2020: System Demonstrations*, pp. 119–126.

Mrkšić, N., Séaghdha, D. Ó., Wen, T.-H., Thomson, B., & Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL 2017*, Vol. 1, pp. 1777–1788.

Mrkšić, N., & Vulić, I. (2018). Fully statistical neural belief tracking. In *Proceedings of ACL 2018*, pp. 108–113.

Mrkšić, N., Vulić, I., Séaghdha, D. Ó., Leviant, I., Reichart, R., Gašić, M., Korhonen, A., & Young, S. (2017). Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *TACL*, *5*, 309–324.

Muise, C., Chakraborti, T., Agarwal, S., Bajgar, O., Chaudhary, A., Lastras-Montano, L. A., Ondrej, J., Vodolan, M., & Wiecha, C. (2019). Planning for goal-oriented dialogue systems. arXiv preprint arXiv:1910.08137.

Namazifar, M., Papangelis, A., Tür, G., & Hakkani-Tür, D. (2020). Language model is all you need: Natural language understanding as question answering. *CoRR*, *abs/2011.03023*.

Nekvinda, T., & Dusek, O. (2020). One model, many languages: Meta-learning for multilingual text-to-speech. In *Proceedings of INTERSPEECH 2020*, pp. 2972–2976.

Nema, P., Shetty, S., Jain, P., Laha, A., Sankaranarayanan, K., & Khapra, M. M. (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)*, pp. 1539–1550, New Orleans, Louisiana. Association for Computational Linguistics.

Ni, J., Dinu, G., & Florian, R. (2017). Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In *Proceedings of ACL 2017*, pp. 1470–1480.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., & Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In *Proceedings of EMNLP 2020*, pp. 4547–4562.

Ohmura, J., & Eskénazi, M. (2018). Context-aware dialog re-ranking for task-oriented dialog systems. In *Proceedings of SLT 2018*, pp. 846–853.

Onoe, Y., & Durrett, G. (2019). Learning to denoise distantly-labeled data for entity typing. In *Proceedings of NAACL-HLT 2019*, pp. 2407–2417.

Pan, Y., Li, X., Yang, Y., & Dong, R. (2020). Multi-source neural model for machine translation of agglutinative language. *Future Internet*, *12*(6), 96.

Pandey, G., Contractor, D., Kumar, V., & Joshi, S. (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)*, pp. 1329–1338, Melbourne, Australia. Association for Computational Linguistics.

Passban, P., Liu, Q., & Way, A. (2018). Improving character-based decoding using target-side morphological information for neural machine translation. In *Proceedings of NAACL-HLT 2018*, pp. 58–68.

Patra, B., Moniz, J. R. A., Garg, S., Gormley, M. R., & Neubig, G. (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*, pp. 184–193, Florence, Italy. Association for Computational Linguistics.

Peng, B., Zhu, C., Li, C., Li, X., Li, J., Zeng, M., & Gao, J. (2020). Few-shot natural language generation for task-oriented dialog. In *Findings of EMNLP 2020*, pp. 172–182.

Perez, J., & Liu, F. (2017). Dialog state tracking, a machine reading approach using memory network. In *Proceedings of EACL 2017*, Vol. 1, pp. 305–314.

Peskov, D., Clarke, N., Krone, J., Fodor, B., Zhang, Y., Youssef, A., & Diab, M. (2019). Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 4518–4528.

Pfeiffer, J., Goyal, N., Lin, X. V., Li, X., Cross, J., Riedel, S., & Artetxe, M. (2022). Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of NAACL 2022*.

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2020). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of EMNLP 2020*, pp. 7654–7673.

Pfeiffer, J., Vulić, I., Gurevych, I., & Ruder, S. (2021). UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of EMNLP 2021*, pp. 10186–10203.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT?. In *Proceedings of ACL 2019*, pp. 4996–5001.

Pittermann, J., Pittermann, A., & Minker, W. (2010). Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, *13*(1), 49–60.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online. Association for Computational Linguistics.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, *45*(3), 559–601.

Ponti, E. M., Reichart, R., Korhonen, A., & Vulić, I. (2018a). Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of ACL 2018*, pp. 1531–1542.

Ponti, E. M., Vulić, I., Glavaš, G., Mrkšić, N., & Korhonen, A. (2018b). Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of EMNLP 2018*, pp. 282–293.

Ponti, E. M., Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). Cross-lingual semantic specialization via lexical relation induction. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 2206–2217.

Pratap, V., Sriram, A., Tomasello, P., Hannun, A., Liptchinsky, V., Synnaeve, G., & Collobert, R. (2020a). Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters. In *Proceedings of INTERSPEECH 2020*, pp. 4751–4755.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020b). MLS: A large-scale multilingual dataset for speech research. In *Proceedings of INTERSPEECH 2020*, pp. 2757–2761.

Price, P. (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*.

Qin, L., Ni, M., Zhang, Y., & Che, W. (2021). Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of IJICAI 2021*, pp. 3853–3860.

Qin, L., Xu, X., Che, W., Zhang, Y., & Liu, T. (2020). Dynamic fusion network for multi-domain end-to-end task-oriented dialog. In *Proceedings of the ACL 2020*, pp. 6344–6354.

Quan, J., Zhang, S., Cao, Q., Li, Z., & Xiong, D. (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*, pp. 930–940, Online. Association for Computational Linguistics.

Raiman, J., & Miller, J. (2017). Globally normalized reader. In *Proceedings of EMNLP 2017*, pp. 1059–1069.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pp. 2383–2392.

Ramos-Soto, A., Janeiro-Gallardo, J., & Diz, A. B. (2017). Adapting simplenlg to spanish. In *Proceedings of the 10th international conference on natural language generation*, pp. 144–148.

Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of COLING 2020*, pp. 6838–6855.

Raux, A., Langner, B., Bohus, D., Black, A. W., & Eskenazi, M. (2005). Let's go public! taking a spoken dialog system to the real world. In *Proceedings of INTERSPEECH 2005*.

Ravuri, S., & Stolcke, A. (2015). Recurrent neural network and lstm models for lexical utterance classification. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Razumovskaia, E., Vulić, I., & Korhonen, A. (2022). Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue. In *Findings of ACL 2022*, pp. 2017–2033.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of EMNLP 2020*, pp. 4512–4525.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87.

Reiter, E., & Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press.

Ren, F., & Xue, S. (2020). Intention detection based on siamese neural network with triplet loss. *IEEE Access*, *8*, 82242–82254.

Ren, L., Xie, K., Chen, L., & Yu, K. (2018). Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2780–2786, Brussels, Belgium. Association for Computational Linguistics.

Rönnqvist, S., Kanerva, J., Salakoski, T., & Ginter, F. (2019). Is multilingual BERT fluent in language generation?. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pp. 29–36.

Rotman, G., Vulić, I., & Reichart, R. (2018). Bridging languages through images with Deep Partial Canonical Correlation Analysis. In *Proceedings of ACL 2018*, pp. 910–921.

Rücklé, A., Pfeiffer, J., & Gurevych, I. (2020). MultiCQA: Zero-shot transfer of self-supervised text matching models on a massive scale. In *Proceedings of EMNLP 2020*, pp. 2471–2486.

Ruder, S., Constant, N., Botha, J., Siddhant, A., Firat, O., Fu, J., Liu, P., Hu, J., Garrette, D., Neubig, G., & Johnson, M. (2021). XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP 2021*, pp. 10215–10245.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631.

Ruder, S., Vulić, I., & Søgaard, A. (2022). Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Proceedings of ACL 2022*, pp. 2340–2354.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2021). How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of ACL 2021*, pp. 3118–3135.

Şahin, G. G., & Steedman, M. (2018). Data augmentation via dependency tree morphing for low-resource languages. In *Proceedings of EMNLP 2018*, pp. 5004–5009.

Salesky, E., Abdullah, B. M., Mielke, S., Klyachko, E., Serikov, O., Ponti, E. M., Kumar, R., Cotterell, R., & Vylomova, E. (2021). SIGTYP 2021 shared task: Robust spoken language identification. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pp. 122–129.

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

Sap, M., Shwartz, V., Bosselut, A., Choi, Y., & Roth, D. (2020). Commonsense reasoning for natural language processing. In *Proceedings of ACL 2020: Tutorial Abstracts*, pp. 27–33.

Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2019a). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT 2019*, Vol. 1, pp. 3795–3805.

Schuster, T., Ram, O., Barzilay, R., & Globerson, A. (2019b). Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of NAACL-HLT 2019*, pp. 1599–1613.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016a). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI 2016*.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., & Bengio, Y. (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)*, pp. 588–598, Berlin, Germany. Association for Computational Linguistics.

Shan, Y., Li, Z., Zhang, J., Meng, F., Feng, Y., Niu, C., & Zhou, J. (2020). A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of ACL 2020*, pp. 6322–6333.

Shekhar, R., Venkatesh, A., Baumgärtner, T., Bruni, E., Plank, B., Bernardi, R., & Fernández, R. (2019). Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat. In *Proceedings of NAACL-HLT 2019*, pp. 2578–2587.

Shen, S.-q., Chen, Y., Yang, C., Liu, Z.-y., Sun, M.-s., et al. (2018). Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(12), 2319–2327.

Siddhant, A., Bapna, A., Cao, Y., Firat, O., Chen, M. X., Kudugunta, S. R., Arivazhagan, N., & Wu, Y. (2020a). Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of ACL 2020*, pp. 2827–2835.

Siddhant, A., Johnson, M., Tsai, H., Ari, N., Riesa, J., Bapna, A., Firat, O., & Raman, K. (2020b). Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of AAAI 2020*, pp. 8854–8861.

Sigurdsson, G. A., Alayrac, J., Nematzadeh, A., Smaira, L., Malinowski, M., Carreira, J., Blunsom, P., & Zisserman, A. (2020). Visual grounding in video for unsupervised word translation. In *Proceedings of CVPR 2020*, pp. 10847–10856.

Smith, E. M., Hsu, O., Qian, R., Roller, S., Boureau, Y., & Weston, J. (2022). Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *CoRR*, *abs/2201.04723*.

Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR 2017*.

Solaiman, I., & Dennison, C. (2021). Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems, 34*.

Song, Y., Li, C.-T., Nie, J.-Y., Zhang, M., Zhao, D., & Yan, R. (2018). An ensemble of retrieval-based and generation-based human-computer conversation systems. In *Proceedings of IJCAI 2018*, pp. 4382–4388.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., & Dolan, B. (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*, pp. 196–205, Denver, Colorado. Association for Computational Linguistics.

Stent, A., Prasad, R., & Walker, M. (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)*, pp. 79–86, Barcelona, Spain.

Su, P.-H., Mrkšić, N., Casanueva, I., & Vulić, I. (2018). Deep learning for conversational AI. In *Proceedings of NAACL 2018: Tutorial Abstracts*, pp. 27–32.

Sun, Y., Tang, D., Xu, J., Duan, N., Feng, X., Qin, B., Liu, T., & Zhou, M. (2019). Knowledge-aware conversational semantic parsing over web tables. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 827–839. Springer.

Susanto, R. H., & Lu, W. (2017). Neural architectures for multilingual semantic parsing. In *Proceedings of ACL 2017*, Vol. 2, pp. 38–44.

Upadhyay, S., Faruqui, M., Tür, G., Dilek, H.-T., & Heck, L. (2018). (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6034–6038. IEEE.

Üstün, A., Bisazza, A., Bouma, G., & van Noord, G. (2020). UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of EMNLP 2020*, pp. 2302–2315.

Van, P. N., Hoang, T. C., Manh, D. N., Minh, Q. N., & Quoc, L. T. (2022). Viwoz: A multi-domain task-oriented dialogue systems dataset for low-resource language..

van der Goot, R., Sharaf, I., Imankulova, A., Üstün, A., Stepanović, M., Ramponi, A., Khairunnisa, S. O., Komachi, M., & Plank, B. (2021). From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of NAACL-HLT 2021*, pp. 2479–2497, Online. Association for Computational Linguistics.

van der Heijden, N., Yannakoudakis, H., Mishra, P., & Shutova, E. (2021). Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of EACL 2021*.

van Miltenburg, E., Elliott, D., & Vossen, P. (2017). Cross-linguistic differences and similarities in image descriptions..

van Miltenburg, E., Kádár, Á., Koolen, R., & Krahmer, E. (2018). DIDEC: The Dutch image description and eye-tracking corpus. In *Proceedings of COLING 2018*, pp. 3658–3669.

Vania, C., Kementchedjhieva, Y., Søgaard, A., & Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 1105–1116.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of NeurIPS 2017*, pp. 6000–6010.

Vaudry, P.-L., & Lapalme, G. (2013). Adapting simplenlg for bilingual english-french realisation. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pp. 183–187.

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118.

Vulić, I., Glavaš, G., Mrkšić, N., & Korhonen, A. (2018). Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In *Proceedings of NAACL-HLT 2018*, pp. 516–527.

Vulić, I., Mrkšić, N., Reichart, R., Ó Séaghdha, D., Young, S., & Korhonen, A. (2017). Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL 2017*, pp. 56–68.

Wan, X., Li, H., & Xiao, J. (2010). Cross-language document summarization based on machine translation quality prediction. In *Proceedings of ACL 2010*, pp. 917–926.

Wang, A., & Cho, K. (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*, pp. 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 6382–6388.

Wen, T.-H., Su, P.-H., Budzianowski, P., Casanueva, I., & Vulić, I. (2019). Data collection and end-to-end learning for conversational AI. In *Proceedings of EMNLP-IJCNLP 2019: Tutorial Abstracts*.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gasic, M., Barahona, L. M. R., Su, P.-H., Ultes, S., & Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL 2017*, Vol. 1, pp. 438–449.

Weston, J., Dinan, E., & Miller, A. (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*, pp. 87–92, Brussels, Belgium. Association for Computational Linguistics.

Williams, A., Nangia, N., & Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT 2018*, Vol. 1, pp. 1112–1122.

Williams, J., Raux, A., Ramachandran, D., & Black, A. (2013). The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pp. 404–413, Metz, France. Association for Computational Linguistics.

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. arXiv preprint arXiv:1901.08149.

Wu, C.-S., Hoi, S. C., Socher, R., & Xiong, C. (2020a). TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of EMNLP 2020*, pp. 917–929.

Wu, Q., Lin, Z., Karlsson, B., Jian-Guang, L., & Huang, B. (2020b). Single-/multi-source cross-lingual ner via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6505–6514.

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 833–844.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130.

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (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)*, pp. 496–505, Vancouver, Canada. Association for Computational Linguistics.

Xie, Q., Dai, Z., Hovy, E. H., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In *Proceedings of NeurIPS 2020*.

Xu, P., & Sarikaya, R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 ieee workshop on automatic speech recognition and understanding*, pp. 78–83.

Xu, W., Haider, B., & Mansour, S. (2020). End-to-end slot alignment and recognition for cross-lingual nlu. In *Proceedings of EMNLP 2020*, pp. 5052–5063.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT 2021*, pp. 483–498.

Yang, L., Hu, J., Qiu, M., Qu, C., Gao, J., Croft, W. B., Liu, X., Shen, Y., & Liu, J. (2019). A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1341–1350.

Yang, L., Qiu, M., Qu, C., Guo, J., Zhang, Y., Croft, W. B., Huang, J., & Chen, H. (2018). Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of SIGIR 2018*, pp. 245–254.

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

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, *101*(5), 1160–1179.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, *33*, 17283–17297.

Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., & Chen, J. (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*, pp. 109–117, Online. Association for Computational Linguistics.

Zhang, H., & Lin, Y. (2020). Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages. In *Proceedings of INTERSPEECH 2020*, pp. 3161–3165.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. In *Proceedings of ACL 2018*, pp. 2204–2213.

Zhang, W.-N., Chen, Z., Che, W., Hu, G., & Liu, T. (2017). The first evaluation of chinese human-computer dialogue technology. arXiv preprint arXiv:1709.10217.

Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., & Dolan, B. (2018). Generating informative and diverse conversational responses via adversarial information maximization. In *Proceedings of NeurIPS 2018*, pp. 1815–1825.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, W. B. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of ACL 2020*, pp. 270–278.

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., Jia, Y., Rosenberg, A., & Ramabhadran, B. (2019a). Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. In *Proceedings of INTERSPEECH 2019*, pp. 2080–2084.

Zhang, Z., Zhang, Z., Chen, H., & Zhang, Z. (2019b). A joint learning framework with bert for spoken language understanding. *IEEE Access*, *7*, 168849–168858.

Zhang, Z., Li, J., Zhu, P., Zhao, H., & Liu, G. (2018). Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhao, M., Zhu, Y., Shareghi, E., Reichart, R., Korhonen, A., & Schütze, H. (2020). A closer look at few-shot crosslingual transfer: Variance, benchmarks and baselines. *CoRR*, *abs/2012.15682*.

Zhong, V., Xiong, C., & Socher, R. (2018). Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of ACL 2018*, Vol. 1, pp. 1458–1467.

Zhou, H., Zheng, C., Huang, K., Huang, M., & Zhu, X. (2020). Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of ACL 2020*, pp. 7098–7108.

Zhu, C., Zeng, M., & Huang, X. (2019). Multi-task learning for natural language generation in task-oriented dialogue. In *Proceedings of EMNLP-IJCNLP 2019*, pp. 1261–1266.

Zhu, Q., Huang, K., Zhang, Z., Zhu, X., & Huang, M. (2020a). Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *TACL*, *8*, 281–295.

Zhu, S., Cheng, X., & Su, S. (2020b). Conversational semantic parsing over tables by decoupling and grouping actions. *Knowledge-Based Systems*, *204*, 106237.

Ziser, Y., & Reichart, R. (2018). Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance. In *Proceedings of EMNLP 2018*, pp. 238–249.

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of EMNLP 2016*, pp. 1568–1575.

Zuo, L., Qian, K., Yang, B., & Yu, Z. (2021). AllWOZ: Towards Multilingual Task-Oriented Dialog Systems for All..