Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

« Est-ce que tu me suis ? » : une revue du suivi de l'état du dialogue

Léo Jacqmin[1,2]
(1) Orange Innovation, Lannion, France
(2) Aix-Marseille Université, LIS UMR 7020, Marseille, France
leo.jacqmin@orange.com

RÉSUMÉ

Tout en communiquant avec un utilisateur, un système de dialogue orienté tâche doit suivre les besoins de l'utilisateur à chaque étape selon l'historique de la conversation. Ce procédé appelé suivi de l'état du dialogue est primordial car il informe directement les actions du système. Cet article présente dans un premier temps la tâche du suivi de l'état du dialogue, les jeux de données disponibles et les approches modernes. Ensuite, compte tenu du nombre important de publications des dernières années, il vise à recenser les point saillants et les avancées des recherches. Bien que les approches neuronales aient permis des progrès notables, nous argumentons que certains aspects critiques liés aux systèmes de dialogue sont encore trop peu explorés. Pour motiver de futures études, plusieurs pistes de recherche sont proposées.

ABSTRACT

"Do you follow me ?" : a review of dialogue state tracking

While communicating with a user, a task-oriented dialogue system has to track the user needs at each step according to the conversation history. This process called dialogue state tracking is crucial because it directly informs the system's actions. This paper first presents dialogue state tracking, available datasets and modern approaches. Then, considering the large number of publications in the last few years, it aims at listing the highlights and advances of research. Although neural approaches have allowed significant progress, we argue that some critical aspects related to dialogue systems are still underexplored. To motivate future studies, several research avenues are proposed.

MOTS-CLÉS : Suivi de l'état du dialogue, systèmes de dialogues orienté tâches.

KEYWORDS: Dialogue state tracking, task-oriented dialogue systems.

1 Introduction

La conversation humaine étant par nature complexe et ambiguë, l'apprentissage d'un agent conversationnel à domaine ouvert capable d'effectuer des tâches arbitraires est un problème ouvert. Par conséquent, la pratique s'est concentrée sur la construction de systèmes de dialogue orientés tâche limités à des domaines spécifiques comme la réservation de vols ou l'achat d'un produit. Ces systèmes sont typiquement implémentés à travers une architecture modulaire qui fournit plus de contrôle et permet l'interaction avec une base de données, traits désirables pour une application commerciale. Cette architecture, illustrée en Figure 1, comprend les composantes suivantes : la compréhension du langage (NLU pour *Natural Language Understanding*), le suivi de l'état du dialogue (DST pour

Dialogue State Tracking), la politique du dialogue et la génération de langage naturel. Le DST cherche à extraire à chaque tour de parole une représentation des besoins de l'utilisateur en prenant en compte l'historique du dialogue. Au-delà du traitement d'un tour de parole isolé, il doit être capable d'accumuler avec précision l'information au cours d'une conversation et d'ajuster sa prédiction de l'état du dialogue en fonction des observations et du contexte. Il permet ainsi d'obtenir un résumé de tout l'historique du dialogue afin d'en suivre le progrès.

Différents articles de revue ont dressé l'état de l'art de la tâche de DST sous l'angle d'un paradigme spécifique ou d'une période donnée.[1] Young *et al.* (2013) ont donné une vue d'ensemble des systèmes de dialogue statistiques à base de POMDP.[2] Ces approches génératives modélisent le dialogue sous la forme d'un réseau bayésian dynamique pour suivre un état de croyance. Elles permettaient de prendre en compte l'incertitude de l'entrée liée à la transcription de la parole et offraient une alternative aux systèmes déterministes conventionnels qui étaient coûteux à mettre en place et souvent fragiles en fonctionnement. Toutefois, les systèmes à base de POMDP présupposaient certaines indépendances pour être calculables et ont peu à peu été délaissés en faveur des méthodes discriminatives d'apprentissage automatique qui modélisent directement la distribution de l'état du dialogue. Ces méthodes, recensées par Henderson (2015), tirent profit des corpus de dialogue annotés comme les corpus du *Dialogue State Tracking Challenge* (DSTC), premier cadre d'évaluation standardisé pour cette tâche (Williams *et al.*, 2016). Les dernière années dans le domaine du DST ont été marquées par l'utilisation de modèles neuronaux qui ont permis des avancées notables que couvrent Balaraman *et al.* (2021). Depuis, de nombreux travaux touchant à des problématiques clés comme la capacité d'adaptation ont été publiés. Le présent article propose donc une synthèse des avancées récentes afin d'identifier les réalisations majeures du domaine.

Le développement récent des modèles d'apprentissage profond a permis de répondre à des problématiques fondamentales des systèmes de dialogue. Les modèles de DST peuvent désormais être dissociés d'un domaine donné et partagés entre des domaines similaires (Wu *et al.*, 2019). Ils obtiennent même des résultats prometteurs lorsqu'aucun dialogue annoté n'est disponible (Lin *et al.*, 2021b). Malgré ces progrès, les approches modernes restent limitées à un scénario de dialogue spécifique. En effet, dans la plupart des jeux de données dédiés au DST, les dialogues consistent à remplir un formulaire : le système demande des contraintes jusqu'à ce qu'il puisse interroger la base de données et renvoyer les résultats à l'utilisateur. De plus, ces approches à partir des données sont rigides et ne correspondent pas au besoin de contrôler finement les agents conversationnels dans un contexte applicatif. Dans un scénario réel, l'accès à des données annotées est limité et les modèles de DST récents semblent avoir une capacité de généralisation faible (Li *et al.*, 2021b). Ces limitations montrent qu'il reste des défis majeurs au développement d'agents conversationnels polyvalents qui puissent être adoptés par le public. Une deuxième contribution de cet article est de proposer plusieurs directions de recherche potentielles pour aborder ces défis.

2 Suivi de l'état du dialogue

Cette section fournit quelques éléments de contexte pertinents pour le DST, délimite la portée de l'étude dans le contexte des systèmes de dialogue, et définit la tâche du DST. Elle donne en dernier lieu un aperçu des jeux de données et des métriques d'évaluation utilisés.

1. Pour le lecteur intéressé, d'autres articles de revue adoptent un angle plus large et donnent une vue d'ensemble sur les systèmes de dialogue de manière générale (Chen *et al.*, 2017; Gao *et al.*, 2018; Zhang *et al.*, 2020b).
2. *Partially Observable Markov Decision Process* ou processus de décision markovien partiellement observable.

2.1 Contexte

Le langage se présente comme une interface naturelle pour interagir avec une machine. Ainsi, un système de dialogue abouti permet d'automatiser l'exécution d'une tâche donnée en interaction avec un utilisateur. Ces dernières années ont vu un accroissement des travaux de recherche dans ce domaine qui va de pair avec l'intérêt grandissant des entreprises à implémenter des solutions à base de systèmes de dialogue orientés tâche pour réduire leurs coûts de support client.

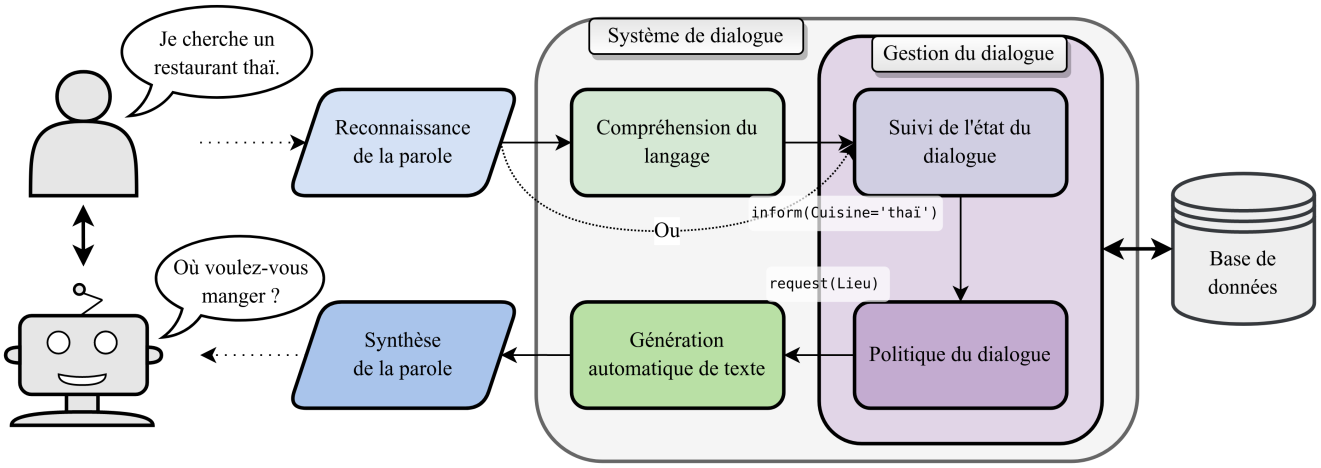

FIGURE 1 – Architecture modulaire pour un système de dialogue parlé orienté tâche.

La Figure 1 illustre l'architecture modulaire communément adoptée pour ces systèmes. Cette architecture permet d'intégrer différentes composantes et de comprendre leur effet sur le système dans son ensemble. Le NLU se compose de deux sous-tâches principales, à savoir la détection d'intention qui consiste à identifier l'intention de l'utilisateur, comme la réservation d'un hôtel, et l'étiquetage de slots qui consiste à identifier les concepts sémantiques pertinents dans un énoncé de l'utilisateur, comme le prix et le lieux. Le DST vise à mettre à jour l'état du dialogue, une représentation des besoins de l'utilisateur exprimés au fil de la conversation. La politique du dialogue cherche à apprendre l'action du système en fonction de l'état actuel. Il forme avec le DST le module de gestion du dialogue qui est en interface avec l'ontologie, une représentation structurée de la base de données en back-end qui contient les informations nécessaires à la réalisation de la tâche. En dernier lieu, la génération de langage naturel transforme l'action du système en langage naturel.

Dans le cas d'un système de dialogue parlé, des composantes de reconnaissance automatique de la parole (ASR pour *Automatic Speech Recognition*) et de synthèse de la parole sont intégrées pour passer de la parole au texte et inversement. Le NLU opère alors sur les hypothèses de l'ASR et est dénoté SLU (*Spoken Language Understanding*). [3] Traditionnellement, le DST opère sur la sortie du SLU pour mettre à jour l'état du dialogue en traitant le bruit provenant de l'ASR et du SLU. Par exemple, le SLU peut fournir une liste d'hypothèses des représentations sémantiques basée sur la liste d'hypothèses de phrases de l'ASR. Le DST gère toutes ces incertitudes pour mettre à jour l'état du dialogue. Cependant, les jeux de données récents sont collectés sous la forme de texte sans prendre en compte les entrées vocales bruitées, ce qui a fait que la tâche d'étiquetage de slots du SLU et la tâche de DST sont étudiées séparément. Cet article recense essentiellement les progrès autour du DST.

3. À noter qu'il existe également des approches *end-to-end* pour obtenir une représentation sémantique à partir de la parole.

2.2 État du dialogue

L'état du dialogue e_t constitue un résumé des besoins de l'utilisateur à un tour de parole donné destiné à guider le système de dialogue dans l'exécution d'une ou plusieurs tâches. Plus précisément, e_t est extrait à partir de l'historique du dialogue jusqu'au tour t de telle sorte à ce que e_t contienne des informations suffisantes pour que le système choisisse l'action suivante. Ce résumé est généralement représenté sous la forme de paires de (*slot, valeur*). [4] Les slots sont des concepts sémantiques auxquels une valeur doit être attribuée afin d'accomplir la tâche. Traditionnellement, les slots informables (contraintes fournie par l'utilisateur, par ex. `fourchette_prix`) sont distingués des slots demandables (informations que l'utilisateur peut demander, par ex. `numéro_téléphone`). À noter qu'un slot peut être à la fois informable et demandable. L'ensemble des slots possibles est prédéfini dans une ontologie spécifique à un domaine, et les valeurs des slots informables sont spécifiées par l'utilisateur au cours du dialogue. Par exemple, pour la réservation d'une chambre d'hôtel, l'état du dialogue au tour t pourrait être $e_t = \{(\texttt{date}, \texttt{15 janvier}), (\texttt{nombre\_nuitées}, 2)\}$. Les slots informables peuvent également prendre une valeur spéciale : soit *dontcare* lorsque que l'utilisateur n'a pas de préférence, soit *none*, lorsque l'utilisateur n'a pas encore spécifié d'objectif pour le slot. La Figure 2 montre un exemple du DST à chaque tour de parole.

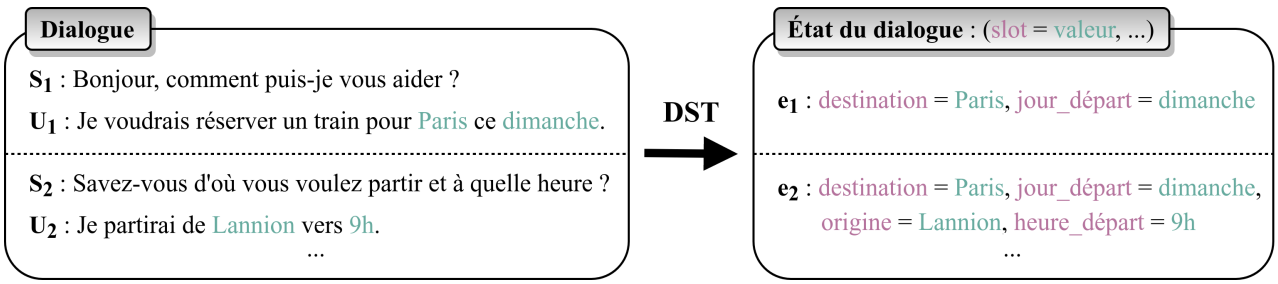

FIGURE 2 – Tâche du suivi de l'état du dialogue (DST pour *Dialogue State Tracking*). **S** et **U** représentent le système et l'utilisateur. Le DST cherche à mettre à jour l'état du dialogue sous la forme de paires de (slot, valeur) à chaque tour de parole. Les valeurs correspondent aux besoins exprimés dans l'énoncé.

La tâche du DST consiste donc à estimer l'état du dialogue actuel e_t sous la forme d'un ensemble de paires de (*slot, valeur*) pour chaque tour t. Cette estimation peut être faite soit à partir du tour de parole précédent ou à partir de l'historique du dialogue complet. Dans le premier cas, le système prend en compte l'état du dialogue précédent e_{t-1} ainsi que les énoncés du tour précédent $t-1$ afin de mettre à jour e_{t-1}. La mise à jour de e_{t-1} peut être faite selon des règles prédéfinies ou via une fonction qui cherche à approximer cette mise à jour à partir de la prédiction $\hat{e}_t$ et de e_{t-1}. Dans le deuxième cas, un nouvel état du dialogue e_t est prédit à chaque tour à partir de l'historique du dialogue complet. Cette approche permet d'éviter l'accumulation d'erreurs au fil du dialogue mais peut potentiellement prédire un état du dialogue e_t qui ne concorde pas avec e_{t-1}.

La représentation de l'état du dialogue en paires de (*slot, valeur*) est utilisée pour traiter un domaine unique. Lorsqu'il s'agit d'un corpus multidomaine, cette approche est généralement étendue en combinant domaines et slots pour extraire des triplets (*domaine*, *slot*, *valeur*). [5]

4. À noter qu'il s'agit d'une représentation possible et que d'autres ont été envisagées. Par exemple, les systèmes de bout en bout génèrent une réponse directement à partir de l'énoncé de l'utilisateur en utilisant une représentation continue implicite.
5. Par souci de simplicité, dans la suite de cet article, nous utilisons le terme "slot" pour désigner les "domaine-slots".

2.3 Jeux de données

De nombreux jeux de données publics ont été publiés pour faire avancer les approches d'apprentissage automatique pour le DST. L'évolution de ces jeux de données est marquée par un niveau de complexité des dialogues croissant, notamment avec l'arrivée de corpus multidomaines. On distingue deux approches principales pour la collecte de données de dialogues : (i) Les approches Wizard-of-Oz (WOZ ou H2H pour *human-to-human*) (Kelley, 1984) où deux humains jouent respectivement les rôles d'utilisateur et système selon une description de la tâche. Cette approche permet d'obtenir des dialogues naturels et variés, mais l'annotation qui suit peut être source d'erreurs. (ii) Les approches à base de simulation (M2M pour *machine-to-machine*) où deux systèmes jouent les rôles d'utilisateur et de système et interagissent l'un avec l'autre pour générer des squelettes de conversations qui sont ensuite paraphrasés par des humains. L'avantage de cette méthode réside dans le fait que les annotations sont obtenues automatiquement. Cependant, la complexité de la tâche et la diversité du langage sont souvent limitées car la simulation du dialogue est réalisée à l'aide de règles.

La Table 1 recense les jeux de données principaux ainsi que les jeux de données récents pertinents pour les problématiques évoquées en Section 4. En dehors des jeux de données mentionnés, il existe d'autres corpus moins récents recensés par Balaraman *et al.* (2021) que nous ne reprenons pas ici par souci de clarté. Ce qui suit est une description des jeux de données principaux. Les jeux de données récents, à savoir SMCalFlow (Andreas *et al.*, 2020), ABCD (Chen *et al.*, 2021), SIMMC 2.0 (Kottur *et al.*, 2021a), BiToD (Lin *et al.*, 2021d) et DSTC10 (Kim *et al.*, 2021), seront abordés en Section 4.

DSTC2 (Henderson *et al.*, 2014a) Les premières éditions du *Dialogue State Tracking Challenge* (DSTC) ont introduit les premiers jeux de données et métriques d'évaluation partagés pour la tâche du DST et ont ainsi catalysé la recherche dans ce domaine. Le corpus de la deuxième édition (DSTC2) est resté une référence aujourd'hui. Il comprend des dialogues entre des participants rémunérés et divers systèmes de dialogue téléphonique (collecte H2M). L'utilisateur a pour objectif de trouver un restaurant en spécifiant des contraintes comme le type de cuisine et peut demander des informations spécifiques comme le numéro de téléphone. Le corpus WOZ 2.0 (Wen *et al.*, 2017; Mrkšić *et al.*, 2017) se base sur la même ontologie que celle utilisée dans DSTC2.

MultiWOZ (Budzianowski *et al.*, 2018) [6] Actuellement le jeu de données de référence, MultiWOZ fut le premier corpus multidomaine à grande échelle. Il contient des dialogues entre un touriste et un employé d'un centre d'information qui peuvent parfois s'étendre sur plusieurs domaines. [7] Un problème majeur lié à la manière dont les données ont été récoltées est l'inconsistance et les erreurs d'annotations. Quatre autres versions ont été publiées ultérieurement pour essayer de corriger ces erreurs (Eric *et al.*, 2019; Zang *et al.*, 2020; Han *et al.*, 2021; Ye *et al.*, 2021a). Des versions en mandarin, coréen, vietnamien, hindi, français, portugais et thaï ont été obtenues par un procédé de traduction automatique suivi d'une correction manuelle (Gunasekara *et al.*, 2020; Zuo *et al.*, 2021).

SGD (Rastogi *et al.*, 2020b) Le *Schema-Guided Dataset* est le jeu de données actuellement le plus utilisé après MultiWOZ. Il a été créé pour solliciter les recherches sur l'indépendance au domaine grâce à l'utilisation de schémas. Les schémas décrivent les domaines, slots et intentions en langage naturel et peuvent être utilisés par un modèle générique pour traiter différents domaines. Le jeu d'évaluation comprend des schémas non vus pour favoriser la généralisation des modèles. Une extension étudie la robustesse des modèles à différentes formulations des schémas (Lee *et al.*, 2021b).

6. Un tableau des scores à jour est disponible au lien suivant : https://github.com/budzianowski/multiwoz.
7. L'ensemble des dialogues couvre les domaines suivants : attraction, hôpital, police, hôtel, restaurant, taxi et train.

	DSTC2	WOZ2.0	MultiWOZ	SGD	SMCalFlow	ABCD	SIMMC2.0	BiToD	DSTC10
Langue(s)	en	en, de, it	en, 7 lang.*	en, ru, ar, id, sw	en	en	en	en, zh	en
Collecte	H2M	H2H	H2H	M2M	H2H	H2H	M2M	M2M	H2H
Modalité(s)	parole	texte	texte	texte	texte	texte	texte, image	texte	parole
Nb domaines	1	1	7	16	4	30	1	5	3
Nb dialogues	1612	600	8438	16 142	41 517	8034	11 244	5787	107
Nb tours	23 354	4472	115 424	329 964	155 923	177 407	117 236	115 638	2292
Moy. tours / dial.	14,5	7,5	13,7	20,4	8,2	22,1	10,4	19,9	12.6
Moy. tokens / tour	8,5	11,5†	13,8†	9,7†	10,2	9,2	12,8	12,2†	17.8
État du dialogue	slot-val	slot-val	slot-val	slot-val	dataflow	action	slot-val	slot-val	slot-val
Slots	8	4	25	214	-	231	12	68	18
Valeurs	212	99	4510	14 139	-	12 047	64	8206	327

TABLE 1 – Caractéristiques des jeux de données principaux (gauche) et récents (droite) disponibles pour la tâche du suivi de l'état du dialogue. La modalité "parole" prend la forme d'hypothèses issues de l'ASR. * Traductions automatiques du corpus original. † Moyenne pour l'anglais.

2.4 Métriques d'évaluation

Comme il existe une correspondance stricte entre l'historique du dialogue et l'état du dialogue, des métriques d'exactitudes permettent de mesurer les performances des modèles de DST. Introduites par Williams *et al.* (2013), deux métriques sont couramment utilisées pour une évaluation conjointe ou individuelle.

Joint Goal Accuracy Métrique principale qui fait référence à l'ensemble des besoins de l'utilisateur. Elle indique la performance du modèle à prédire correctement l'état du dialogue à un tour donné et équivaut à la proportion de tours de parole où toutes les valeurs de slots prédites (y compris pour les slots non affectés) correspondent exactement aux valeurs de référence.

Slot Accuracy Contrairement au *Joint Goal Accuracy*, cette métrique évalue individuellement à chaque tour la valeur prédite pour chaque slot. Elle est obtenue par la macro-moyenne de l'*accuracy* des slots : $SA = \frac{\sum_i^n acc_i}{n}$, où n représente le nombre de slots. Pour une évaluation plus détaillée, elle peut être décomposée selon le type de slots (par ex. les slots demandés par l'utilisateur).

3 Approches modernes

La complexité croissante des jeux de données pour le DST va de pair avec les progrès récents liés aux modèles d'apprentissage profond. L'utilisation de ces modèles a permis d'aborder des problématiques cruciales du DST comme le traitement de dialogues multidomaines et de domaines non vus. Cette section donne en premier lieu une vue d'ensemble des méthodes courantes pour le DST. Elle recense ensuite plus en détail les avancées récentes des deux dernières années (2020 et 2021). Une sélection de modèles récents est présentée en Table 2.

3.1 Méthodes courantes

Une caractéristique qui permet de catégoriser les modèles de DST est la manière dont ils prédisent les valeurs de slots. La prédiction peut se faire soit à partir d'un ensemble de valeurs prédéfini (ontologie fixe), soit à partir d'un ensemble ouvert de valeurs (vocabulaire ouvert).

Ontologie fixe Dans la continuité des approches discriminatives qui les ont précédées, les approches neuronales traditionnelles se basent sur une ontologie fixe et traitent la tâche du DST comme un problème de classification en classes multiples (Henderson *et al.*, 2014b; Mrkšić *et al.*, 2017). Les prédictions pour un slot donné sont estimées par une distribution de probabilités sur un ensemble de valeurs prédéfini, ce qui permet de restreindre le champ de prédiction à un vocabulaire fermé et ainsi de simplifier considérablement la tâche. En termes de *Joint Goal Accuracy*, les performances de cette approche de classification sont donc relativement élevées (Chen *et al.*, 2020), toutefois son coût est proportionnel à la taille du vocabulaire car toutes les valeurs potentielles doivent être évaluées. En pratique, le nombre de valeurs peut être élevé et une ontologie prédéfinie est rarement disponible. De manière générale, cette approche reste limitée car elle ne permet pas de prendre en compte les valeurs non vues durant l'apprentissage ni les slots dont les valeurs ne sont pas prédéfinies (par ex. `nom_hôtel` ou `heure_départ`). Plus important encore, cette dépendance à une ontologie prédéfinie rend le modèle rigide et incapable de traiter des domaines non vus.

Vocabulaire ouvert Pour pallier ces problèmes, des approches permettant de prédire sur un ensemble de valeurs ouvert ont été proposées. Une première méthode consiste à extraire la valeur d'un slot directement depuis l'énoncé de l'utilisateur, supprimant ainsi la dépendance à une ontologie fixe. Par exemple, Gao *et al.* (2019) reformule le DST comme une tâche de compréhension. Une question du type "Quelle est la valeur du slot x ?" est posée et l'historique du dialogue est considéré comme l'extrait où la réponse doit être trouvée. Un mécanisme d'attention est utilisé pour comparer la question au passage et identifier la position de début et de fin de la valeur dans l'historique du dialogue. Une limite de cette méthode est qu'elle dépend uniquement du contexte du dialogue pour la prédiction de valeurs de slots. Cependant, de nombreuses valeurs peuvent ne pas y apparaître ou avoir des descriptions différentes selon les utilisateurs (par ex. la valeur "onéreux" peut être exprimée sous la forme "haut de gamme"). Une alternative est la génération de valeurs qui fait appel à une architecture encodeur-décodeur pour générer l'état du dialogue. Par exemple, le modèle TRADE (Wu *et al.*, 2019) utilise un mécanisme de copie pour générer une valeur pour chaque slot en se basant sur une représentation de l'historique du dialogue. Il incorpore également un classifieur (*slot gate*) qui prédit *ptr* si la valeur d'un slot donné est exprimée dans l'entrée et doit être générée par le décodeur, *none* si le slot est inactif et ne prend pas de valeur, et *dontcare* si n'importe quelle valeur convient. Les paramètres du modèle sont partagés pour tous les slots, TRADE ne dépend donc pas d'une ontologie et a ouvert la voie à l'application de modèles à des domaines non vus. Une autre approche est de décoder l'état du dialogue complet sous la forme d'une chaîne de caractères à l'aide d'un modèle seq2seq (Feng *et al.*, 2021). L'inconvénient des modèles génératifs est qu'ils ont tendance à produire des valeurs invalides, par exemple par des répétitions ou des omissions de mots.

Méthodes hybrides Un compromis semble exister entre le niveau d'indépendance des valeurs dans un modèle et les performances du DST. Certains travaux ont cherché à combiner les approches à ontologie fixe avec la prédiction à vocabulaire ouvert pour bénéficier des avantages des deux méthodes. Cette approche se base sur la distinction entre slots catégoriques pour lesquels un ensemble de valeurs est prédéfini, et slots non catégoriques pour lesquels l'ensemble de valeurs n'est pas limité. Parmi les modèles hybrides, TripPy (Heck *et al.*, 2020) est probablement le plus abouti. Basé sur un encodeur BERT, il garde en mémoire un ensemble de valeurs potentielles et utilise un mécanisme de copie pour extraire la valeur d'un slot selon trois scénarios : (i) la valeur peut être extraite directement dans l'énoncé de l'utilisateur, (ii) la valeur a été mentionnée par le système et confirmée par l'utilisateur, et (iii) la valeur est identique à celle d'un autre slot. Un classifieur est utilisé pour les slots booléens.

Modèle	Décodeur	Contexte	Supervision additionnelle	Onto.	MWOZ2.1
DSTreader (Gao *et al.*, 2019)	Extractif	Historique complet	-	✗	36,40
TRADE (Wu *et al.*, 2019)	Génératif	Historique complet	-	✗	45,60
DSTQA (Zhou & Small, 2019)	Classifieur	Historique complet	Graphe de connaissance	✓	51,17
TOD-BERT (Wu *et al.*, 2020)	Classifieur	Historique complet	-	✓	48,00
NADST (Le *et al.*, 2020)	Génératif	Historique complet	-	✗	49,04
DS-DST (Zhang *et al.*, 2020a)	Extractif + classif.	Tour précédent	-	Slots cat.	51,21
SOM-DST (Kim *et al.*, 2020)	Génératif	Historique, état précédent	-	✗	52,57
MinTL (Lin *et al.*, 2020)	Génératif	Historique, état précédent	Génération de réponses	✗	53,62
SST (Chen *et al.*, 2020)	Génératif	Tour et état précédents	Graphe de schémas	✓	55,23
TripPy (Heck *et al.*, 2020)	Extractif + classif.	Historique complet	Slots mentionnés	✗	55,30
SimpleTOD (Hosseini-Asl *et al.*, 2020)	Génératif	Historique complet	Génération de réponses	✗	55,76
Seq2Seq-DU (Feng *et al.*, 2021)	Génératif	Historique complet	Schémas	Slots cat.	56,10
SOLOIST (Peng *et al.*, 2021a)	Génératif	Historique complet	Génération de réponses	✗	56,85
TripPy + COCO (Li *et al.*, 2021b)	Extractif + classif.	Historique complet	Augmentation de données	✗	60,53
D3ST (Zhao *et al.*, 2022)	Génératif	Historique complet	Schémas	✗	57,80

TABLE 2 – Caractéristiques des modèles de DST récents et performance en termes de *Joint Goal Accuracy*. "Onto." dénote l'accès à un ensemble de valeurs prédéfini pour chaque slot.

3.2 Avancées récentes

Ces dernières années, plusieurs thématiques importantes ont été abordées notamment grâce à l'utilisation des modèles de langage pré-entraînés (PLM pour *Pre-trained Language Model*) appris sur de grandes quantités de texte non annoté via un objectif auto-supervisé comme la prédiction de tokens masqués (Devlin *et al.*, 2019).

Modélisation des relations entre les slots Les méthodes mentionnées jusqu'à présent traitent les slots individuellement sans prendre en compte leurs relations. Or, les slots ne sont pas conditionnellement indépendants (par ex. les slots d'un même domaine auront tendance à apparaître ensemble). Une alternative est de considérer explicitement les corrélations entre les slots Ye *et al.* (2021b) modélisent les relations entre slots en utilisant un mécanisme d'auto-attention, alors que Chiang & Yeh (2021) utilisent un champ aléatoire de Markov. Ces relations peuvent aussi être représentées comme des arêtes dans des modèles à base de graphes, où les slots (et potentiellement les valeurs) sont représentés par des nœuds dans le graphe. Chen *et al.* (2020) construisent d'abord un graphe de schémas et utilisent ensuite un réseau de graphe à base d'attention (*graph attention network*, GAT) pour fusionner les informations des énoncés du dialogue et du graphe de schéma. Lin *et al.* (2021a) adoptent une architecture hybride pour permettre une prédiction séquentielle des valeurs à partir d'un modèle GPT-2 tout en modélisant les relations entre les slots et les valeurs avec un GAT.

Adaptation des PLM aux données de dialogues Les travaux récents de DST ont considérablement bénéficié des avancées permises par les PLM. Cependant, les PLM existants sont pré-entraînés sur des données textuelles de forme libre en utilisant des objectifs de modélisation du langage. Cela limite leur capacité à modéliser le contexte ou la dynamique multi-tours des dialogues. Des études ont montré qu'il était avantageux d'adapter un PLM au domaine ou à la tâche cible en continuant l'apprentissage auto-supervisé (Gururangan *et al.*, 2020). Cette méthode a été appliquée aux systèmes de dialogue orientés tâche et au DST. Il y a deux questions sous-jacentes derrière cette approche : la sélection des données d'adaptation et l'élaboration de fonctions d'objectif auto-supervisées qui permettent l'apprentissage de meilleures représentations des dialogues pour la tâche en aval. Wu *et al.* (2020) rassemblent neuf corpus de dialogues orientés tâche et continuent l'apprentissage auto-supervisé d'un modèle BERT avec comme fonctions d'objectif la prédiction de mots masqués et la sélection de réponse. Le modèle obtenu nommé TOD-BERT fournit une amélioration importante par rapport

à un modèle BERT standard sur plusieurs tâches de dialogue dont le DST. Avec une configuration similaire, Zhu *et al.* (2021) contrastent ces résultats et constatent que l'adaptation au domaine cible est surtout avantageuse lorsque peu de données annotées sont disponibles. Sur base de TOD-BERT, Hung *et al.* (2021) montrent qu'il est avantageux non seulement d'adapter un PLM aux données de dialogues, mais aussi au domaine cible. Pour ce faire, ils utilisent des données conversationnelles issues de Reddit filtrées pour qu'elles contiennent des termes spécifiques au domaine cible. Enfin, Yu *et al.* (2021) présentent deux fonctions d'objectif destinées à injecter des biais inductifs dans un PLM afin de représenter conjointement les énoncés dynamiques de dialogues et la structure de l'ontologie. Ils évaluent leur méthode sur plusieurs tâches d'analyse sémantique conversationnelle dont le DST et améliorent la performance de TripPy.

Apprentissage avec peu de données Le manque de données annotées empêche le développement de modèles de DST performants et robustes. Cependant, le processus de collecte de données est coûteux et prend du temps. Une approche pour aborder ce problème est d'apprendre un modèle sur des domaines riches en ressources et de l'appliquer à un domaine non vu avec peu ou pas de données (*few-shot* et *zero-shot learning*). Dingliwal *et al.* (2021) adoptent le méta-apprentissage et utilisent les domaines sources pour méta-apprendre les paramètres du modèle utilisé et initialiser l'affinage pour le domaine cible. Les travaux autour des jeux de données basés sur des schémas. Rastogi *et al.* (2020a) utilisent les description de slots pour traiter des domaines et slots non vus (Lin *et al.*, 2021c; Zhao *et al.*, 2022). L'une des contraintes de ces méthodes est qu'elles reposent sur la similarité entre le domaine non vu et les domaines utilisés pour l'apprentissage initial. Une autre série d'approches tente d'exploiter des connaissances externes. Hudeček *et al.* (2021) utilisent l'analyse sémantique FrameNet comme supervision faible pour identifier des slots potentiels. La formulation du DST comme une tâche de compréhension permet d'accéder à des ressources plus importantes. Gao *et al.* (2020); Li *et al.* (2021a); Lin *et al.* (2021b) proposent différentes méthodes pour pré-entraîner un modèle sur des données de compréhension abondantes en vue de l'appliquer au DST. À noter que les approches d'adaptation des PLM vues plus haut permettent un apprentissage plus efficace lorsque peu de données sont disponibles et constituent aussi une des solutions potentielles au problème du manque de données (Wu *et al.*, 2020). Dans ce sens, Mi *et al.* (2021) présentent une méthode d'auto-apprentissage complémentaire aux PLM adaptés au dialogue comme TOD-BERT.

Modèles seq2seq à base de *prompts* Récemment, l'approche *text-to-text* où les entrées et sorties du modèle sont des chaînes de caractères a permis d'aborder toutes les tâches de NLP avec un modèle unique (Raffel *et al.*, 2020). Cette approche repose sur la formulation de requêtes textuelles adéquates appelés *prompts*. Elle a été appliquée avec succès au DST et a permis de combler l'écart de performance entre les méthodes de classification et de génération. En plus des noms du slot et du domaine, Lee *et al.* (2021a) intègrent la description des slots dans la requête pour générer indépendamment les valeurs de chaque slot. Ils ajoutent également une liste de valeurs possibles pour les slots catégoriques. Zhao *et al.* (2021, 2022) développent cette approche en générant l'état du dialogue complet d'une traite. Ces travaux se basent sur les slots pour formuler une requête et générer les valeurs correspondantes. Yang *et al.* (2022) inversent cette méthode et formulent une requête à partir de valeurs extraites depuis l'énoncé pour générer leur slot respectif. Ils cherchent ainsi à pallier le manque de données, argumentant que les slots qui figurent dans un corpus réduit ne représentent pas tous les besoins potentiels alors que les valeurs sont souvent mentionnées explicitement. Les modèles seq2seq ont également été utilisés pour une approche unifiée du dialogue orienté tâche qui génère conditionnellement l'état du dialogue, l'action du système puis la réponse du système (Lin *et al.*, 2020; Hosseini-Asl *et al.*, 2020; Peng *et al.*, 2021a).

Application à d'autres langues que l'anglais Jusqu'à récemment, la plupart des travaux autour du DST étaient cantonnés à l'anglais à cause du manque de données dans d'autres langues, empêchant la création de modèles véritablement multilingues. Ces dernières années, plusieurs travaux se sont penchés sur cette problématique. L'un des défis de DSTC9 visait à étudier le DST cross-lingue. Pour ce faire, une version chinoise de MultiWOZ et une version anglaise de CrossWOZ ont été obtenues par un procédé de traduction automatique suivi d'une correction manuelle (Gunasekara *et al.*, 2020). Zuo *et al.* (2021) ont utilisé la même méthode pour traduire MultiWOZ en chinois, coréen, vietnamien, hindi, français, portugais et thaï. Le problème des traductions automatiques est qu'elles manquent de naturel et ne sont pas localisées. Deux jeux de données en chinois ont été obtenus par une collecte H2H : CrossWOZ et RiSAWOZ (Zhu *et al.*, 2020; Quan *et al.*, 2020), toutefois ce type de collecte est coûteux. L'approche M2M permet d'obtenir des corpus multilingues adéquats en adaptant des squelettes de conversation selon la langue cible. Lin *et al.* (2021d) ont tiré profit de cette méthode pour créer BiToD, un corpus bilingue anglais-chinois. Majewska *et al.* (2022) ont utilisé les schémas issus de SGD pour créer des squelettes de conversations qui ont ensuite été adaptés en russe, arabe, indonésien et swahili. Enfin, pour pré-entraîner un modèle de DST cross-lingue, Moghe *et al.* (2021) tirent profit des données de sous-titres de films parallèles et conversationnelles.

4 Défis et futures directions

Malgré les avancées récentes, il reste de nombreux défis à relever pour obtenir des modèles de DST capables de saisir avec précision les besoins de l'utilisateur de manière dynamique et dans des scénarios variés. Il existe de nombreuses pistes intéressantes, cette section articule les besoins autour de trois axes : généralisation, robustesse et pertinence des modèles.

4.1 Modèles généralisables

Les systèmes de dialogue orientés tâches sont destinés à être déployés dans des environnements dynamiques qui peuvent impliquer des paramètres et des domaines différents. Dans la pratique, les domaines d'application de ces systèmes sont nombreux et variés (par ex. service à la clientèle des télécommunications, des banques, assistance technique, etc.), ce qui rend difficile voir impossible l'annotation manuelle de corpus pour chaque domaine. L'efficacité de l'apprentissage supervisé traditionnel est donc réduite et c'est une des raisons pour lesquelles la plupart des systèmes en production sont élaborés à partir de règles.

L'apprentissage avec peu ou pas de nouvelles données annotées offre une alternative pour garantir la flexibilité des systèmes. L'importance de cet aspect se reflète par les nombreux travaux récents qui abordent cette problématique, comme présenté en Section 3. Bien que des progrès notables aient été réalisés, ces approches restent limitées. Les modèles qui se basent sur des ressources de DST existantes sont incapables de traiter des domaines dont la distribution s'écarte de celle des données d'apprentissage (Dingliwal *et al.*, 2021). Les modèles qui utilisent des connaissances extérieures offrent une approche plus générique mais obtiennent des performances relativement faibles (Lin *et al.*, 2021b). L'apprentissage de modèles généralisables reste donc un problème ouvert. Une piste intéressante est l'apprentissage continu qui permet d'ajouter de nouvelles compétences à un système au fil du temps après le déploiement. Sans ré-entraînement avec toutes les données, le modèle doit être capable d'accumuler des connaissances (Madotto *et al.*, 2021; Liu *et al.*, 2021).

Dans un scénario réel, des entités qui ne sont pas observées durant l'apprentissage sont vouées à apparaître. Un modèle de DST doit pouvoir extrapoler à partir d'entités similaires vues durant l'apprentissage. De même, de par sa nature contrainte, un système de dialogue orienté tâche peut être perturbé par des énoncés hors domaine. Il est désirable qu'il puisse reconnaître de tels énoncés afin de fournir une réponse appropriée. La capacité de généralisation d'un modèle à des scénarios nouveaux est donc liée à sa robustesse.

4.2 Modèles robustes

Comme les utilisateurs peuvent exprimer des besoins similaires de différentes manières, un modèle de DST doit être capable d'interpréter différentes formulations d'une demande de manière cohérente. En d'autres termes, il doit être robuste aux variations de l'entrée. Des travaux d'analyse ont montré que la performance des modèles chute lorsqu'ils sont confrontés à des exemples réalistes qui s'écartent de la distribution du jeu d'évaluation (Huang *et al.*, 2021; Li *et al.*, 2021b). Lié à cela, un autre aspect peu étudié est le traitement d'énoncés qui dévient de la norme dans le cas d'un système de dialogue écrit.

Un modèle de DST doit être capable de prendre en compte tout l'historique et d'ajuster ses prédictions en utilisant toute l'information disponible. De nombreux travaux ont constaté que les performances se dégradent rapidement lorsque la longueur du dialogue augmente. Un autre aspect critique est donc la gestion efficace de dialogues longs (Zhang *et al.*, 2021). L'état du dialogue condense les informations importantes, mais il peut être difficile de corriger une erreur faite à un tour antérieur. Pour pallier ce phénomène de propagation des erreurs, Tian *et al.* (2021) utilisent une génération de l'état du dialogue en deux passes pour corriger les erreurs potentielles. Manotumruksa *et al.* (2021) proposent une fonction d'objectif basée sur les tours pour pénaliser le modèle en cas de prédiction incorrecte dans les premiers tours. Enfin, d'autres travaux simulent des dialogues plus longs en insérant des énoncés provenant de corpus de dialogue ouvert (*chit-chat*) dans des dialogues orientés tâche (Kottur *et al.*, 2021b; Sun *et al.*, 2021).

Un autre phénomène encore peu étudié dans le contexte des approches récentes est la robustesse aux entrées vocales. Pour les systèmes de dialogue parlé, plusieurs nouveaux défis se posent, par exemple les erreurs de l'ASR ou les disfluences verbales. Les premières éditions du DSTC avaient fourni des corpus de la parole qui incluaient une transcription et une liste des hypothèses de l'ASR. Depuis, les corpus de dialogues orientés tâche sont essentiellement issus du langage écrit. L'un des défis de DSTC10 considère de nouveau cet aspect et propose une tâche de DST où les jeux de développement et d'évaluation contiennent les hypothèses de l'ASR (Kim *et al.*, 2021).

Pour apprendre des modèles robustes, il est nécessaire de disposer de jeux de données diversifiés qui représentent les défis du monde réel. Dans ce sens, plusieurs cadres d'évaluation ont été publiés pour servir de plate-forme d'analyse pour les systèmes de dialogue orientés tâches robustes (Lee *et al.*, 2021b; Peng *et al.*, 2021b; Cho *et al.*, 2021). L'augmentation de données est une solution potentielle au problème du manque de variété dans les corpus (Campagna *et al.*, 2020; Li *et al.*, 2021b), notamment pour simuler des erreurs de l'ASR (Wang *et al.*, 2020).

4.3 Modèles pertinents

Comme on l'a vu, les jeux de données existants ne reflètent pas vraiment les conditions réelles ce qui résulte en une tâche plutôt artificielle. Il est important de garder une vision holistique pour le

développement de modèles de DST afin de s'assurer de la pertinence de leur application dans un système de dialogue.

Depuis les premiers systèmes de dialogues orientés tâche, l'état du dialogue est considéré comme un formulaire à remplir sous la forme de paires de (*slot*, *valeur*). Cette représentation fixe convient pour des tâches simples comme la réservation de vols, mais elle ne permet pas de traiter des domaines avec des structures relationnelles riches et un nombre variable d'entités. En effet, la gestion de la composition n'est pas possible (par ex. "itinéraire pour ma prochaine réunion"), et les connaissances ne sont pas directement partagées entre les slots. La Section 3 a présenté des approches qui tentent de répondre à ce dernier point en utilisant des graphes pour la représentation de l'état du dialogue. Pour promouvoir le traitement de scénarios plus réalistes, d'autres travaux proposent des représentations plus riches avec un corpus associé. (Andreas *et al.*, 2020) encodent l'état sous la forme d'un programme exécutable et proposent le corpus SMCallFlow. (Cheng *et al.*, 2020) quant à eux utilisent une structure arborescente. Dans le corpus ABCD, (Chen *et al.*, 2021) adoptent une représentation des procédures qu'un employé d'un service client doit suivre conformément aux politiques d'entreprise. Avec SIMMC 2.0, (Kottur *et al.*, 2021a) étendent la tâche du DST à une conversation multimodale. Dans l'état du dialogue, les slots se basent sur le contexte multimodal, ce qui nécessite la manipulation d'objets multimodaux (par opposition aux slots textuels).

Les modèles de DST sont presque essentiellement évalués en isolation. Or le dialogue est dynamique. Dans un scénario réel, la prédiction d'un état du dialogue erroné aurait dévié le cours de la conversation par rapport au dialogue de référence. Dès lors, une question importante est l'impact que le module de DST peut avoir sur le système de dialogue dans son ensemble. La plupart des études évaluent les modèles de manière isolée, en partant du principe qu'il est toujours possible d'assembler un ensemble de modules performants pour construire un bon système de dialogue. La performance globale d'un système est rarement prise en compte. (Takanobu *et al.*, 2020) ont mené des évaluations simulées et humaines de systèmes de dialogue avec une grande variété de configurations et de paramètres sur le corpus MultiWOZ. Ils ont constaté une chute du taux de réussite lorsqu'un modèle de DST est utilisé plutôt qu'un modèle de NLU suivi d'une mise à jour de l'état du dialogue à base de règles. Ils expliquent ce résultat par le fait que le NLU extrait les intentions de l'utilisateur en plus des paires de (*slot*, *valeur*). Cette étude est rare dans son genre et appelle à davantage de travaux similaires.

5 Conclusion

Le suivi de l'état du dialogue est une composante cruciale d'un agent conversationnel qui permet d'identifier les besoins de l'utilisateur à chaque étape de la conversation. Un nombre croissant de travaux s'intéressent à cette tâche et cet article a dressé l'état de l'art des avancées récentes. Après avoir donné une vue d'ensemble de la tâche et des différents jeux de données disponibles, nous avons catégorisé les approches neuronales modernes selon l'inférence de l'état du dialogue. Malgré des résultats encourageants sur les jeux d'évaluation comme MultiWOZ, ces systèmes manquent de flexibilité et de robustesse, aptitudes critiques pour un système de dialogue. Ces dernières années, de nombreux travaux ont cherché à aborder ces limitations et nous en avons synthétisé les avancées. Toutefois, il reste encore des défis importants à relever à l'avenir. Il existe de nombreuses pistes intéressantes et nous avons proposé trois caractéristiques essentielles des modèles de DST pour guider les futures recherches : généralisables, robustes et pertinents.

Remerciements

Je tiens à remercier les relecteurs anonymes ainsi que mes encadrants Lina Rojas-Barahona et Benoit Favre pour leurs conseils et commentaires pertinents concernant cet article.

Références

ANDREAS J., BUFE J., BURKETT D., CHEN C., CLAUSMAN J., CRAWFORD J., CRIM K., DELOACH J., DORNER L., EISNER J., FANG H., GUO A., HALL D., HAYES K., HILL K., HO D., IWASZUK W., JHA S., KLEIN D., KRISHNAMURTHY J., LANMAN T., LIANG P., LIN C. H., LINTSBAKH I., MCGOVERN A., NISNEVICH A., PAULS A., PETTERS D., READ B., ROTH D., ROY S., RUSAK J., SHORT B., SLOMIN D., SNYDER B., STRIPLIN S., SU Y., TELLMAN Z., THOMSON S., VOROBEV A., WITOSZKO I., WOLFE J., WRAY A., ZHANG Y. & ZOTOV A. (2020). Task-Oriented Dialogue as Dataflow Synthesis. *Transactions of the Association for Computational Linguistics*, **8**, 556–571. DOI : 10.1162/tacl_a_00333.

BALARAMAN V., SHEIKHALISHAHI S. & MAGNINI B. (2021). Recent Neural Methods on Dialogue State Tracking for Task-Oriented Dialogue Systems : A Survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 239–251, Singapore and Online : Association for Computational Linguistics.

BUDZIANOWSKI P., WEN T.-H., TSENG B.-H., CASANUEVA I., ULTES S., RAMADAN O. & GAŠIĆ M. (2018). MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 5016–5026, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1547.

CAMPAGNA G., FORYCIARZ A., MORADSHAHI M. & LAM M. (2020). Zero-Shot Transfer Learning with Synthesized Data for Multi-Domain Dialogue State Tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 122–132, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.12.

CHEN D., CHEN H., YANG Y., LIN A. & YU Z. (2021). Action-Based Conversations Dataset : A Corpus for Building More In-Depth Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3002–3017, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.239.

CHEN H., LIU X., YIN D. & TANG J. (2017). A Survey on Dialogue Systems : Recent Advances and New Frontiers. *ACM SIGKDD Explorations Newsletter*, **19**(2), 25–35. DOI : 10.1145/3166054.3166058.

CHEN L., LV B., WANG C., ZHU S., TAN B. & YU K. (2020). Schema-Guided Multi-Domain Dialogue State Tracking with Graph Attention Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(05), 7521–7528. DOI : 10.1609/aaai.v34i05.6250.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

p. 8107–8117, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.651.

CHIANG T.-R. & YEH Y.-T. (2021). Improving Dialogue State Tracking by Joint Slot Modeling. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, p. 155–162, Online : Association for Computational Linguistics.

CHO H., SANKAR C., LIN C., SADAGOPAN K. R., SHAYANDEH S., CELIKYILMAZ A., MAY J. & BEIRAMI A. (2021). CheckDST : Measuring Real-World Generalization of Dialogue State Tracking Performance. *arXiv :2112.08321 [cs]*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DINGLIWAL S., GAO S., AGARWAL S., LIN C.-W., CHUNG T. & HAKKANI-TUR D. (2021). Few Shot Dialogue State Tracking using Meta-learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1730–1739, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.148.

ERIC M., GOEL R., PAUL S., KUMAR A., SETHI A., KU P., GOYAL A. K., AGARWAL S., GAO S. & HAKKANI-TUR D. (2019). MultiWOZ 2.1 : A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. *arXiv :1907.01669 [cs]*.

FENG Y., WANG Y. & LI H. (2021). A Sequence-to-Sequence Approach to Dialogue State Tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1714–1725, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.135.

GAO J., GALLEY M. & LI L. (2018). Neural Approaches to Conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, p. 1371–1374, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3209978.3210183.

GAO S., AGARWAL S., JIN D., CHUNG T. & HAKKANI-TUR D. (2020). From Machine Reading Comprehension to Dialogue State Tracking : Bridging the Gap. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, p. 79–89, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.nlp4convai-1.10.

GAO S., SETHI A., AGARWAL S., CHUNG T. & HAKKANI-TUR D. (2019). Dialog State Tracking : A Neural Reading Comprehension Approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, p. 264–273, Stockholm, Sweden : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5932.

GUNASEKARA C., KIM S., D'HARO L. F., RASTOGI A., CHEN Y.-N., ERIC M., HEDAYATNIA B., GOPALAKRISHNAN K., LIU Y., HUANG C.-W., HAKKANI-TÜR D., LI J., ZHU Q., LUO L., LIDEN L., HUANG K., SHAYANDEH S., LIANG R., PENG B., ZHANG Z., SHUKLA S., HUANG M., GAO J., MEHRI S., FENG Y., GORDON C., ALAVI S. H., TRAUM D., ESKENAZI M., BEIRAMI A., EUNJOON, CHO, CROOK P. A., DE A., GERAMIFARD A., KOTTUR S., MOON S., PODDAR S. & SUBBA R. (2020). Overview of the Ninth Dialog System Technology Challenge : DSTC9. *arXiv :2011.06486 [cs]*.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8342–8360, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.740.

HAN T., LIU X., TAKANOBU R., LIAN Y., HUANG C., WAN D., PENG W. & HUANG M. (2021). MultiWOZ 2.3 : A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. *arXiv :2010.05594 [cs]*.

HECK M., VAN NIEKERK C., LUBIS N., GEISHAUSER C., LIN H.-C., MORESI M. & GASIC M. (2020). TripPy : A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 35–44, 1st virtual meeting : Association for Computational Linguistics.

HENDERSON M. (2015). Machine Learning for Dialog State Tracking : A Review. In *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing*.

HENDERSON M., THOMSON B. & WILLIAMS J. D. (2014a). The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, p. 263–272, Philadelphia, PA, U.S.A. : Association for Computational Linguistics. DOI : 10.3115/v1/W14-4337.

HENDERSON M., THOMSON B. & YOUNG S. (2014b). Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, p. 292–299, Philadelphia, PA, U.S.A. : Association for Computational Linguistics. DOI : 10.3115/v1/W14-4340.

HOSSEINI-ASL E., MCCANN B., WU C.-S., YAVUZ S. & SOCHER R. (2020). A simple language model for task-oriented dialogue. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 20179–20191 : Curran Associates, Inc.

HUANG Y., FENG J., WU X. & DU X. (2021). Counterfactual Matters : Intrinsic Probing For Dialogue State Tracking. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, p. 1–6, Online : Association for Computational Linguistics.

HUDEČEK V., DUŠEK O. & YU Z. (2021). Discovering Dialogue Slots with Weak Supervision. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 2430–2442, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.189.

HUNG C.-C., LAUSCHER A., PONZETTO S. P. & GLAVAŠ G. (2021). DS-TOD : Efficient Domain Specialization for Task Oriented Dialog. *arXiv :2110.08395 [cs]*.

KELLEY J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, **2**(1), 26–41. DOI : 10.1145/357417.357420.

KIM S., LIU Y., JIN D., PAPANGELIS A., GOPALAKRISHNAN K., HEDAYATNIA B. & HAKKANI-TÜR D. Z. (2021). "How Robust R U ?" : Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. DOI : 10.1109/ASRU51503.2021.9688274.

KIM S., YANG S., KIM G. & LEE S.-W. (2020). Efficient Dialogue State Tracking by Selectively Overwriting Memory. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, p. 567–582, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.53.

KOTTUR S., MOON S., GERAMIFARD A. & DAMAVANDI B. (2021a). SIMMC 2.0 : A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4903–4912, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

KOTTUR S., SANKAR C., YU Z. & GERAMIFARD A. (2021b). DialogStitch : Synthetic Deeper and Multi-Context Task-Oriented Dialogs. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 21–26, Singapore and Online : Association for Computational Linguistics.

LE H., SOCHER R. & HOI S. C. H. (2020). Non-Autoregressive Dialog State Tracking. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.

LEE C.-H., CHENG H. & OSTENDORF M. (2021a). Dialogue State Tracking with a Language Model using Schema-Driven Prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4937–4949, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LEE H., GUPTA R., RASTOGI A., CAO Y., ZHANG B. & WU Y. (2021b). SGD-X : A Benchmark for Robust Generalization in Schema-Guided Dialogue Systems. *arXiv :2110.06800 [cs]*.

LI S., CAO J., SRIDHAR M., ZHU H., LI S.-W., HAMZA W. & MCAULEY J. (2021a). Zero-shot Generalization in Dialog State Tracking through Generative Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 1063–1074, Online : Association for Computational Linguistics.

LI S., YAVUZ S., HASHIMOTO K., LI J., NIU T., RAJANI N., YAN X., ZHOU Y. & XIONG C. (2021b). CoCo : Controllable Counterfactuals for Evaluating Dialogue State Trackers. *arXiv :2010.12850 [cs]*.

LIN W., TSENG B.-H. & BYRNE B. (2021a). Knowledge-Aware Graph-Enhanced GPT-2 for Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7871–7881, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LIN Z., LIU B., MADOTTO A., MOON S., ZHOU Z., CROOK P., WANG Z., YU Z., CHO E., SUBBA R. & FUNG P. (2021b). Zero-Shot Dialogue State Tracking via Cross-Task Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7890–7900, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LIN Z., LIU B., MOON S., CROOK P., ZHOU Z., WANG Z., YU Z., MADOTTO A., CHO E. & SUBBA R. (2021c). Leveraging Slot Descriptions for Zero-Shot Cross-Domain Dialogue StateTracking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5640–5648, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.448.

LIN Z., MADOTTO A., WINATA G. I. & FUNG P. (2020). MinTL : Minimalist Transfer Learning for Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 3391–3405, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.273.

LIN Z., MADOTTO A., WINATA G. I., XU P., JIANG F., HU Y., SHI C. & FUNG P. (2021d). 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)*.

LIU Q., CAO P., LIU C., CHEN J., CAI X., YANG F., HE S., LIU K. & ZHAO J. (2021). Domain-Lifelong Learning for Dialogue State Tracking via Knowledge Preservation Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 2301–2311, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

MADOTTO A., LIN Z., ZHOU Z., MOON S., CROOK P., LIU B., YU Z., CHO E., FUNG P. & WANG Z. (2021). Continual Learning in Task-Oriented Dialogue Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7452–7467, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

MAJEWSKA O., RAZUMOVSKAIA E., PONTI E. M., VULIĆ I. & KORHONEN A. (2022). Cross-Lingual Dialogue Dataset Creation via Outline-Based Generation. *arXiv :2201.13405 [cs]*.

MANOTUMRUKSA J., DALTON J., MEIJ E. & YILMAZ E. (2021). Improving Dialogue State Tracking with Turn-based Loss Function and Sequential Data Augmentation. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 1674–1683, Punta Cana, Dominican Republic : Association for Computational Linguistics.

MI F., ZHOU W., KONG L., CAI F., HUANG M. & FALTINGS B. (2021). Self-training Improves Pre-training for Few-shot Learning in Task-oriented Dialog Systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1887–1898, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

MOGHE N., STEEDMAN M. & BIRCH A. (2021). Cross-lingual Intermediate Fine-tuning improves Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1137–1150, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

MRKŠIĆ N., Ó SÉAGHDHA D., WEN T.-H., THOMSON B. & YOUNG S. (2017). Neural Belief Tracker : Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1777–1788, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1163.

PENG B., LI C., LI J., SHAYANDEH S., LIDEN L. & GAO J. (2021a). SOLOIST : Building Task Bots at Scale with Transfer Learning and Machine Teaching. *arXiv :2005.05298 [cs]*.

PENG B., LI C., ZHANG Z., ZHU C., LI J. & GAO J. (2021b). RADDLE : An Evaluation Benchmark and Analysis Platform for Robust Task-oriented Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 4418–4429, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.341.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 930–940, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.67.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

RASTOGI A., ZANG X., SUNKARA S., GUPTA R. & KHAITAN P. (2020a). Schema-Guided Dialogue State Tracking Task at DSTC8. *arXiv :2002.01359 [cs]*.

RASTOGI A., ZANG X., SUNKARA S., GUPTA R. & KHAITAN P. (2020b). Towards Scalable Multi-Domain Conversational Agents : The Schema-Guided Dialogue Dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, p. 8689–8696 : AAAI Press.

SUN K., MOON S., CROOK P., ROLLER S., SILVERT B., LIU B., WANG Z., LIU H., CHO E. & CARDIE C. (2021). Adding Chit-Chat to Enhance Task-Oriented Dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1570–1583, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.124.

TAKANOBU R., ZHU Q., LI J., PENG B., GAO J. & HUANG M. (2020). Is Your Goal-Oriented Dialog Model Performing Really Well ? Empirical Analysis of System-wise Evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 297–310, 1st virtual meeting : Association for Computational Linguistics.

TIAN X., HUANG L., LIN Y., BAO S., HE H., YANG Y., WU H., WANG F. & SUN S. (2021). Amendable Generation for Dialogue State Tracking. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, p. 80–92, Online : Association for Computational Linguistics.

WANG L., FAZEL-ZARANDI M., TIWARI A., MATSOUKAS S. & POLYMENAKOS L. (2020). Data Augmentation for Training Dialog Models Robust to Speech Recognition Errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, p. 63–70, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.nlp4convai-1.8.

WEN T.-H., VANDYKE D., MRKŠIĆ N., GAŠIĆ M., ROJAS-BARAHONA L. M., SU P.-H., ULTES S. & YOUNG S. (2017). A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 438–449, Valencia, Spain : Association for Computational Linguistics.

WILLIAMS J., RAUX A., RAMACHANDRAN D. & BLACK A. (2013). The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, p. 404–413, Metz, France : Association for Computational Linguistics.

WILLIAMS J. D., RAUX A. & HENDERSON M. (2016). The Dialog State Tracking Challenge Series : A Review. *Dialogue & Discourse*, **7**(3), 4–33. DOI : 10.5087/dad.2016.301.

WU C.-S., HOI S. C., SOCHER R. & XIONG C. (2020). TOD-BERT : Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 917–929, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.66.

WU C.-S., MADOTTO A., HOSSEINI-ASL E., XIONG C., SOCHER R. & FUNG P. (2019). Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 808–819, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1078.

YANG Y., LEI W., CAO J., LI J. & CHUA T.-S. (2022). Prompt Learning for Few-Shot Dialogue State Tracking. *arXiv :2201.05780 [cs]*.

YE F., MANOTUMRUKSA J. & YILMAZ E. (2021a). MultiWOZ 2.4 : A Multi-Domain Task-Oriented Dialogue Dataset with Essential Annotation Corrections to Improve State Tracking Evaluation. *arXiv :2104.00773 [cs]*.

YE F., MANOTUMRUKSA J., ZHANG Q., LI S. & YILMAZ E. (2021b). Slot Self-Attentive Dialogue State Tracking. In *Proceedings of the Web Conference 2021*, WWW '21, p. 1598–1608, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3442381.3449939.

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. DOI : 10.1109/JPROC.2012.2225812.

YU T., ZHANG R., POLOZOV A., MEEK C. & AWADALLAH A. H. (2021). SCoRe : Pre-Training for Context Representation in Conversational Semantic Parsing. In *ICLR*.

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*, p. 109–117, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.nlp4convai-1.13.

ZHANG J., HASHIMOTO K., WU C.-S., WANG Y., YU P., SOCHER R. & XIONG C. (2020a). Find or Classify ? Dual Strategy for Slot-Value Predictions on Multi-Domain Dialog State Tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, p. 154–167, Barcelona, Spain (Online) : Association for Computational Linguistics.

ZHANG Y., CAO Y., MAHDIEH M., ZHAO J. & WU Y. (2021). Improving Longer-range Dialogue State Tracking. *arXiv :2103.00109 [cs]*.

ZHANG Z., TAKANOBU R., ZHU Q., HUANG M. & ZHU X. (2020b). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, **63**(10), 2011–2027. DOI : 10.1007/s11431-020-1692-3.

ZHAO J., GUPTA R., CAO Y., YU D., WANG M., LEE H., RASTOGI A., SHAFRAN I. & WU Y. (2022). Description-Driven Task-Oriented Dialog Modeling. *arXiv :2201.08904 [cs]*.

ZHAO J., MAHDIEH M., ZHANG Y., CAO Y. & WU Y. (2021). Effective Sequence-to-Sequence Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7486–7493, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

ZHOU L. & SMALL K. (2019). Multi-domain Dialogue State Tracking as Dynamic Knowledge Graph Enhanced Question Answering. *ArXiv*.

ZHU Q., GU Y., LUO L., LI B., LI C., PENG W., HUANG M. & ZHU X. (2021). When does Further Pre-training MLM Help ? An Empirical Study on Task-Oriented Dialog Pre-training. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, p. 54–61, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

ZHU Q., HUANG K., ZHANG Z., ZHU X. & HUANG M. (2020). CrossWOZ : A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. *Transactions of the Association for Computational Linguistics*, **8**, 281–295. DOI : 10.1162/tacl_a_00314.

ZUO L., QIAN K., YANG B. & YU Z. (2021). AllWOZ : Towards Multilingual Task-Oriented Dialog Systems for All. *arXiv :2112.08333 [cs]*.