UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Human-Machine Dialogue in the Medical Field. Using Dialog to Collect Important Patient Information

THÈSE

présentée et soutenue publiquement le 20 Octobre 2022

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Anna Liednikova

Composition du jury

Rapporteurs :	Anne Vilnat	Professeure, Université Paris-Saclay, France
	Sandra Bringay	Professeure, Université Paul-Valéry Montpellier 3, France
Président du jury :	Yannick Toussaint	Professeur, Université de Lorraine, France
Invité :	Philippe Jolivet	CEO et co-fondateur d'ALIAE, France
Directeur de thèse :	Claire Gardent	Directrice de recherche, CNRS, LORIA, France

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Acknowledgments

This thesis CIFRE was financed by ALIAE.

I would like to express gratitude to my supervisors, who guided me through this long and challenging journey, and all the people who contributed directly or indirectly to it.

I would like to thank Philippe for this inspiring mission and project, for being available for discussions and synchronization that helped to better identify business needs that should be taken into account for research, for constant support and encouragement

I would like to thank Claire for navigating me in academia world, inculcating useful habits and routines, for enframing our ambitions into the "paper" ground, giving constructive feedback for the ideas and helping develop writing and thinking stage by stage.

I would like to thank Alexandre for being part of important and turning discussions, Alice, Célian and Philippe for sharing their expertise on following the patients that significantly influenced the start of the thought process of this research, and Barbara for giving valuable feedbacks on the prototypes.

Our work in the team started before the PhD studies during my Erasmus program in Mines Nancy. I would like to thank Yannick for making this valuable connection, my alma mater for giving me an opportunity to study abroad and Mines Nancy for welcoming.

I thank my thesis committee and reviewers, Anne Vilnat, Sandra Bringay and Yannick Toussaint for showing interest in my research work and their generous time.

I would like to thank my both team members! It was a pleasure to be part of SYNALP team as well as LORIA community, I appreciate support and inspiration giving by them. A lot of exchanges, meeting and scientific events definitely nourished my research journey both on personal and professional levels. I would like to thank ALIAE team for our discussions and collaborations, for precious moments of common achievements. I would like also to thank Farnaz for long, pleasant and enriching cooperation.

I would like to thank my friends for being by my side, for our gathering together, journeys and discussions that gave fuel to live and move forward.

I would like to thank my family who reminded me to stay open to academia perspectives, and my mum for showing an example of a strong and caring woman expressing her creativity and intelligence despite chaos and obstacles on the route. I wish she could join me in the end of this journey, just as we discussed, when she was still alive.

While trauma keeps us dumbfounded,
the path out of it is paved with words,
carefully assembled, piece by piece,
until the whole story can be revealed.
"The Body Keeps The Score",
Bessel Van der Kolk

Abstract

Healthcare dialogue systems are developed to automate and simplify routine tasks such as collecting patient information or making an appointment. Often, these models are trained to mimic doctor-patient interaction as their constant availability is a key feature for patients, in particular with chronic conditions. Chronic patients regularly visit their doctor and are asked to repeatedly fill in standardized questionnaires, which may trigger repetitive, incorrect input. In collaboration with the ALIAE company, we focus on developing novel dialogue models which can maintain a conversation with the patient while collecting both specific answers to a set of pre-defined questions and serendipitous information about the patient condition that we conjecture, may be useful for the patient treatment. Specifically, we propose a dialogue system and automatic questionnaire filling pipeline that should complement existing routine between doctors and patients.

This thesis makes three main contributions. We first propose an approach to flexibly guide the user through a pre-defined medical decision tree using naturally written user input – this allows the health-bot to collect answers to a set of pre-defined questions. To improve user engagement, increase the probability of all required medical topics being addressed and allow for further information about the patient condition to be collected, we then extend this initial model by integrating additional bots designed to handle health-related follow-up questions and maintain small talk. Finally, we introduce novel zero-shot Question Answering models and pre-processing techniques so that standard, clinical questionnaires can be automatically filled in based on the content of collected human-bot dialogues.

Keywords: natural language generation, human-machine dialog, healthbot, dialogue retrieval models, dialogue-based question answering

Résumé

Les systèmes de dialogue pour les soins de santé sont développés pour automatiser et simplifier les tâches de routine telles que la collecte d'informations sur les patients ou la prise de rendez-vous. Souvent, ces modèles sont entraînés à imiter l'interaction entre le médecin et le patient, en étant disponibles à tout moment. C'est particulièrement important pour les patients souffrant de maladies chroniques. Ils consultent régulièrement un médecin et remplissent des questionnaires standardisés, tout en s'habituant à donner les mêmes réponses et en étant confrontés à des difficultés entre les visites. En collaboration avec l'entreprise ALIAE, nous proposons un système de dialogue et de remplissage automatique de questionnaires qui devrait compléter la routine existante entre médecins et patients.

Dans cette thèse, nous proposons une approche pour guider de manière flexible le dialogue à travers un arbre de décision médical prédéfini basé sur les entrées naturelles de l'utilisateur plutôt que sur des réponses oui/non. Ensuite, nous abordons la nécessité d'avoir des bots supplémentaires pour poser des questions de suivi liées à la santé et maintenir de petites discussions afin d'améliorer l'engagement de l'utilisateur et d'augmenter la probabilité de revenir sur les sujets de l'arbre requis. Enfin, nous explorons des modèles et des techniques de prétraitement pour le remplissage automatique de questionnaires sur la base des dialogues obtenus dans un cadre zéro-shot.

Mots-clés: génération de langage naturel, dialogue homme-machine, healthbot, modèles de recherche de dialogue, réponse aux questions basée sur le dialogue

Contents

List of Figures

List of Tables

French Abstract

Dialogue homme-machine dans le domaine médical. Utilisation de Dialogue pour collecter les informations importantes sur le patient

Les systèmes de dialogue ont récemment commencé à faire leur apparition dans le secteur des soins de santé. Ces Health-Bots sont utilisés, par exemple, pour fournir des informations relatives à la santé, pour aider l'utilisateur à fixer un rendez-vous, pour donner un diagnostic probable sur la base d'une liste de symptômes, ou pour fournir des réponses spécifiques (FAQ améliorée). Ces agents de dialogue sont souvent des systèmes de reconnaissance de formes fragiles ou mettent en œuvre un arbre de dialogue prédéfini et stable. Ils couvrent habituellement un domaine très spécifique et sont généralement incapables de maintenir une conversation durable telle qu'elle se produit habituellement, par exemple, entre un médecin et un patient.

Pouvoir créer de telles interactions en marge des visites médicales pourrait être particulièrement important pour accompagner les patients atteints de maladies chroniques. Des études [1] ont montré qu'il est important de comprendre la qualité de vie des patients de leur propre point de vue. La collecte d'informations périphériques aux sujets requis peut donner une compréhension plus holistique du contexte du patient et aider à comprendre comment les choix de style de vie influencent la dynamique de la pathologie. C'est un sujet très difficile qui n'a pas encore été abordé. Dans ce travail, nous aimerions faire quelques premiers pas dans cette direction.

Entre les visites, la façon traditionnelle de mesurer l'état actuel des patients est de recourir à des questionnaires standardisés. Les limites de ces questionnaires sont leur manque de personnalisation, leur répétitivité et leur perspective unidirectionnelle (celle du médecin). Les nouvelles technologies qui permettent une communication bidirectionnelle sont plus appropriées pour obtenir des évaluations et un suivi plus pertinents des patients, et pour mieux personnaliser les soins à terme. En particulier, les robots de santé gagnent rapidement du terrain en tant que technologie clé qui permettrait de compléter les questionnaires standardisés utilisés traditionnellement pour recueillir des informations sur les patients par une interaction plus informelle et personnalisée, reproduisant et élargissant le dialogue patient-médecin dans un nouveau type de relation.

Cette thèse a bénéficié d'un contrat CIFRE et les modèles de robots de santé que nous avons développés visaient à répondre à un objectif clé de la société ALIAE [2], à savoir :

> Comment développer des robots de santé qui peuvent compléter les questionnaires standards utilisés dans les programmes de soins aux patients et être utilisés pour collecter des informations à la fois spécifiques et fortuites sur l'état du patient ?

Notre objectif à long terme est de développer un système de dialogue homme-machine qui compléterait les questionnaires cliniques standard traditionnellement utilisés dans les études cliniques ou les programmes de soins aux patients en engageant régulièrement le patient dans un dialogue sur les sujets du questionnaire. Comme nos utilisateurs cibles sont des patients chroniques, il est plus important de les garder engagés pendant une longue période pour obtenir une évolution de leur qualité de vie au fil du temps plutôt que d'obtenir toutes les informations immédiatement lors de la première conversation. Cependant, dans ce travail, nous commençons par les premières étapes : développer un système de dialogue pour la première interaction et remplir le questionnaire sur la base de cette première conversation.

Il est également important de souligner que notre système n'est pas destiné à remplacer le médecin, mais plutôt à le compléter. Dans ce travail, nous ne cherchons pas à reproduire le comportement du médecin, mais plutôt à l'enrichir sur les aspects de la qualité de vie liés à

[1] https://www.fda.gov/media/117890/download
[2] aliae.io

la pathologie. Voici les principales différences que nous allons aborder dans cette thèse : la flexibilité du changement de sujet, la possibilité de parler au-delà des sujets médicaux, et le fait d'éviter les énoncés répétitifs du questionnaire.

L'approche que nous utilisons pour recueillir des informations consiste à donner la liberté à l'utilisateur de parler de quelque chose qui le dérange réellement. Une conversation décontractée peut être un lien naturel entre les sujets requis dans le questionnaire, et une occasion de découvrir le contexte de la vie du patient et de créer une relation avec lui. Elle peut nous aider à recevoir des informations supplémentaires sur le patient qui ne correspondent pas à l'arbre médical prédéfini.

Dès que le système de dialogue est opérationnel, il est nécessaire d'extraire automatiquement de ces dialogues les informations du patient qui sont pertinentes pour son traitement, mais qui ne seraient pas nécessairement détectées dans l'approche du questionnaire standardisé. Aussi, au lieu de donner le même énoncé de questionnaire à répondre par l'utilisateur, nous développons le système de dialogue de telle sorte qu'il est possible de collecter le contexte nécessaire afin de dériver les réponses à ces énoncés de manière implicite et automatisée.

Cette thèse propose un pipeline complémentaire à la routine habituelle médecin-patient avec un système de dialogue comme acteur de base.

Les pathologies chroniques sont impactées par l'environnement des patients, leur qualité de vie, leur vie sociale, leur état mental. Afin de mieux personnaliser les soins, il est de plus en plus important de comprendre la personne dans son environnement, d'analyser son état, ses émotions, ses comportements et ses interactions particulières.

La partie I (Contexte) présente le domaine des systèmes de dialogue. Dans le chapitre 1, nous discutons de la manière dont les systèmes de dialogue peuvent différer en fonction de leur objectif (axé sur une tâche, bavardage ou leur combinaison) et nous présentons la manière dont ils sont utilisés pour les applications médicales et de santé. Dans la section 2, nous examinons de plus près leurs architectures modulaires et de bout en bout. Dans la section 3, nous donnons un aperçu des modèles de réseaux neuronaux de base pour les systèmes de recherche et de dialogue génératif. Enfin, dans la section 4, nous présentons différentes approches pour l'évaluation humaine et automatique du dialogue.

Chatbot médical pour la collecte d'informations sur les patients

La partie II se concentre sur le développement d'un système de dialogue pour les patients.

Tout d'abord, dans le chapitre 1, nous proposons un modèle de dialogue qui introduit une flexibilité entre l'arbre médical prédéfini et les sujets introduits par l'utilisateur. Puisque l'absence de données de dialogue est le principal goulot d'étranglement pour l'entraînement d'un modèle de génération de réponses, nous avons introduit une approche d'augmentation des données basée sur un arbre de décision médical et des phrases de forum et nous avons testé des modèles de récupération et de génération. Sur la base des expériences, nous nous sommes rendus compte que, étant donné la liberté dont ils disposent, les utilisateurs ont tendance à utiliser le small talk et à s'écarter des sujets médicaux proposés.

Création de corpus de dialogue

Des études ont montré que les questionnaires fermés traditionnellement utilisés dans le cadre d'études cliniques ne permettent pas de recueillir des informations correctes et précises sur l'état des patients, car ceux-ci s'habituent aux questions et donnent systématiquement les mêmes réponses d'une interaction à l'autre. Notre objectif à long terme est de développer un système de dialogue homme-machine qui compléterait les questionnaires cliniques standard en engageant régulièrement le patient dans un dialogue sur les sujets du questionnaire. Comme nos utilisateurs cibles sont des patients souffrant de douleurs chroniques, il est plus important de les garder engagés pendant une longue période que d'obtenir toutes les informations lors de la première interaction.

Collecte d'un corpus initial auprès d'un expert. Pour créer notre corpus de dialogue, nous avons demandé à un médecin de formaliser des interactions patient-médecin typiques survenant dans le contexte d'une étude clinique sous la forme d'un arbre de dialogue décrivant les questions à poser et, pour chaque question, les réponses possibles. Les interactions couvrent quatre domaines, à savoir le sommeil, l'humeur, l'anxiété et les loisirs, et l'arbre de dialogue compte 58 nœuds. Un fragment de l'arbre de dialogue créé pour le domaine du sommeil est présenté dans la Figure 1 à gauche, et un exemple de dialogue pour le domaine du sommeil dans la même figure à droite. Nous appelons les données recueillies auprès de l'expert D_{init}.

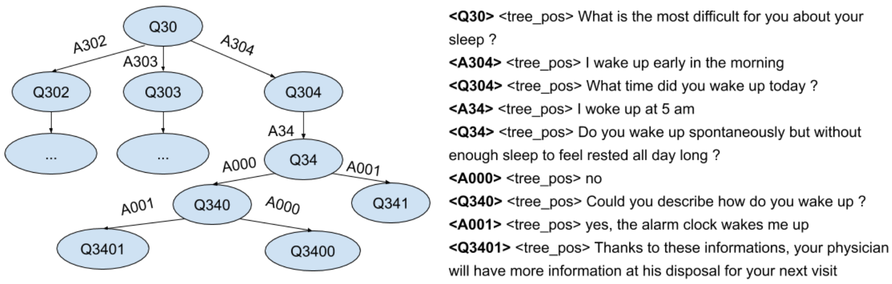


Figure 1: Fragment d'arbre de dialogue pour le domaine du sommeil et dialogue correspondant

Extraction des paraphrases. Nous extrayons les paraphrases des réponses des patients fournies par l'expert à partir du forum HealthBoard [3]. Comme les réponses du patient sont principalement des réponses affirmatives aux questions du médecin, nous commençons par filtrer les questions des données du forum pour ne garder que les énoncés qui sont des affirmations. Nous comparons ensuite chaque tour du patient dans D_{init} avec son contexte (P, le tour précédent du médecin) avec les énoncés assertifs extraits du forum. Pour chaque séquence $D + P$ de tours de patients contextualisés dans D_{init} et chaque énoncé (assertif) U dans le forum, nous créons un encastrement S-BERT (Reimers and Gurevych, 2019), extraire du forum tous les énoncés sémantiquement similaires et utiliser la pertinence marginale maximale (MMR) (Goldstein and Carbonell, 1998) pour sélectionner dans ce pool de candidats un sous-ensemble de paraphrases qui maximise à la fois la similarité (les paraphrases doivent être sémantiquement similaires autour d'entrée) et la diversité (l'ensemble de paraphrases résultant doit être le plus diversifié possible).

Comme l'illustre la figure 2, nous appliquons ce processus d'extraction de paraphrases non

[3] healthboards.com

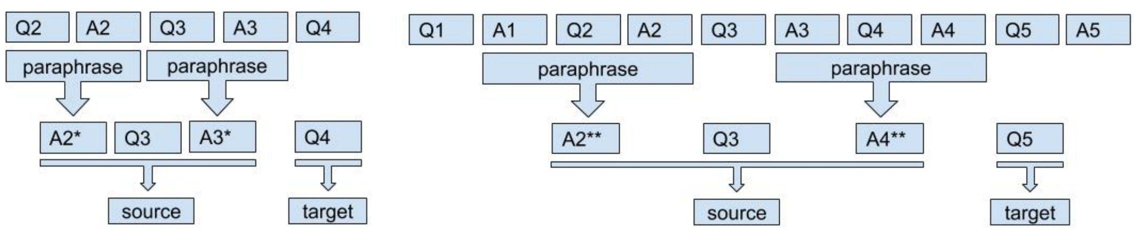


Figure 2: Paraphrasement de paires source-cible à partir de $INIT_{long}$ (gauche) et $INIT_{short}$ (droite)

seulement pour créer des paraphrases pour un seul tour, mais aussi pour créer des paraphrases qui résument 3 tours consécutifs. De cette façon, nous pouvons dériver des versions compressées des dialogues initiaux. Par exemple, nous pouvons dériver le court dialogue à partir de l'interaction de dialogue plus longue présentée :

Q1 : Dormez-vous bien ?
A1 : Non
Q2 : Qu'est-ce qui vous empêche de dormir ?
A2 : J'ai des douleurs dans les jambes

Q1 : Dormez-vous bien ?
A1 Q2 A2 : Non, j'ai mal aux jambes et ça m'empêche de dormir.

Nous désignons l'ensemble des paraphrases qui résument trois tours consécutifs comme SHORT et celles qui résument un seul tour comme LONG.

Med Bot Modèles

Notre objectif est d'apprendre un modèle qui imite un médecin dans le type d'interaction médecin-patient qui est typique des conversations des études cliniques.

Comme nous dérivons les données d'apprentissage de l'arbre de dialogue, chaque tour de parole du patient et chaque requête du médecin sont associés à un état de dialogue (un nœud dans cet arbre de dialogue). Nous utilisons cette double information (tour de dialogue et état du dialogue) pour entraîner et comparer deux modèles de génération de réponses : un modèle de classification qui, étant donné les trois derniers tours d'une interaction médecin-patient, prédit un état de dialogue et produit la requête correspondante du médecin ; et un modèle génératif séquence-à-séquence qui génère de manière auto-régressive une réponse tout en conditionnant les trois derniers tours de dialogue. Pour les deux modèles, nous utilisons une approche de pré-entraînement et de réglage fin similaire à celle présentée par Radford (2018).

L'entrée du modèle consiste en trois tours $\langle a_1 q_1 a_2 \rangle$. Nous concaténons ces trois tours, en préfixant chaque tour avec son identifiant d'état de dialogue et en les séparant avec un jeton de délimitation. Chaque jeton est représenté par la somme de trois incorporations : une incorporation de mot et une incorporation de position qui sont apprises dans la phase de pré-entraînement ; et une incorporation de tour (apprise pendant le réglage fin) indiquant si le jeton appartient à un tour de patient ou de médecin. L'entrée du modèle est la somme des trois types d'embeddings - mot, position et tour - pour chaque jeton de la séquence d'entrée.

ComBot, un modèle d'ensemble pour les interactions répétées basées sur les tâches

Dans le chapitre 2, nous présentons ComBot, un ensemble de trois modèles de dialogue : MedBot, destiné à suivre l'arbre de décision médical de manière flexible (voir le chapitre 1), FollowUpBot, destiné à chatter et à poser des questions de suivi sur des sujets de santé, même s'ils ne sont pas directement liés à l'arbre de décision médical, et EmpathyBot, un modèle de bavardage destiné à apporter de la compassion et à maintenir la conversation sur tout autre sujet non lié à la santé.

Chaque robot fournit un seul candidat. Pour les classer, nous encodons l'ensemble du contexte de dialogue actuel et chaque réponse du candidat en utilisant l'encodeur ConveRT, nous calculons la similarité (produit scalaire) pour chaque paire candidat/contexte, et nous sélectionnons le candidat ayant le score de similarité le plus élevé. Si tous les scores des candidats sont inférieurs à 0.1, nous considérons qu'il n'y a pas de bonne réponse et nous mettons fin à la conversation.

Medical Bot MedBot est un modèle de récupération qui utilise le modèle de sélection de réponses de dialogue ConveRT pré-entraîné (Henderson et al., 2019b) pour récupérer une requête à partir du Corpus MedTree (Liednikova et al., 2020). Il est conçu pour recueillir des informations auprès de l'utilisateur sur la base d'un ensemble prédéfini de questions contenues dans un questionnaire. Pour extraire du corpus MedTree la question qui correspond le mieux au contexte actuel du dialogue, le modèle MedBot compare les trois derniers tours du dialogue actuel avec les contextes du corpus MedTree. Le modèle identifie le contexte du corpus MedTree avec le score de similarité le plus élevé et produit la question qui suit ce contexte.

Follow-Up Bot L'une des principales motivations derrière l'utilisation d'un robot de santé dans les études cliniques est de compléter les informations traditionnellement recueillies par un questionnaire fixe rempli chaque semaine par les patients avec des informations fortuites, c'est-à-dire des informations qui ne sont pas activement demandées par le questionnaire, mais qui sont utiles pour analyser les résultats de la cohorte.

Le modèle MedBot présenté dans la section précédente est contraint de ne traiter que les sujets présents dans l'arbre de dialogue, ce qui revient à modéliser un questionnaire fermé. Pour permettre la collecte d'informations de santé fortuites, nous avons développé le FollowUpBot dont la fonction est de générer des questions de santé qui ne sont pas prédites par l'arbre de dialogue mais qui découlent naturellement de l'entrée de l'utilisateur. Plutôt que de restreindre artificiellement le dialogue à l'ensemble limité de sujets prédéfinis par l'arbre de dialogue, le modèle combiné (MedBot + FollowUpBot) permet des transitions basées soit sur l'arbre de dialogue, soit sur des questions de suivi liées à la santé. En ce sens, FollowUpBot permet non seulement la collecte d'informations fortuites liées à la santé, mais aussi des transitions de dialogue plus fluides.

Comme MedBot, FollowUpBot a utilisé le modèle ConveRT pré-entraîné pour récupérer des requêtes appropriées au contexte à partir d'un ensemble de données de dialogue. Dans ce cas, cependant, les requêtes sont extraites du jeu de données HealthBoard, un nouveau jeu de données que nous avons créé pour prendre en charge les questions de suivi dans le domaine de la santé.

Empathy Bot Comme son nom l'indique, le rôle de l'EmpathyBot est d'engager l'utilisateur en lui témoignant de l'empathie. Pour ce robot, nous utilisons le modèle génératif de Roller et al. (2021) qui a été pré-entraîné sur une variante de discussion Reddit (Baumgartner et al., 2020) et affiné sur le ConvAI2 (Zhang et al., 2018b), Wizard of Wikipedia (Dinan et al., 2018), Empathetic Dialogues (Rashkin et al., 2019), et Blended Skill Talk datasets (BST) (Smith et al., 2020) pour optimiser l'engagement et l'humanité dans la conversation en domaine ouvert.

Remplissage automatique et non supervisé de questionnaires cliniques

La partie III propose différentes approches pour remplir automatiquement les questionnaires à partir des dialogues obtenus avec ComBot. Étant donné un historique de dialogue humain-robot de santé D et un ensemble de questions q_i extraites d'un questionnaire Q, la tâche consiste à déterminer la réponse correcte a_i à chaque question q_i, y compris l'option "non mentionnée", c'est-à-dire la possibilité que le dialogue n'aborde pas cette question.

Tout d'abord, dans le chapitre 1, nous procédons à une catégorisation des questionnaires cliniques en fonction des types de questions et de réponses et proposons une procédure de collecte de données pour chacun d'entre eux. Ensuite, nous explorons comment le remplissage automatique de questionnaires pourrait être résolu en différentes tâches : réponse aux questions, classification de texte et inférence en langage naturel. Nous fournissons également une analyse détaillée de leurs performances pour chaque type de question et de réponse.

Types de questions Khashabi et al. (2020) a classé les ensembles de données question-response existants en quatre catégories en fonction du format des réponses : Extractives (EX), Abstractives (AB), à choix multiples (MC) et Oui/Non (YN). Dans la section suivante, nous souhaitons également présenter les réponses de type échelle de Likert et échelle visuelle analogique (EVA), qui sont très courantes dans les questionnaires cliniques. Sinha et al. (2017) ont conclu

Question type	Questionnaire
Open Question (OQ)	Morin (Morin, 1993) QCD (R. C. Daut, 1983) MOS-SS (Sherbourne and Stewart, 1991)
Closed Question (CQ)	QPC (C. Thomas-Antérion, 2004) EPICES (Bihan et al., 2005)
Agreement Likert-scale (ALS)	TSK (D.D. Kori, 1990) PBPI (Williams and Thorn, 1989) TAS-20 (Bagby et al., 1994) LOT-R (Scheier et al., 1994) IEQ-CF (R. Sullivan, 2008) JCK (KARASEK, 1985)
Frequency Likert-scale (FLS)	MOS-SS (Sherbourne and Stewart, 1991) MBI (Maslach et al., 1997) PCL-S (Weathers et al., 1993) HAD (Zigmond and Snaith, 1983) SF-12 (Ware Jr et al., 1996)
Visual Analogue Scale (VAS)	QCD (R. C. Daut, 1983) Dallas (Lawlis et al., 1989) LEEDS (Parrott and Hindmarch, 2004) FABQ-W (Waddell et al., 1993)

Table 1: Les questionnaires cliniques et leurs types de questions

dans leur étude que l'échelle visuelle analogique (EVA) et l'échelle d'évaluation numérique (NRS) étaient les échelles de douleur les mieux adaptées pour mesurer la douleur dans l'endométriose. De plus, l'EVA est souvent utilisée en recherche épidémiologique et clinique pour mesurer l'intensité ou la fréquence de divers symptômes (Paul-Dauphin, 1999). Pour chaque type de question présent dans le questionnaire médical, nous fournissons des exemples de question et de questionnaire dans le tableau 1.

Questionnaires Pour les expérimentations, nous avons choisi trois questionnaires sémantiquement proches des thèmes du modèle de chatbot : Morin (OQ), PBPI (ALS, VAS, CQ), Mos-ss (FLS). Pour le questionnaire PBPI, nous demandons aux annotateurs de donner les réponses dans trois types de format en même temps : CQ, ALS, VAS.

Collecte de données Pour créer un chatbot d'interaction, nous avons suivi l'ensemble ComBot (Liednikova et al., 2021). La conversation commence toujours par l'ouverture "Qu'est-ce qui est le plus difficile pour vous dans votre sommeil ?". Nous recueillons les dialogues en utilisant la plateforme Amazon Mechanical Turk et en demandant aux Turkers d'interagir avec le bot, tout en se comportant comme s'ils avaient besoin de questions. Pour éviter que les Turkers n'introduisent textuellement les questions dans le dialogue, on leur a donné une liste de sujets à mentionner plutôt que les questions elles-mêmes. De cette façon, nous nous assurons que les dialogues recueillis traitent des questions auxquelles il faut répondre tout en encourageant leur paraphrase diversifiée au cours de la conversation. Étant donné que la création d'un dialogue suivant les sujets du questionnaire est une tâche très consommatrice de ressources, nous avons décidé d'atténuer le risque que le système propose une direction non souhaitée de la conversation avec une fonction de rejet: ils pouvaient rejeter le candidat actuel, auquel cas, le tour avec le score de confiance le plus élevé suivant serait affiché par le bot. Les annotateurs étaient ensuite invités à remplir le questionnaire sur la base de leur conversation et à sélectionner l'option "Non mentionné" (NA) si l'historique du dialogue ne permettait pas de répondre à la question en cours.

Exploration de modèles pour les applications de type "zero-shot

Nous avons demandé à 10 annotateurs d'interagir avec le chatbot une fois pour chacun des trois questionnaires (Morin, PBPI, Mos-ss). Au total, nous avons donc recueilli 30 dialogues et les réponses aux questions correspondantes. Pour le questionnaire PBPI, nous demandons aux annotateurs de donner les réponses dans trois types de formats en même temps : CQ, ALS, VAS.

Modèles :

- **Modèle de Questions-Réponses** UnifiedQA (Khashabi et al., 2020) [4] est un modèle de Questions-Réponses pré-entraîné unique, qui donne de bons résultats sur 20 ensembles de données de Questions-Réponses couvrant 4 formats différents. En affinant ce modèle de Questions-Réponses pré-entraîné en modèles spécialisés, on obtient un nouvel état de l'art sur 10 ensembles de données de Questions-Réponses factoïdes et de sens commun, établissant UnifiedQA comme un point de départ solide pour la construction de systèmes de Questions-Réponses. Dans nos expériences, nous avons utilisé la version UnifiedQA-t5-3b.

- **Modèle MNLI** Nous utilisons le modèle xlarge DeBERTa V2(He et al., 2020)[5] ajusté avec le jeu de données MNLI (Williams et al., 2018) pour la tâche NLI. Nous transmettons au modèle l'historique du dialogue concaténé comme prémisse et la question sous forme déclarative comme hypothèse. La sortie est constituée de probabilités pour trois classes : Entaillement, Contradiction, Neutralité.

- **Modèle ZeroShot-TC** Nous utilisons le modèle Bart-large (Lewis et al., 2020)[6] pour la classification de texte zéro-shot entraînée sur le corpus MNLI (Williams et al., 2018). Dans ce cadre, nous passons l'historique des dialogues concaténés comme contexte et formulons les étiquettes cibles comme des modèles remplis, de sorte que les données d'entrée puissent être plus proches du format d'implication. Nous ajoutons ensuite une étiquette cible supplémentaire "NA" pour prendre en compte la situation où la réponse n'est pas mentionnée dans le dialogue. Le modèle fournit des scores de probabilité pour chaque candidat, et le candidat ayant la probabilité la plus élevée est choisi comme réponse finale.

Exploration de l'influence du format d'entrée du dialogue

Dans le chapitre 2, nous étudions comment la sélection et le prétraitement du contenu peuvent influencer la performance du modèle d'inférence en langage naturel dans un contexte de zéro coup pour le remplissage automatique de questionnaires. Pour l'évaluation, nous fournissons un ensemble de test composé de 100 dialogues et de leurs questions et réponses associées. Nous considérons deux types de questions : Les questions fermées (CQ) et les questions à échelle de Likert d'accord (ALS). Les CQ ont trois réponses possibles (oui, non ou Non Applicable, c'est-à-dire que le dialogue ne répond pas à la question) et les ALS en ont cinq (totalement en désaccord, plutôt en désaccord, d'accord, totalement d'accord, NA).

Suivant Ghassemi Toudeshki et al. (2021), nous modélisons la réponse aux questions comme une tâche NLI où la prémisse est dérivée du dialogue, l'hypothèse de la question et la réponse du

[4] https://github.com/allenai/unifiedqa
[5] https://github.com/microsoft/DeBERTa
[6] https://huggingface.co/facebook/bart-large-mnli

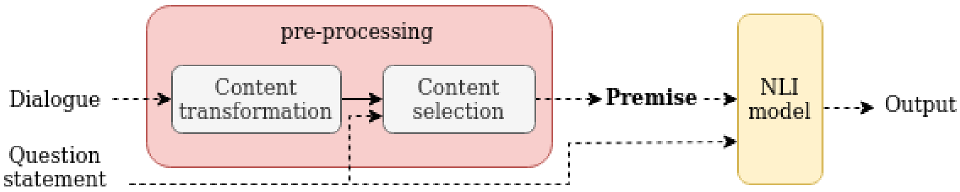


Figure 3: Schéma de prétraitement du dialogue

résultat NLI. Étant donné une question et un dialogue, notre modèle, illustré dans la Figure 3, répond à la question en trois étapes comme suit :

- **Dérivation d'une prémisse NLI à partir du dialogue** La prémisse NLI est dérivée du dialogue d'entrée en utilisant d'abord la transformation du contenu et ensuite la sélection du contenu. Comme détaillé dans la section 2.5, nous expérimentons différentes manières de transformer et de sélectionner le contenu.

- **Dériver une hypothèse NLI à partir d'une question** Pour dériver une hypothèse NLI d'une question, nous représentons simplement les questions sous forme d'énoncés (par exemple, "J'ai régulièrement des douleurs" au lieu de "Avez-vous régulièrement des douleurs ?"). Les questions du questionnaire PBPI étant déjà sous forme d'énoncé, nous ne les avons pas modifiées et les avons utilisées telles quelles.

- **Dériver la réponse** Nous utilisons RoBERTa large (Liu et al., 2019a)[7] affiné sur le jeu de données MNLI (Williams et al., 2018) pour déterminer la relation d'implication. Nous dérivons ensuite la réponse à partir de la relation d'implication entre le dialogue et la question comme suit.

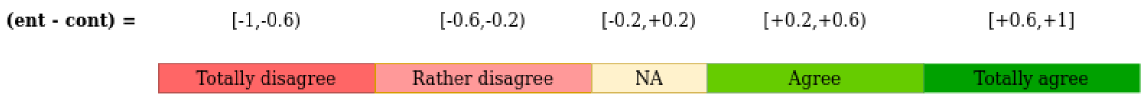


Figure 4: Faire correspondre les scores NLI aux types de réponses ALS

Nous envisageons différentes manières de transformer et de sélectionner le contenu des dialogues. Nous étudions également l'impact du modèle NLI utilisé, en comparant DeBERTa, le modèle utilisé dans Ghassemi Toudeshki et al. (2021), avec RoBERTa (Liu et al., 2019a), le modèle utilisé dans notre approche.

Transformation du contenu

- **Null Transformation** (CT_{null}) Une ligne de base de transformation nulle où nous concaténons simplement les tours du dialogue d'entrée. Pour encoder les informations sur le locuteur dans chaque tour, l'énoncé est accompagné du rôle du locuteur (patient/bot) au début.

- **Summary** (CT_{sum}) Les paires de tours adjacents sont résumées, et les résumés résultants sont concaténés. De cette façon, le dialogue d'entrée est transformé en une séquence de résumés de deux tours. Nous avons également essayé de résumer l'ensemble du dialogue

[7] https://huggingface.co/roberta-large-mnli

en une seule fois, mais nous avons constaté que l'application du résumé sur chaque paire de tours plutôt que sur l'ensemble du dialogue donne de meilleurs résultats. Nous utilisons le modèle **BART-large**[8] (Lewis et al., 2020) affiné sur le corpus de résumé de nouvelles XSUM (Narayan et al., 2018) et sur le corpus de résumé de dialogues SAMSum (Gliwa et al., 2019).

- **Long Answers** (CT_{answer}) Dans un dialogue de recherche d'information, les tours adjacents sont souvent des paires question-réponse. Sur la base de cette observation, nous transformons chaque paire de tours adjacents dans le dialogue en une seule phrase déclarative en supposant que le premier tour est une question (par exemple, " Quel médicament avez-vous pris ? "), le second est une courte réponse à cette question (par exemple, " Doliprane ") et la phrase dérivée de la transformation est une longue réponse à la question (par exemple, " J'ai pris du Doliprane "). Ce modèle affiné a été appliqué à chacun des deux tours suivants des dialogues d'entrée, et les phrases déclaratives résultantes ont ensuite été concaténées pour former la transformation déclarative du dialogue entier.

Sélection du contenu

- **Sélection de contenu nul** (CS_{null}) Une ligne de base de sélection de contenu nul où la prémisse est la concaténation de toutes les unités d'entrée produites par les opérations de transformation du contenu (tours de dialogue, séquence de résumés de deux tours, séquence de réponses complètes).

- **Unit-Based** (CS_{units}) Chaque question est évaluée par rapport à chaque élément d'entrée. Étant donné une séquence d'entrée I_n de longueur n, la réponse a_i à une question q est ensuite déterminée en agrégeant les probabilités d'implication résultantes.

- **Similarité** (CS_{sim}) Pour chaque question q, nous sélectionnons un sous-ensemble d'unités d'entrée qui sont sémantiquement similaires à q. Nous codons la question et les unités d'entrée en utilisant SBERT[9] (Reimers and Gurevych, 2019) et calculons similarité cosinus pour chaque paire (q, unité d'entrée). Nous sélectionnons ensuite les éléments dont le score de similarité est supérieur à 0,5, nous les concaténons et nous utilisons le résultat comme prémisse de la NLI.

- **NLI** (CS_{nli}) Pour chaque question q du questionnaire, nous sélectionnons les unités d'entrée qui sont liées à q en utilisant le modèle NLI (RoBERTa-Large). Plus précisément, nous sélectionnons les phrases qui ont un score d'implication ou de contradiction supérieur à 0,5. Toutes les phrases sélectionnées sont ensuite concaténées pour former la prémisse NLI.

[8] https://huggingface.co/Salesforce/bart-large-xsum-samsum
[9] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2

Introduction

Introduction

Dialogue systems have recently started to emerge in the healthcare sector. These Health-Bots are used, for instance, to provide health-related information, to help the user to set up an appointment or to give a likely diagnosis based on a list of symptoms, or to provide with specific answers (improved FAQ). Those dialogue agents are often brittle pattern-matching systems or implement a pre-defined and stable dialogue tree. They usually cover a very specific domain, and they are generally unable to maintain a sustainable conversation such as typically occurs, for instance, between a physician and a patient.

Being able to create such interactions beside the medical visits could be especially important to accompany patients with chronic conditions. Studies[10] have shown that it is important to understand the quality of life of the patients from their own perspective. Collecting information that is peripheral to the required topics can give a more holistic understanding of a patient's context and help to understand how lifestyle choices are influencing pathology dynamics. It's a very challenging topic and hasn't been addressed yet. In this work, we would like to make a few first steps in this direction.

In between visits, the traditional way to measure the current state of patients is standardized questionnaires. The limits of such questionnaires are their lack of personalization, their repetitiveness, and their one way perspective (the physician's one). Newer technologies that drive two-way communication are more appropriate to get more relevant assessments and follow-up of patients, and to better personalize the care eventually. In particular, health-bots are rapidly gaining traction as a key technology which would permit complementing the current standardized questionnaires traditionally used to collect patient information with a more informal and customized interaction, replicating and expanding patient-physician dialogue into a new type of relationship.

This thesis benefited from a CIFRE contract and the health-bot models we developed aimed at answering a key goal of the ALIAE[11] company, namely:

> How to develop health-bots which can complement the standard questionnaires used in patient care program and be used to collect both specific and serendipitous information about the patient condition ?

Our long-term goal is to develop a human-machine dialogue system that would complement standard clinical questionnaires traditionally used in clinical studies or patient care programs by regularly engaging the patient in a dialogue about the questionnaire topics. Since our target users are chronic patients, it is more important to keep them engaged for a long period to get evolution of their quality of life though time rather than getting all information right away though the first conversation. Though, in this work, we start with the first steps: developing a dialogue system for the first-time interaction and filling in the questionnaire based on this first conversation.

It's also important to point out that our system is not dedicated to replace the doctor, but rather complement. In this work, we are not aiming to reproduce doctor behaviour, but rather to enrich it on the aspects of quality of life connected with pathology. Here are the key differences that we are going to address in this thesis: topic switch flexibility, opportunity to talk beyond medical subjects, and avoiding repetitive questionnaire statements.

The approach we use to collect information is to give freedom to the user to talk about something that actually bothers her or him. Casual chitchat conversation can be a natural glue

[10] https://www.fda.gov/media/117890/download

[11] aliae.io

between subjects required in the questionnaire, and an opportunity to discover the context of the patient's life and to create a relationship with him or her. It can help us to receive additional information about the patient that does not fit predefined medical tree.

As soon as the dialogue system is up and running, it is necessary to automatically extract from these dialogues patient information that is relevant to its treatment, but that would not necessarily be detected in the standardized questionnaire approach. Also, instead of giving the same questionnaire statement to answer by the user, we develop the dialogue system in such a way that it is possible to collect the necessary context in order to derive the answers to these statements in implicit and automated way.

This thesis proposes the complementary pipeline to usual doctor-patient routine with a dialogue system as a base actor.

Chronic pathologies are impacted by the environment of the patients, their quality of life, their social life, and mental state. In order to better personalize the care, it is more and more important to understand the person in her/his environment, analyse his/her state, emotions, behaviours, and their special interactions.

Thesis Outline

Part I (Background) introduces the field of Dialogue Systems. In Chapter 1, we discuss how dialogue systems can differ on the purpose (task-based, chit-chat or their combination) and provide the ways they are used for healthcare and medical applications. In Section 2, we take a closer look at their modular and end-to-end architectures. In Section 3, we make an overview of basic neural networks models for retrieval and generative dialogue systems. Finally, in Section 4 we provide different approaches for human and automatic dialogue evaluation.

Part II focuses on developing dialogue system for patients.

First, in Chapter 1 we propose a dialogue model that introduces flexibility between pre-defined medical tree and topics introduces by the user. The key research questions addressed in this chapter are:

> RQ1: How to create training and test dialogue data for the health domain ?

> RQ2: Which neural architecture is best suited to create a health bot from this data?

Since absence of dialogue data is the main bottleneck for training a response generation model, we introduced data augmentation approach based on medical decision tree and forum sentences and test retrieval and generative models. Based on this data, we trained both retrieval and generative dialogue models. The evaluation indicated that given freedom, the users tend to use small talk and to deviate from proposed medical topics.

Chapter 2 addresses that issue, focusing on the following research question:

> RQ3: How do we design a health bot that collects both specific medical information and serendipitous information about the patient general state ?

We introduce ComBot, an ensemble of three dialogue models: MedBot, dedicated to follow medical decision tree in a flexible way (See Chapter 1), FollowUpBot, dedicated to chat and ask follow-up question on healthcare topics even if they are not directly related to medical decision tree, and EmpathyBot, a chit-chat model that is dedicated to provide compassion and keep up the conversation on any other topic which not related to health.

Part III proposes different approaches to automatically fill the questionnaires based on dialogues obtained with ComBot. In this part, our research questions are:

RQ4: Which NLP method is best suited to extract relevant medical information from human-healthbot dialogue ?

RQ5: How much does dialogue pre-processing help improve information extraction?

First, in Chapter 1 we do a clinical questionnaires' categorization based on questions and answers types and propose a data collection procedure for each of them. Later, we explore how automatic questionnaire filling could be resolved as different tasks: question answering, text classification and natural language inference. Also, we provide a detailed analysis of their performance with respect to each question and answer type.

Second, in Chapter 2 we investigate how content selection and content preprocessing can influence the performance of natural language inference model in zero shot setting for automatic questionnaire filling.

Publications

Parts of this thesis have appeared in the following publications and reports:

- Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning Health-Bots from Training Data that was Automatically Created using Paraphrase Detection and Expert Knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 638–648, Barcelona, Spain (Online). International Committee on Computational Linguistics.

- Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2021. Gathering Information and Engaging the User ComBot: A Task-Based, Serendipitous Dialog Model for Patient-Doctor Interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, Online. Association for Computational Linguistics.

- Farnaz Ghassemi Toudeshki, Philippe Jolivet, Alexandre Durand-Salmon, and Anna Liednikova. 2021. Zero-Shot Clinical Questionnaire Filling From Human-Machine Interactions. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Farnaz Ghassemi Toudeshki, Anna Liednikova, Philippe Jolivet, and Claire Gardent. 2022. Exploring the Influence of Dialog Input Format for Unsupervised Clinical Questionnaire Filling. In *Proceedings of The 13th International Workshop on Health Text Mining and Information Analysis*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Part I

Background

Chapter 1

Chatbot purposes

Contents

> Dialogue is a way of taking the energy of our differences and channeling it toward something that has never been created before. ... In essence, a dialogue is a flow of meaning. ... In the most ancient meaning of the word, logos meant 'to gather together', and suggested an intimate awareness of the relationships among things in the natural world.
> "Dialogue and the Art of Thinking Together"
> William Isaacs

In this chapter, we briefly survey the evolution of dialogue systems in Natural Language Processing (NLP) and in the health domain.

The technology of dialogue systems, started in the 1960s with the aim of seeing if the users would believe that they were talking to a real human instead of a machine. However, chatbot systems are built to mimic human conversation not only for entertainment, but also for providing some information and helping in decision-making. Especially, they could be successfully applied to some restricted and well predictable contexts. Today, most of the small-scale commercial dialogue systems are hard-coded to support a very limited functionality fitted to a specific use case, such as make a purchase or check the weather forecast (Jacques et al., 2019; Jain et al., 2018). Zadrozny et al. (2000) expressed that the best way to facilitate human computer interaction is by allowing users "to express their interest, wishes, or queries directly and naturally, by speaking, typing, and pointing". This idea became a driver behind the development of conversational agents (Shawar and Atwell, 2007). The research still continues to move forward the frontiers of situations when the dialogue systems could be successfully applied. With advances in natural language generation technologies, conversational agents experience a surge of interest from both industry and academia.

Based on dialogue system application, we can consider their two major types: chit-chat dialogues and task-oriented dialogues.

1.1 Chit-chat

The idea of an automated system communicating through a natural language goes back to Alan Turing's seminal Imitation Game (Turing, 1950), which introduced the so-called Turing Test in a form of a human-machine dialogue. This test challenges people to determine whether they are talking to another human or to a machine.

The history of chatbots began in 1966 with a rule-based and parser-based system, ELIZA (Weizenbaum, 1966), that mimicked a psychotherapy interview, and PARRY (Colby et al., 1971), that simulated a schizophrenic patient. While these systems were based on simple pattern matching and substitutions, they succeeded to create an illusion of understanding between humans and machines.

However, usage of rules can be successfully applied only in a very restricted context. The next stage of chatbots development has been driven by training open-domain chit-chat models with dialogue examples from movie scripts like OpenSubtitles (Tiedemann, 2009) or from forum-like post-reply pairs such as Twitter dialogue corpus (Ritter et al., 2011), Chinese Weibo dataset (Wang et al., 2013) and Reddit Forum posts (Al-Rfou et al., 2016). However, these datasets are noisy and sometimes do not reflect realistic conversations, so Li et al. (2017b) proposed a new high-quality multi-turn dialogue dataset, DailyDialogues, with manually labelled communication intention and emotion information.

Although in the beginning this kind of dialogue systems was used predominantly for entertainment (Ram et al., 2018), with the development of social skills in chatbots, the idea of replacing a human companion with a conversational agent such as Replika (Fedorenko et al., 2017) started to emerge. However, most of these systems are still limited for real world application, though they might perform well in some particular conditions, like maintaining a coherent and engaging conversation for 20 minutes to pass the Alexa Prize challenge (Hakkani-Tur, 2018). The way of evaluating dialogue systems has also changed: instead of guessing if the system is a bot or a human as in the Turing Test, now the users are asked to evaluate the system by some specific criteria, knowing from the very beginning that they are talking to a machine.

To maintain a meaningful and long conversation with a human, a chatbot should be capable to imitate human dialogue in all its facets: empathy, knowledge, personality, etc. There are different datasets that address each of these facets. Personalized dialogue generation started with the PERSONA-CHAT dataset (Zhang et al., 2018b), empathetic and compassionate response generation was firstly introduced with the Empathetic Dialogues dataset (Rashkin et al., 2019). Holding conversation on different topics requires from a dialogue model to have a certain amount of world knowledge and common-sense reasoning that is partially addressed with Wizard of Wikipedia dataset (Dinan et al., 2018). Later, personality, knowledge-grounding and empathy were combined all together in Blended Skill Talk dataset (Smith et al., 2020) and used for both retrieval and generative chit-chat models (Roller et al., 2021).

While emotional intelligence in conversational agents has been widely explored, there is still a lack of explicit modelling of values based on social domain theory. Qiu et al. (2021) proposed another stage of development by incorporating a value model, VALUENET. It is the first large-scale text dataset, organized in ten dimensions, that conform to the basic human value theory in intercultural research. Firstly, they created a knowledge base of human values and identified scenarios related to these values. Then, they ask annotators to decide whether these scenarios might affect someone's value in a positive or negative way. The idea comes from the assumption that values underlie our attitudes and guide our way of evaluating the events.

Even now, having human-like conversation is still regarded as an ambitious goal and drives the development of systems capable of holding an open-domain (chit-chat) conversation through the

Figure 1.1: A cherry-picked example of a conversation between a crowdworker (left speaker) and the Generative BST 2.7B model (right speaker) (Roller et al., 2021)

incorporation of better communicative skills like perception of natural language, constructing intelligent decisions as well as producing proper replies (Jafari et al., 2020).

1.2 Task-oriented

While open-domain chatbots seek to maximize user engagement and keep the conversation going, task-oriented dialogue systems are targetting at accomplishing some specific task as efficiently as possible, saving people from tedious work.

These dialogue systems are usually characterized by a clearly defined and measurable goal, a structured dialogue behaviour and a closed domain to work on. Usually, the task involves retrieving some information from the knowledge base or/and the user in order to perform some specific action based on this information.

Classical examples of the tasks are in the restaurant reservation, trip planning by booking flights tickets or hotels, finding public transportation routes between locations or making an appointment (Williams et al., 2016; Zhao et al., 2017a). These dialogue systems are dedicated to replace the process of navigating through menus and user interfaces by conversation.

Task-oriented dialogues usually rely on a domain specification, which can be defined in terms of an ontology, a table or a set of annotated dialogue samples that provide a frame for linking slots and intents to possible replies. Such an ontology should enumerate all concepts and attributes (slots) that a user can specify or request information for (Mrkšić et al., 2017). The dialogue models are then designed to jointly perform the tasks of parsing the input utterances, matching or filling the slots and tracking the belief state of the dialogue (Williams et al., 2016).

Though these dialogue systems may be limited in their conversational capabilities to respond to arbitrary utterances, they are very robust at executing task-specific commands and requirements. However, sometimes to take the best from two worlds, task-oriented agents are combined with their chitchatting counterparts.

1.3 Chatbot ensembles

D'Haro et al. (2020) argued that real human-to-human dialogues often combines both task-oriented and chitchat-oriented communication strategies, rather than using only one. Some studies (Yu et al., 2017; Sun et al., 2021) showed that the user is more likely to engage with a task-oriented dialogue agent if it sounds more natural and handles out of domain responses well. Recently, an emphasis has been put on integrating the two types of agents to model different conversation approaches and benefit the positive aspects of both types, like the robust abilities of task-oriented dialogue systems to perform tasks and the human-like chattiness of open-domain chatbots (Zhao et al., 2017a; Yu et al., 2017; Serban et al., 2017). For example, Yu et al. (2017) introduces a dialogue model which interleaves both bots with a help of a statistical reinforcement learning policy that selects a response among social and task-oriented candidates that will lead to natural and engaging interactions.

Conversely, Gunson et al. (2020) showed that success of inclusion of open-domain chat depends on the context and that in a public "walk-up" setting or workplace, the users haven't shown significant preference for combined system over purely task-oriented bot. Actually, the majority of participants did not engage in social conversation with the system, some of them didn't feel comfortable making small talk in a working building, though some of them liked this option but preferred to decrease its ratio. Gunson et al. (2020)'s work showed how different human-machine interactions can be in more formal context in comparison with in-home and more private "companion/assistant" case.

A dialogue system can also be an ensemble of several conversational agents. Humans lose interest in communication if they don't feel understood. It might happen when a chatbot is unable to interpret a user's text accurately that results in giving absurd responses (Abedin et al., 2021). So having a chatbot for each specific topic or representing some particular skill in an ensemble could be a safe solution to make the system cover most of the cases. Such kind of ensembles were boosted and promoted by Amazon Alexa Challenge (Khatri et al., 2018). Digital assistants embedded in smartphones and smart home devices, such as Siri, Alexa and Cortana (by Statista Research Department and 17, 2022) play the role of a hub that can serve different chatbots (skills) from the same interface.

One of the examples from this challenge is the ALANA ensemble model (Papaioannou et al., 2017b,a), developed for open domain chitchat that combines domain-specific bots used to provide information from different sources with social bots to smooth the interactions by asking for clarification, expressing personal views or handling profanities.

While the previously mentioned approaches model different interlocutor explicitly, Zhao et al. (2022) proposed a unified dialogue system (UniDS) framework based on pre-trained language model DialoGPT (Zhang et al., 2020b) to handle chit-chat and task-oriented strategies in an end-to-end manner with the help of a unified dialogue data schema: belief state, database result and system act.

The components of the combined dialogue system and their design in terms of explicitly and interpretability may be very different from one domain to another depending on the applicative and the performance requirements.

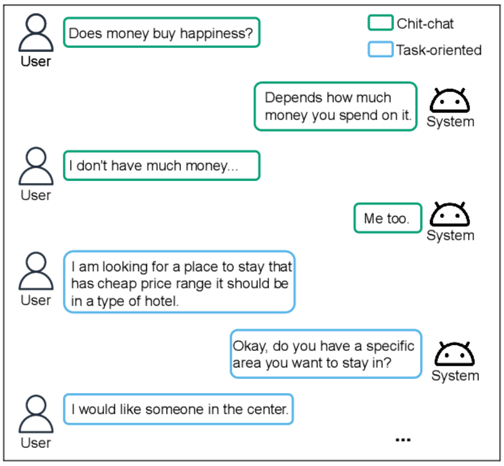


Figure 1.2: Illustration of users being interested to chit-chat with the dialogue system before booking a hotel (Zhao et al., 2022)

1.4 Chatbots in healthcare

A healthcare dialogue system can be developed both for doctors and patients. In the first case, the main goal is educating the juniors (Campillos Llanos et al., 2015) or automate doctor routine like auto-transcribing the conversations and extracting the valuable information from it (Yim and Yetisgen, 2021). In the second case, the dialogue systems can propose the patient with ongoing and highly personalized conversations about their health and well-being, aiming to meet their specific needs and preferences.

Furthermore, conversational agents being available 24 hours/7 days a week from any location can help to balance proposal and demand by relieving some pressure from the healthcare system. On the one hand, health services can delegate and automate some processes before actual decision-making, like administrative tasks (Palanica et al., 2019) or assistance to physicians in basic interaction with patients (Ni et al., 2017). On the other hand, patients can receive immediate health service information and support from a chatbot without waiting for a dedicated appointment with a health provider unless it is really needed.

Conversational agents can also offer the safety of anonymity. Users will sometimes feel more comfortable interacting with an "anonymous" conversation agent compared to a healthcare professional (Meadows et al., 2020), for example, in the case of post-traumatic stress disorder (Lucas et al., 2017).

Car et al. (2020b) showed a shifting trend in conversational agents' goal from monitoring and tracking to connecting patients to healthcare services, supporting them in achieving long-term behaviour changes and chronic disease management. In clinical psychology, embodied conversational agents were used for social skills training, cognitive behaviour therapy, counselling for depression and anxiety, or autism treatment (Provoost et al., 2017). A prominent example is Woebot (WoeBot, 2022) that applies cognitive behavioural therapy for users with anxiety and depression.

Car et al. (2020b) identified five distinct themes in terms of conversational agent content: *treatment and monitoring* (treatment implementation, management, adherence, support, and monitoring), *health service support* (connecting patients to health care services), *education* (providing health care–related information), *lifestyle behaviour change* (supporting users in tackling various modifiable health risk factors), and *diagnosis* (identification of the nature of a disease or a condition).

In this work, we focus on data collection (i.e. monitoring) and giving advice (i.e. education).

Data Collecting information The screening agents control the conversation flow through structured questions and scripts with symptom checklists or scales to assess the risk of diseases, such as cancers, mental health disorders and risk of chronic diseases (Spänig et al., 2019; Ghosh et al., 2018).

Most of the medical conversation systems collect information for triage or diagnosis (Middleton et al., 2016; Agrawal et al., 2017; Xu et al., 2019). In this case, dialogue flow is close to decision tree, when the following question is chosen to narrow the conversation and in such a way that it should better differentiate between final states. The fewer questions, the more efficient the dialogue system. Examples of such medical trees could be found in American Medical Association Family Guide (Kunz, 1982), part of which was implemented by Fischer and Lam (2016).

On the other hand, monitoring a chronic condition may be more efficient in the opposite way by expanding the context of existing symptoms to get the full picture of a patient's quality of life. In this thesis, we will address this strategy.

Effective patient monitoring includes continuous, unobtrusive data collection with explicit check-ins by mimicking human interaction between healthcare providers and patients. For example, the Babylon chatbot (Middleton et al., 2016) offer AI-empowered triage based on users' self-reported symptoms, personal medical history and a medical knowledge database reviewed by clinical experts. The chatbot triages users and provide them with an action plan to link further appointment-making with general practitioners or specialists.

Conversational agents could provide consultations to empower patients to engage more fully before and during the clinical appointment. By exchanging information and providing live responses, conversational agents may collect patients' information before their appointment to provide tailored counselling and improve the quality of visits (Adams WG, 2014).

Giving advice The purpose of training and education is achieved mainly by providing information to users via conversations and coaching users to acquire skills. Moreover, it's something that users may expect from the healthcare bot in exchange for provided information.

> "I expected to get some advice from the chatbot, but it didn't. I was a little disappointed because I felt the chatbot did not care what I shared."

As illustrated by this comment issued by a patient during an experiment by (Lee et al., 2020), feedback is something highly expected by users talking to a healthcare bot. This is highly correlated with DeCapua and Dunham (1993)'s finding on how advice offers support by helping patients clarify their problem and finding possible solutions.

Much work has been done on advice giving in the context of recommendation systems. Wiesner and Pfeifer (2014) focuses on providing patients with relevant and trustworthy information, while Farrell et al. (2012) targets recommendations that are actionable. Some work has focused on extracting relevant health information from gray literature (Turoff and Hiltz, 2008), health social networks (Song and Marsh, 2012), YouTube videos, expert findings and personal health records (Rivero-Rodriguez et al., 2013; Ati et al., 2015). Other work has proposed recommender systems to help with diagnosis (Thong et al., 2015; Pattaraintakorn et al., 2007; Lafta et al., 2015).

Going further, Pereira and Díaz (2019) reviewed studies that used health chatbots for behaviour change to promote patients' self-management and discourage harmful practices in chronic conditions, such as asthma (Kadariya et al., 2019) and diabetes (Gong et al., 2020), mental and neurological disorders and addictions. They emphasized the role of chatbots in monitoring biosignals and mood, coaching and advising users by improving their self-efficacy and cognition through distraction and encouragement.

Healthcare chatbot challenges In medical and healthcare domain till recent times, dialogue systems were basically task-oriented conversational agents (Car et al., 2020b). One of the few exceptions is a combination of medical advice model and the compassion model trained on Facebook AI Empathetic Dialogue dataset (Rashkin et al., 2019) to maintain a conversation in order to help people with major depression (Harilal et al., 2020).

Another interesting issue to address is that the developed chatbot should be stable and welcoming enough, so that the user won't be influenced by the second-hand anxiety from the chatbots (Dimitroff et al., 2017). Shan et al. (2022) argue that the chatbots' mental health should be assessed before releasing the chatbots online to avoid negative psychological impact on users. Their study establish the corresponding assessment dimensions for chatbots for depression, anxiety, alcohol addiction, and empathy.

Furthermore, using a bot in the psychological field needs to be even more accurate and human-like by giving the vibe of human-connection and empathy. That's why, even today, most users prefer human agents to solve their problems. For a virtual agent to respond like a human, it needs to understand the problem of the user and provide sensible replies to humans.

Chapter 2

Chatbot Architectures

Contents

Before conversational data became available, early work on dialogue systems were mostly either rule-based (Weizenbaum, 1966) or learning-based methods (Schatzmann et al., 2006; Misu et al., 2012). Rule-based systems don't require any data and largely rely on manual effort in rules designing. Matching response to the context based on keywords or their combinations is one of early simple approaches' example. Learning-based systems rely on automatic training of a model (e.g., reinforcement learning) with little data. Both approaches require much hand-crafted designing that makes such dialogue systems unsatisfactory in coverage and difficult to be extended to other domains or tasks.

With the emergence of datasets, dialogue systems architectures have been divided into another two types: 1) modular or pipeline architecture, when an input message passes through a sequence of separate modules each trained for a specific subtask in order to obtain a final text response, 2) end-to-end architecture, when a single model is trained to output text response directly from input textual message. Most real-world commercial system benefit interpretability and stability of modular systems. Though, their creation requires large-scale datasets annotated to train each component separately. End-to-end systems require less effort to be built, though being black-box give risks of uncontrollable behaviour (Gao et al., 2018).

In this chapter, we review three key pillars of today's dialogue models: the modular or pipeline architecture mostly used in pre-neural NLP, the end-to-end models common in neural dialogue modelling and some datasets commonly used for training conversational agents and dialogue models.

2.1 Modular Dialogue System

Traditionally, modular dialogue systems consist of three units: Natural Language Understanding (NLU), Dialogue Manager (DM) and Natural Language Generation (NLG).

Natural Language Understanding module processes the raw user input and extracts useful information and features from the dialogue. It usually performs different type of utterance classification and information extraction. The most common utterance classification tasks are intent classification (Gao et al., 2018), dialogue act classification (Jurafsky and Shriberg, 1997), domain identification. The most common information extraction task is slot filling (Gorin et al., 1997). For this type of task, it's common to use sequence labelling methods such as Conditional Random Field (CRF) (Hahn et al., 2011) or recurrent neural network, typically bi-LSTM with CRF layer (Yao et al., 2014; Mesnil et al., 2015). Some of these tasks could be performed jointly to boost the performance, for example, putting together intent classification and entity extraction (Guo et al., 2014; Zhang and Wang, 2016; Bunk et al., 2020) or sentiment and dialogue act classification (Cerisara et al., 2018).

Dialogue Manager These features can be used by the **Dialogue Manager** (Young, 2002) which defines the content of the next utterance and thus the behaviour of the dialogue system. There are three main types of dialogue management: finite-state-based, frames-based or agent-based.

The first dialogue managers were *finite-state-based models* (Sutton et al., 1998), that allowed to define a dialogue flow as a sequence of predefined states (questions of the system) and transitions between them (possible user's answers). Being rigid, inflexible and machine-initiated, they are still relatively straightforward, predictable and can be efficient when the domain and the task are simple. For example, finite-state model can be successfully applied for a form-filling task or guiding the conversation through a predefined decision tree, when yes/no or short answers are expected with limited alternatives at branch points. Such a flow looks similar to a patient-doctor conversation, when a doctor asks precise questions and classify patient answers to ask another appropriate question or to state a diagnosis or to make a prescription (Goddeau et al., 1996).

Frame-based dialogue management (Nestorovic, 2009) controls the dialogue strategy by keeping track of the interaction history within topic frames. Each of these frames is defined as several slots for related pieces of information. So, instead of representing what the system has to do and when, it relied on schemas specifying what the system has to solve. Such kind of dialogue management can then be cast as a slot-filling problem.

So it's simple to implement state-based dialogue management, but it is rigid and limited during runtime. Frame-based management resolves inflexibility, but dialogue flow optimization is left as an open problem (Nestorovic, 2011). *Agent-based dialogue systems* (Nguyen and Wobcke, 2005) succeeded to overcome these limitations by enabling complex communication between the system, the user and the specific applications like knowledge base or other external resources. Basically, this dialogue manager performs two tasks. On the one hand, it keeps track of the conversation's context by updating internal belief states to conduct a coherent dialogue with a user, which is referred to as a *Dialogue State Tracking* task (Young et al., 2010; Williams et al., 2016). On the other hand, it learns to select the next system actions (e.g. retrieving data from a database, asking for missing information, etc.) based on accumulated dialogue state and coordinates smaller modules performing these selected actions, which is called *Dialogue Policy Learning* (Mohan and Li, 2009). It is also possible to combine both tasks into one deep reinforcement learning model based on end-to-end framework when both tasks are learned jointly (Zhao and Eskenazi, 2016).

Natural Language Generation module Stent and Bangalore (2014) acts inversely to the natural language understanding one: it receives features and information from the Dialogue Manager to generate a response that, finally, will be presented to the user. Initially, this module was based on templates, though recently it is mostly generation or retrieval model.

Template-based systems map non-linguistic input directly to the linguistic surface structure (Reiter and Dale, 1997), for example, through replacing gaps (van Deemter et al., 2005). In the dialogue system, information extracted by natural language understanding module is used by dialogue manager to decide how to fill gaps in the templates to give to the user a well-formed response. Template-based systems are a cost-effective solution in the early stages of prototyping (Galley et al., 2001), when limited variability is required in the system responses. Also, templates are usually more intuitive for domain experts than complex mechanisms (Reiter and Dale, 1997). However, they are difficult to maintain and update (Reiter and Dale, 1997) as the system grows, since a number of templates needed to cover all necessary cases might become unreasonably large (Galley et al., 2001).

Recent natural language generation modules can be mainly represented as a retrieval-based (also called response selection) model (Wang et al., 2013), when human-to-human conversation datasets are mined to copy a response from a similar context, or a generative-based (Ritter et al., 2011) model, when an encoder-decoder system generates a response from a user utterance token-by-token.

Building a modular system requires costly creation of data annotated for each module training (Wen et al., 2017). It also means that changes in domain or scaling to the new ones leads not only to collecting new data, but also to redesigning and retraining of corresponding modules (Bordes and Weston, 2017), as well as to adaptation of the subsequent ones (Zhao and Eskenazi, 2016). Interdependence of the pipelined modules also leads to propagation of errors (Li et al., 2017a; Liu et al., 2018).

2.2 End-to-End models

The idea behind end-to-end models is in simplifying complexity of modular dialogue systems and avoiding numerous data annotation for each unit. The first end-to-end chatbots were chitchat-oriented due to the emergence of social media websites. Later, these approaches have been applied for task-oriented models to party or fully replace modules by a single end-to-end model.

End-to-End Non-Task-Oriented Dialogue Systems The increase of conversational exchanges available on social media websites raised the perspectives of training the chatterbot models that could converse in the similar manner without any hand-coding. Ritter et al. (2011) proposed a first single-turn dialogue response generation system trained on their Twitter post-reply dataset. Later, Sordoni et al. (2015) proposed an embedding-based model to encode semantic and syntactic similarity for context-sensitive response generation with capturing long-span dependencies. Sutskever et al. (2014) introduced the idea of sequence-to-sequence learning, when input text is used to generate a response text of variable length. It has inspired several efforts to build end-to-end trainable non-task-oriented conversational systems (Vinyals and Le, 2015; Shang et al., 2015; Serban et al., 2015).

Though such end-to-end chatbots were producing the responses that are conversationally appropriate, they were still often bland (Li et al., 2016; Gao et al., 2019) and lack some capacity of task-oriented systems like grounding in the real world and access to external knowledge (D'Haro et al., 2020), usage of entities and factual content in response generation (D'Haro et al., 2020) or interaction with databases (Sukhbaatar et al., 2015) performed by different modules. Pretty often, this capacity comes at a cost of significant hand-coding specific to some particular domain.

End-to-End Task-Oriented Dialogue Systems constituting all modules in one model (Serban et al., 2016) have become one of the major research topics with the rise of deep learning models (Schmidhuber, 2015) and companies' interest in automating some tasks through dialogue systems. A single deep network was trained to reproduce conversations from a large dataset by generating an answer directly from raw user input, combining inside the tasks of natural language understanding and generation as well as dialogue management.

Wen et al. (2017) proposed as a dialogue sequence-to-sequence mapping with a help of a modularized end-to-end model. Each component is a separate neural network that makes the whole model differentiable, but still modularly connected. The model is end-to-end trainable: each system module is trainable from data, except for a database operator. It does not directly model the user goal, but has a distributed representation of his or her intents in a form of dialogue acts. It has an explicit representation of database attributes in the form of slot-value pairs, which it uses to achieve a high task success rate.

Bordes and Weston (2017) formalized the task-oriented dialogue as a reading comprehension task by regarding the dialogue history as context, user utterance as the question, and system response as the answer. In this work, they utilized end-to-end memory networks for multi-turn inference. Madotto et al. (2018) took a similar approach and further feed the knowledge base information into the memory networks (Weston et al., 2015; Sukhbaatar et al., 2015).

Lei et al. (2018) proposed a two-step sequence-to-sequence generation model which bypassed the structured dialogue act representation, and only retain the dialogue state representation. In their method, the model first encodes the dialogue history and then generates a dialogue state using LSTM (Hochreiter and Schmidhuber, 1997) and CopyNet (Gu et al., 2016) and then generates the final natural language response given this state.

Andreas et al. (2021) found that BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) outperform GPT-based models and achieve state-of-the-art performance among other task-oriented dialogue systems.

The Seventh Dialogue System Technology Challenge (DSTC7) (D'Haro et al., 2020) focused on developing and pushing state-of-the-art of end-to-end dialogue systems. One of the tracks is an end-to-end response selection task when a model needs to select the correct next utterances from a set of possible answer candidates for the multi-turn context, or detecting when none candidate was suitable. Another track provides a test bed for a knowledge-grounded response generation task.

2.3 Dialogue Data Creation

Discussed works have relied on already available dialogue data or on question/answer pairs extracted from online forums (Wei et al., 2018; Lin et al., 2019a; Xu et al., 2019). Starting a new task or application faces us with a problem of absence of training data. In some domains, like healthcare one, such data is extremely scarce and difficult to obtain. When obtainable, it also requires extensive pre-processing due to anonymization constraints. Various methods have been proposed to facilitate the creation of training dialogue data. Normally, existing dialogue datasets can be split into real (written or edited by humans) or synthetic (generated by models) ones.

Natural Human-Human Dialogues Real dialogue datasets are commonly collected and free-form annotated thanks to crowdsourcing. The task can be defined as a Wizard-of-Oz setup (Wen et al., 2017; El Asri et al., 2017a), when two humans interact based on some pre-defined scenario and the resulting dialogues are collected (Green and Wei-Haas, 1985; El Asri et al., 2017a), or by asking crowdsourcing workers to provide continuations to incomplete dialogues (Wen et al., 2017). Both approaches are time intensive and expensive. Moreover, human-human dialogues collected by both approaches may be very different from the human-machine interactions that complicates learning to support efficient human-machine communication where, typically, chat messages are restricted in length.

Synthetic Machine-Machine Dialogues Another line of research is dedicated to acquire data through machine-machine simulations (Xu et al., 2019; Majumdar et al., 2019; Shah et al., 2018a). In particular, Majumdar et al. (2019) combines pre-defined dialogue outlines with template-based verbalization of dialogue turns to automatically create a synthetic dialogue corpus.

For generating synthetic dialogue datasets, one models user behaviour simulator based on a compact representation of the user goal and a stack-like user agenda (Schatzmann et al., 2007). This idea was evolved into dialogue self-play (Shah et al., 2018b), when two or more conversational agents interact by choosing discrete conversational actions to exhaustively generate dialogue histories for a task-oriented system. They presented an agenda-based user simulator agent and a finite state machine-based system agent for the self-play step. In medical domain, Liu et al. (2019b) constructed a simulated nurse-patient dialogue dataset to bootstrap such a screening bot prototype. Later, Majumdar et al. (2019) proposed to combine self-play with Memory Networks (Weston et al., 2015) for low-resource domains data generation by challenging conversational agents to perform on unseen patterns.

Chapter 3

Neural Dialogue Models

Contents

The dialogue response methods can be divided into three main types: retrieval, generative and hybrid retrieve-and-refine approaches which combine both.

Retrieval models select the most suitable response by ranking the candidates from the conversational dataset given a context based on various matching features. Since performance of retrieval model heavily rely on available data, the responses are always "as it is" and are hard to customize for the particular text or requirement from the task, like style or attitude.

Generative models usually they follow encoder-decoder architecture that firstly represent the context with the encoder and later form a response utterance token-by-token based on received representation with the decoder.

While retrieval model propose reliable, well-formed and informative responses, generative ones are able to deal with new dialogue flows thanks to generalization abilities. *Hybrid models* represent the class of the models that tend to combine the advantages of both retrieval and generative models. Early back, Jafarpour (2010) expressed the idea of generating a better response based on top ranked retrieved responses. Later, Yang et al. (2019) proposed a hybrid neural conversational model that consist of three modules: generative, retrieval and a hybrid ranking one. The generation module uses a sequence-to-sequence model to create a response candidate given a dialogue context. The retrieval module selects a set of response candidates from the historical context-response database by matching context to context. The hybrid ranking module is built on the top of neural ranking models to select the best response among retrieved and generated candidates. The integration of neural ranking models, which can learn representations and matching features for conversation context-response candidate pairs, enables us to minimize feature engineering costs during model development.

Hybrid models are out of the scope of this thesis. Instead, we focus on main architectures of generative and retrieval models. In what follows, we summarize the most prominent approaches proposed for these models.

3.1 Generative models

The dialogue generation task can be formulated as a next utterance prediction problem given a conversation history. In other words, given a source sequence (history) $X = (x_1, x_2, ..., x_{T_X})$ consisting of T_X words and a target sequence (response) $Y = (y_1, y_2, ..., y_{T_Y})$ of length T_Y, the model maximizes the generation probability of Y conditioned on X: $p(y_1, ..., y_{T_Y}|x_1, ..., x_{T_X})$ (Chen et al., 2017a).

The first generative model was proposed by Ritter et al. (2011) as a probabilistic model similar to phrase-based machine translation (Koehn et al., 2003). It generated single-turn dialogue responses based on a collected large-scale conversational post-reply corpus from Twitter. Later, most of the generative models follow Encoder-Decoder architecture. Initially, they were based on Recurrent Neural Networks, though later they transitioned from using attention layers to fully attention-based Transformers. With the rise of Language Models, it became possible to train rich and adaptable models that could be easily fine-tuned for a necessary task or even being used out-of-the-box.

Recurrent Neural Networks Encoder-Decoder Most of the first generative models are based on the Recurrent Neural Networks (RNN) Encoder-Decoder (Cho et al., 2014) inherited from phrased-based statistical machine translation. RNNs were used for scoring source-target phrase-pairs or to generate a target sequence given an input one. Encoder-Decoder represents the dialogue context as an embedding vector via an encoder network and generate the next response via a decoder network based on this contextual vector and the generated part of the sequence. The principal difference from machine translation task lays in semantic meaning difference between source and target sequences. In order to deal with this challenge, **?** proposed a first end-to-end single-round Neural Responding Machine with two decoders: (1) a global decoder calculating the vector that summaries the context, and (2) a local decoder using the attention mechanism to select the important words in context for various suitable responses. Since vanilla Recurrent Neural Networks suffer from vanishing gradients, most researchers were modelling first Encoder-Decoders with long short-term memory (LSTM) networks hidden units (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Chung et al., 2014). For instance, Vinyals and Le (2015) generated dialogue response of undefined length using z long short-term memory based sequence-to-sequence framework (Sutskever et al., 2014) initially introduced in neural machine translation. This model reads the input sequence token by token as well as generates an output one.

Cho et al. (2014) showed that the performance of a basic encoder–decoder deteriorates rapidly with increase of input sentence length. When a conversation is treated as one single temporally ordered sequence (Vinyals and Le, 2015; Shang et al., 2015), a potential issue is that all the necessary information is compressed into a fixed-length vector. Mei et al. (2017) adopts an attention alignment mechanism (Bahdanau et al., 2015) from machine translation for generating dialogue response. The basic idea is that the input sequence is encoded into a sequence of vectors instead of one fixed-length representation, so that the most relevant subset of these vectors can be chosen adaptively for each target word while decoding the dialogue response. Despite the popularity of sequence-to-sequence models, Mei et al. (2017) have chosen to base their approach on the language model, since it learns how the conversation evolves as information progresses, while sequence-to-sequence models learn only how the most recent dialogue response is generated. The dynamic attention model promotes coherence of dialogue responses with presented topics by favouring a flexible combination of salient words from conversation history for every individual generated.

Natural language dialogue involves at least two levels of structure. First, within a single utter-

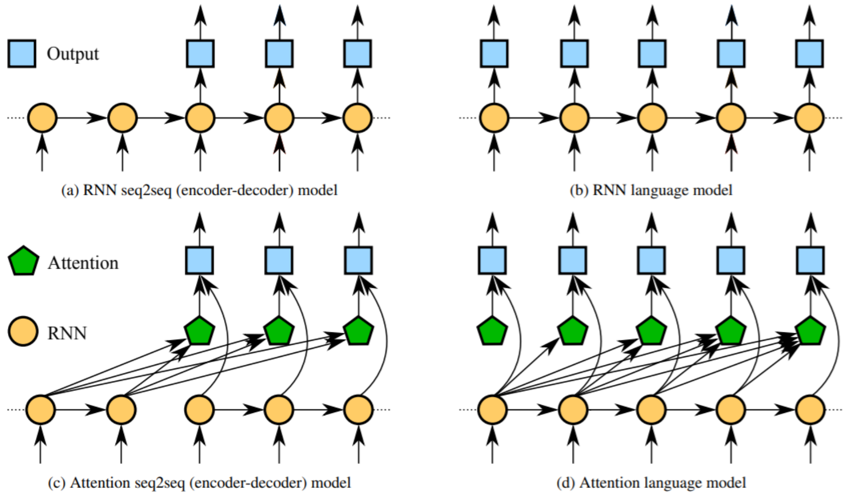


Figure 3.1: Comparing RNN language models to RNN sequence-to-sequence models, with and without attention (Mei et al., 2017)

ance, the structure is dominated by local statistics of the language. Second, dialogue possesses a hierarchical structure with complex dependencies between utterances like conversation topic, speaker goals and style. To explicitly model generative processes that possess multiple levels of variability, Serban et al. (2016) introduced a novel hierarchical neural network architecture with stochastic latent variable that span a variable number of time steps and facilitate the generation of long outputs and maintain the context. Being inspired by success of fully convolutional model over recurrent ones in machine translation, Gehring et al. (2017) and Mangrulkar et al. (2018) were the first to generate dialogue responses with convolutional neural network sequence-to-sequence model. It encapsulates dependencies between words in a better and more computationally efficient way than a recurrent neural network, since representations are built hierarchically.

Transformer-based Pre-trained models The success of attention mechanism (Bahdanau et al., 2015) for computing representation in previous works gave inspiration to create a first sequence transduction model based only on attention (Vaswani et al., 2017), which finally successfully replaced sequence-to-sequence models in plenty of tasks. The recurrent layers, commonly used in encoder-decoder architectures, have been replaced with multi-headed self-attention. Self-attention is an attention mechanism relating different positions of a single sequence in order to compute its representation.

Later Transformer architecture has been used not only for computing representation, but also pre-training them for many-time usage. Radford (2018) introduced a generative pre-training (GPT) and discriminative fine-tuning framework that has advanced the state-of-the-art in various NLP tasks. The main idea is to learn a universal representation from a large corpus of unlabelled text that could be easily transferred with little adaptation to a wide range of tasks by small labelled data with manually annotated training examples of these target tasks. These universal representations should learn world knowledge and long-range dependencies thanks to a diverse corpus. These advantages made GPT favourable to be adapted to natural language generation tasks (Golovanov et al., 2019) and to be extended to neural single-turn dialogue generation model like TransferTransfo (Wolf et al., 2019) and BlenderBot (Roller et al., 2021) for open domain conversation. Later, TransferTransfo model became popular and was used for

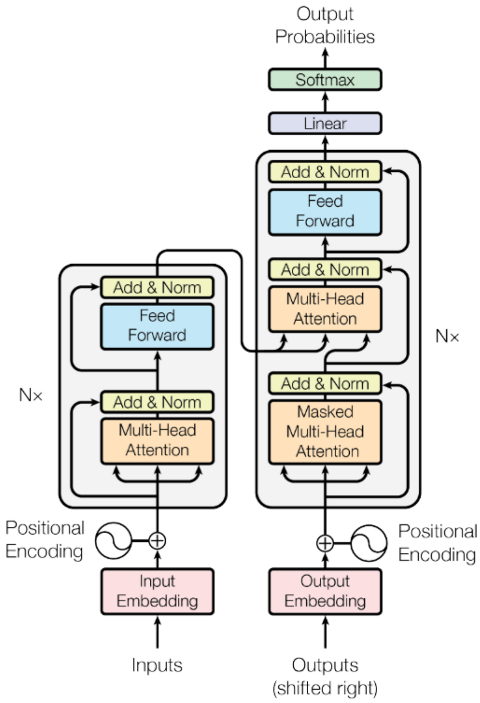


Figure 3.2: The Transformer - model architecture (Vaswani et al., 2017)

other dialogue systems like empathy bot CaiRE (Lin et al., 2019b). GPT-2 (Radford et al., 2019) became a basis for new dialogue response generative models like DLGNet (Oluwatobi and Mueller, 2020) and DialoGPT (Zhang et al., 2020b) to generate conversation responses on large-scale dialogue corpus or complex multi-domain task-oriented dialogue system trained on the MultiWOZ dataset (Budzianowski and Vulić, 2019). Similar to the concept of prompts in GPT-3 (Brown et al., 2020), Thoppilan et al. (2022) precondition Language Models for Dialog Applications (LaMDA) on a few turns of application-specific dialogue to adapt it to the target applications. This model makes use of a single model to perform multiple tasks: it generates potential responses, which are then filtered for safety, grounded on an external knowledge source, and re-ranked to find the highest-quality response. Evolved Transformer (So et al., 2019) became a base for training Meena chatbot (Adiwardana et al., 2020a) with the large-scale multi-turn conversations.

Mentioned language models are usually unidirectional and follow left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). It might not create a problem for sentence-level tasks, though it leads to loss of important information for token-level tasks such as question answering, where it is crucial to incorporate context from both directions (Devlin et al., 2019). It's also true for the dialogue generation task: Andreas et al. (2021) conducted a comparative study on different language models for task-oriented models and find out that Bidirectional and Auto-regressive Transformers (BART) (Lewis et al., 2020) and Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020) models achieve state-of-the-art performance with higher BLEU and F1 scores compared to GPT-2.

Viewing the dialogue context as a linear sequence of tokens for generating the next word through token-level self-attention hinders the exploration of discourse-level coherence among utterances. So next work has been done in the direction of designing dialogue-specific Transformer-based hierarchical architectures that capture its structure. Mehri et al. (2019) proposed masked-utterance retrieval and inconsistency identification pretraining objectives to learn effective representations of dialogue context that captures well both the utterance-level and context-level information. Gu et al. (2021) also proposed masked context regression and distributed utterance order ranking training objectives to efficiently capture discourse-level coherence. Their DialogBERT represent dialogue context by first encoding dialogue utterances through a Transformer encoder and then passes the utterances vectors to a discourse-level Transformer.

3.2 Retrieval models

It is well known that generative model responses may suffer from grammatical errors or hallucination, which is not the case for retrieval models. Instead of generating a new response, a retrieval-based model extracts a relevant response r from a corpus of context-response pairs or simply from a list of candidate responses based on some specific rules, features or certain knowledge.

The first attempt to learn to retrieve an appropriate response from the rising flood of accessible conversational data was the Learning-to-chat method (Jafarpour, 2010). Following a collaborative ranking approach, this model filters and ranks responses from a context that consists of several rounds of conversation using maximum information. To reduce the complexity of the task, Wang et al. (2013) and Ji et al. (2014) consider response retrieval as a simplified problem of a short text conversation where the dialogue context is reduced to the last dialogue turn and apply information retrieval methods previously mostly used for Web search. They showed that retrieval models could be considered more intelligent thanks to a large-scale conversational corpus that should be able to cover most of the possible contexts.

To scale to the often massive retrieval corpus, retrieval based models usually proceed in two steps as follows:

1. Retrieval: Given an input query q, this first step retrieves a number of candidate context-response pairs, forming a reduced candidate set R.

2. Matching and Ranking: the second step returns a matching feature set for each candidate context-response pair in the reduced set R and assigns a ranking score based on these features. The response with the highest score is chosen to respond.

3.2.1 Step 1: Retrieval

The goal of the retrieval stage is to perform a highly efficient and accurate approximate search over a large set of retrieval candidates to select a subset of relevant responses. Partly inspired from Web Search (Huang et al., 2013), retrieval models usually project context and response candidates into a common space and the relevance of a response given a context is computed as the cosine distance or dot-product between them.

Lexical Retrieval Early retrieval models (Ji et al., 2014; Lowe et al., 2015) relied on traditional unsupervised and sparse, term-based information retrieval methods such as BM25 (Robertson and Zaragoza, 2009). BM25 is a commonly-used bag-of-words retrieval function based on token-matching between two high-dimensional sparse vectors with TF-IDF token weights (Manning et al., 2008). Such a sparse model measures similarity using weighted term-matching between context and response candidates without being trained on a particular data distribution. While these hand-crafted features are fast and efficient for lexical matching, they fail to capture the semantics of the context and response candidates (like synonyms and paraphrases), which remained a major shortcoming for the task (Sciavolino et al., 2021).

Dense Retrieval Dense retrieval models map context and responses into a common embeddings space by computing their low-dimensional representations using deep neural networks (Johnson et al., 2021) or pre-trained neural encoders such as BERT (Devlin et al., 2019). Retrieval is then performed using fast maximum inner-product search (Shrivastava and Li, 2014) or nearest neighbour search (Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021). This helps to derive more semantically expressive embeddings and to overcome some limitations of sparse systems relying on discrete bag-of-words representation like vocabulary mismatch (different words are used to express the same meaning) (Nogueira, 2019) and semantic mismatch (the same word has multiple meanings) (Gao et al., 2021). On the other hand, sometimes dense models may show poor performance when the task requires salient phrases or rare entities matching. To address this issue, Chen et al. (2021) introduced the Salient Phrase Aware Retriever, a dense retriever with the lexical matching capacity of a sparse model.

3.2.2 Step 2: Matching and Ranking

After the response candidates R are retrieved for a given sequence of the utterances $C = u_1, ...u_n$, the next step is to rank them and select a final response r. This can be done with a context-response matching model $s(C; r)$ that calculates how likely a candidate r_i to be an appropriate response for the context C. Most of the earliest a context-response matching models $s(C; r)$ were supervised models learned with a set of triples $(y_i; C_i; r_i)$ where y_i is a (binary) label indicating the matching degree between C_i and r_i. As pre-trained language models became available, context-response matching can also be done in a zero-shot setting.

Wu et al. (2017) highlights two main challenges of dialogue response retrieval: (i) identifying context information like words or phrases, which are important for context/response matching, and (ii) modelling the relationships between the context utterances. Relatedly, matching models for response selection generally fall into two frameworks: representation-based where the whole context is handled as a single string and interaction-based, which takes dialogue structure into account and consider each context utterance separately. With the rise of pre-trained language models (PLMs), context-response matching models have also been proposed which started to be built with PLM-based matching framework.

Representation-based Models Early studies of retrieval-based chatbots focus on response selection for single-turn conversation (Wang et al., 2013; Ji et al., 2014; Wang et al., 2015; Wu et al., 2017), leveraging the last utterance in the context as a word sequence view. Later, researchers have begun to pay attention to multi-turn conversation, though still using sequence-to-sequence framework. By concatenating the context utterances as a long one, single-turn matching systems match the context with the response only one time, which makes them quick and robust.

Lowe et al. (2015) proposed a first fully end-to-end single turn response selection model in form of dual encoder: one for concatenated context utterances and second for response. Each input is represented as a fixed-size vector of the final hidden state of the corresponding LSTM (Hochreiter and Schmidhuber, 1997) encoder, fed with pre-trained Common Crawl (Pennington et al., 2014) word embeddings. The final hidden states from both encoder for context and response are used to calculate the probability of being a valid pair. Kadlec et al. (2015) explored variants of this dual encoder approach using Convolutional Neural Networks (LeCun et al., 1998) and Bidirectional LSTMs.

When the context and response are viewed as word sequences, the utterance-level discourse information and dependencies are being ignored. Being inspired by recommendation systems and representation learning (Elkahky et al., 2015), Zhou et al. (2016) introduced multi-view integration to response retrieval system. Their model takes advantage of both word-level and utterance-level semantic representations to compute context-response matching scores.

Interaction-based models During the dialogue, we create the next response heavily relying on the previous dialogue segments at different granularities (words, phrases, sentences, etc) over multiple turns rather than one turn (Lee et al., 2006; Traum and Heeman, 1996). However, representation-based models either ignore relationships among utterances when concatenating them together (Lowe et al., 2015), or lose important information by converting the whole context to a vector without enough supervision from responses (Zhou et al., 2016). In contrast, interaction-based models take dialogue structure into account, considering each dialogue turn in the context separately.

Wu et al. (2017) separately match the response with each utterance based on a convolutional neural network before aggregating these matching score into a final response score. This paradigm is applied in many subsequent works. It is different from previous representation-based models in "matching first" strategy. While matching a response with each utterance, it accumulates this matching information instead of sentences in a chronological order through a recurrent neural network, which models relationships among utterances. The final matching score is calculated with the hidden states of the recurrent neural network. Specifically, for each utterance-response pair, the model constructs a word-word similarity matrix and a sequence-sequence similarity matrix by the word embeddings and the hidden states of a recurrent neural network with gated recurrent units (Chung et al., 2014) respectively.

Inspired by Transformer's (Vaswani et al., 2017) successes in machine translation, Zhou et al. (2018) proposed to select a response for multi-turn context using only attention as well. For measuring the relevance between utterance and response, they exploit both self-attention and cross-attention. Self-attention extracts textual information, i.e. surface relevance, while cross-attention extracts dependency information, i.e. the way segments are semantically and functionally related to each other. Segment-to-segment similarity matrix of these two types are computed for different granularity to capture the matching features of the multi-turn context and response. The most important ones are then extracted via convolution with max-pooling to be further fused via a single-layer perceptron into a final matching score. This approach gave the start for Representation - Matching - Aggregation framework.

Yuan et al. (2019) claimed that excessive context might introduce some noise for response selection models. They propose Multi-hop Selector Network (MSN) that uses a multi-hop selector first to filter the relevant utterances as context using both word and utterance level representations. Then the selected context to fused to generate a better context representation and then feed it to a matching model to rank candidate responses. Initially, the model considers the last utterance as a key to start the selection, sometimes thought it contains something very general (like 'Okay' or 'Thanks'), so the previous utterances are considered as a starting point.

Building on the success of the Enhanced Sequential Inference Model (ESIM) (Chen and Wang, 2019), other approaches have based response retrieval on entailment rather than similarity. The ESIM model was originally developed for natural language inference (Chen et al., 2017b) and outperformed all previous models based on complicated network architectures. The premise and the hypothesis are encoded with a BiLSTM encoder, and the semantic relation between them is modelled by a cross attention mechanism. The final output is a probability that the premise entails the hypothesis obtained through max and mean pooling.

Vig and Ramea (2019) applies the ESIM model with ELMO embeddings to the multi-turn dialogue response selection to encode the relationships between each context utterance and a current response candidate and to return the aggregated result. The aggregation scheme is representing dialogue characteristics, i.e. utterance relevance depends on its position within the context on its speaker. Instead of using class probabilities of natural language inference, Vig and Ramea (2019) use the previous vector output to fuse it with speaker information. Later, the final score is computed based on the weighted average of these output vectors. These weights capture the varying importance of utterances based on their position within the context and learned during the training. In their paper, they provide the weights learned by the model for Ubuntu dataset that shows that utterances of the last turn speaker are more important and that the weights decrease as the distance from the target response increases.

PLM-based models Recently, the emergence of pre-trained language models like BERT (Devlin et al., 2019) that can be fine-tuned with just one additional output layer led to creation of state-of-the-art models for a wide range of tasks, including information retrieval. With BERT it is possible to represent each utterance-response pair and aggregate these representations to calculate the matching score (Vig and Ramea, 2019) or to treat the context as a long sequence and conduct context-response matching with BERT (Whang et al., 2020). During the post-training on dialogue corpus, this model also introduces the next utterance prediction and masked language model tasks borrowed from BERT to incorporate in-domain knowledge for the matching model. Following Whang et al. (2020), Gu et al. (2020) propose to heuristically incorporate speaker-aware embeddings into BERT to promote the capability of context understanding in multi-turn dialogues.

Yang et al. (2018) proposed to obtain sentence-level embeddings for semantic similarity task based on unsupervised multitask training from Reddit conversational data spanning 2007 to 2016 (Al-Rfou et al., 2016), combining dialogue response selection and natural language inference. They suppose that semantically similar sentences should rather have similar distribution of conversational responses (both "How old are you?" and "What is your age?" can be answered with "I am 20 years old") than surface similarity ("How are you?" and "How old are you?") which may result in different meaning.

Henderson et al. (2019c) proposed a compact dual-encoder (bi-encoder) pretraining architecture for response selection and later released its optimized version, ConveRT (Henderson et al., 2019b), pre-trained on large general-domain conversational Reddit dataset (Henderson et al., 2019a). During pre-training, each encoder learns unigram and bi-grams embeddings and applies positional embeddings and self-attention (Vaswani et al., 2017) to each of them separately. The representations are then combined averaged to give a single dimensional representation of the text (input or response). The relevance of each response to the given input is then quantified by the scaled cosine similarity. Later this model can be used in zero-shot setting, or it can be fine-tuned for some target dialogue domain, relying only on the small in-domain dataset that captures its nuances.

Besides bi-encoders, two other approaches are common: Poly-Encoders (Humeau et al., 2020), that encodes the context and response separately for attending over the context using the label candidate, and Cross-Encoders (Whang et al., 2020), that performs full attention over the inputs. Poly-Encoders can precompute candidates' representations for fast real-time inference that is valuable in a production setup, but this method suffers from suboptimal performance. On the other hand, Cross-Encoders' improvement in accuracy comes with a high calculation cost of the full attention for each sample.

Chapter 4

Dialogue Evaluation

Contents

One of the crucial steps in the development of a dialogue system is measurement of their performance. It still rests as a challenging task, since definition of high-quality dialogue may vary in different applications and development of relevant metrics is still an outstanding issue.

Initially, most dialogue evaluation was carried out through human annotation. However, the goal is to develop an automated and repeatable evaluation procedure that highly correlates with human judgments.

There are many ways to do that. We can perform corpus-based static evaluation based on some existing dialogue dataset. While it can be fast, efficient and easily reproducible, the results might be biased by the way the dataset was constructed. Another way to proceed is to conduct interactive evaluation when the crowd workers can chat with a dialogue model. If a conversational agent is a task-based model, probably some defined steps should be completed like mentioning some particular information, or if it's a non-goal oriented model, then crowd workers are usually given freedom to talk about anything they want.

In this section, we review different approaches that have been commonly used for the human and automatic evaluation of dialogue models.

4.1 Human evaluation

Human evaluation is very cost and time intensive. It is also difficult to ensure its reproducibility.

One difficulty is that it requires careful preparation, inviting and paying users to participate. To make the experiment close to real-world conditions, precise annotation guidelines must be provided, and the users should be properly instructed. Sometimes it requires several attempts to came up with a proper task description.

Another issue is that users' behaviour may vary a lot, as well as their understanding of the goal. For instance, Ghazarian et al. (2022) observed that paid and recruited users tend to be more proactive with dialogue system than the real ones. See and Manning (2021) has shown how user behaviour can impact the evaluation. If the user's utterance is unclear, the generative model tends to do some basic errors: being repetitive, ignoring and hallucinating or giving unclear responses (often, this is a vague question such as What is it?). Conversely, if the user's input is clear, the errors relating to reasoning or social abilities (redundant, logical, insulting) are becoming more common.

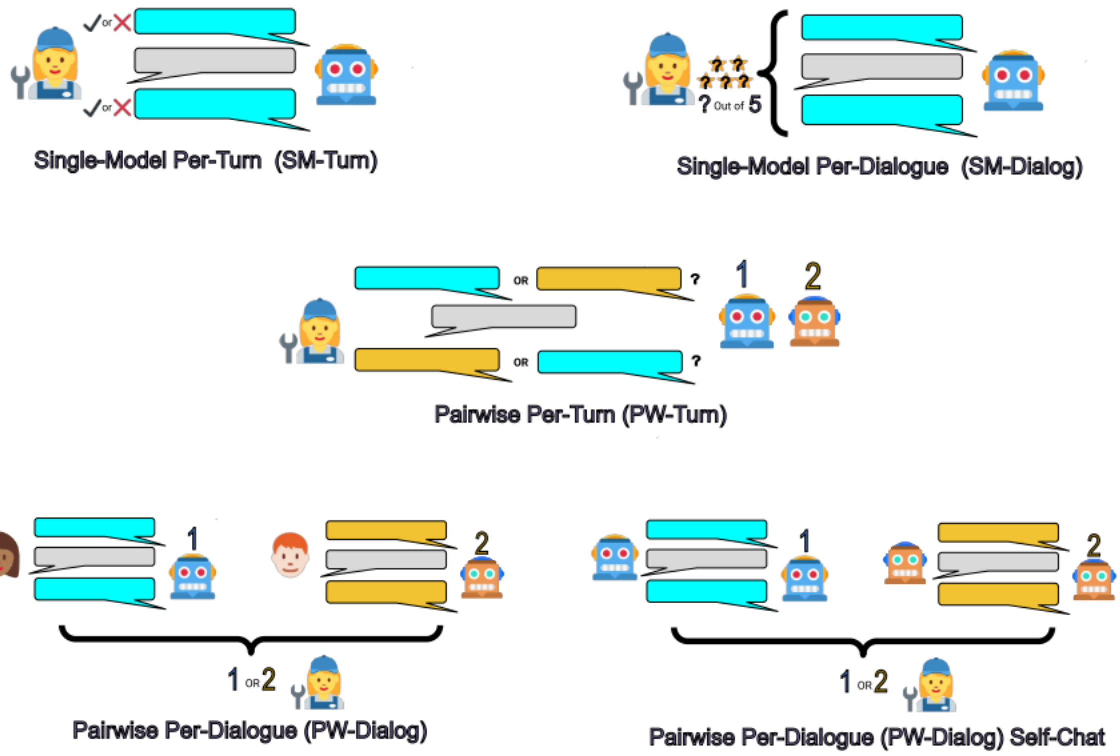


Figure 4.1: Different human evaluation methods (Smith et al., 2022)

Human evaluation of a dialogue can vary over multiple dimensions: single turn vs whole dialogue, static vs interactive, single model vs multiple model output.

Dialogue system evaluation can be performed *per-turn*, when ratings are given after every model response, and *per-dialogue*, when ratings are collected solely at the end of the conversation. Per-turn evaluations have the advantage of being more fine-grained, encouraging annotators to focus on small differences; however, the quality of a conversation is more than the sum of its parts, and it is better captured by global per-dialogue evaluations.

The evaluation can be *static*, looking at the results of the dialogue model, or *interactive*, evaluating turns and dialogue while talking to the model. During an interactive evaluation, humans are allowed to chat freely with bots (Ghandeharioun et al., 2019). Later the users are asked to give some scores to the model based on the whole dialogue to evaluate for quality, fluency, diversity, relatedness, and empathy. Interactive evaluation has been shown to be more reliable than turn-level static evaluation (Mehri and Eskenazi, 2020a).

If there are several models, they could be compared by *pairwise methods* (two models are compared directly by an annotator) or *single-model methods* (the annotator sees and rates only one model at a time). Both approaches can be either per-turn or per-dialogue.

Single-model approaches work well when direct comparison is not paramount, when models differ only slightly in quality but are otherwise similar, for example two models with different numbers of parameters (Smith et al., 2022).

The pairwise model can spot subtle differences apparent when comparing responses, and it can mitigate problems with distribution shift that occur in absolute scoring. *Pairwise per-turn* evaluations are adept at measuring changes in model performance throughout a conversation.

Figure 4.2: ACUTE-EVAL asks humans to compare two multi-turn dialogues, and independent of the gray speakers, choose between Speaker 1 (light blue) and Speaker 2 (dark blue) (Li et al., 2019)

Talking with a dialogue agent, the user must choose between two possible responses from the agent, each coming from two different models, after his or her own message. This technique tends to work well when pairs of models clearly differ in how appropriate their responses are in the context of the previous lines of dialogue, for example, when comparing two models that are trained on different datasets (Smith et al., 2022). *Pairwise per-dialogue* evaluations tend to perform best when differences between models only emerge after several conversation turns, such as when these differences are very subtle, or when noticing patterns in responses that emerge globally across the entire conversation, for example the average length of responses (Smith et al., 2022). ACUTE-EVAL (Li et al., 2019) provides a pairwise relative comparison setup for multi-turn dialogues. In each trial, the annotator is shown two whole conversations, with the second speaker in each conversation highlighted, as the judgment should be independent of the quality of the first speaker. ACUTE-EVAL provides a list of carefully worded questions with two choices: speaker A or B, where the question measures a desired quality such as which speaker is more engaging, interesting or knowledgeable.

4.2 Automatic static evaluation

Automatic metrics are frequently used as the most convenient way to evaluate a dialogue system fast, efficiently and reproducibly, that is especially valued for a quick development cycle. They can be divided into metrics for task-based vs. conversational dialogues, reference-based vs. reference-free method and turn- vs. dialogue-level approaches.

Task-oriented dialogues systems produces well-structured and tailored conversation that can be evaluated by task-success rate that represents percentage of correct actions taken. One of the first frameworks for dialogue evaluation was PARADISE (Walker et al., 1997), that measures task success and interaction costs weighted via linear regression to estimate user satisfaction on the dialogue level. Task-success rate is based on measuring the percentage of the most important building blocks of the interaction used. For example, the proportion of correct entities found (Henderson et al., 2014; Wen et al., 2017), mentions of specific dialogue acts or entities (El Asri et al., 2017b), correct API calls to database or questions answered with extracted from database information (Bordes and Weston, 2017) or satisfied all the associated information requests from the user (Su et al., 2015). The cost of the interactions represents effort needed for the user to invest in order to reach the goal: the number of utterances required in total, how many are needed to correct a misunderstanding, etc (Wahde and Virgolin, 2022). The ideal task-oriented dialogue satisfies the user goal with as little interactions as possible.

Conversational dialogues of non-task-oriented systems tend to be more broad and diverse and have been evaluated at the dialogue turn or on the whole dialogue, with and without a reference text.

References-based metrics are used for performing a corpus-based evaluation, in which the model is used to predict each system response in the held-out test set. They are computed by comparing the system response with a reference text. This can be done in several ways, based on word overlap, embeddings or language model.

Word-overlap metrics were originally proposed by the machine translation and the summarization community to measure the appropriateness of an utterance:

- **BLEU** (Papineni et al., 2002) compare the occurrence of n-grams of different order that are common between a candidate and multiple references.
- **METEOR** (Lavie and Agarwal, 2007) is calculated not only on exact matches like BLEU but also on stem, synonym, and paraphrase matches.
- **ROUGE-L** (Lin, 2004) measures sentence-level structure similarity by identifying the longest common subsequence of n-grams between the candidate and the reference utterances.

However, these standard word-overlap metrics do not correlate well with human judgements (Deriu et al., 2020; Liu et al., 2016), which could be explained by the open-ended one-to-many nature of conversations (Zhao et al., 2017b). In contrast with machine translation and automatic summarization, in dialogue generation task many possible responses could be relevant to a given context, so usage of these metrics is very limited (Liu et al., 2016). Even if the test dataset contains multiple references, typically they might not cover all responses that are possible in a given dialogue context, so it makes no sense to penalize a valid system response that deviates from these ground truth references.

In order to overcome these limitations, Liu et al. (2016) proposed a set of *embedding-based* metrics, where the cosine similarity between the predicted and the reference sentence is calculated over a continuous space representation:

- **Skip-Thought** (Kiros et al., 2015) maps a sentence into an embedding representation to predict the preceding and following sentences and to estimate semantic relatedness.

- **Embedding average** (Foltz et al., 1998; Wieting et al., 2015) computes a sentence-level embedding by mean pooling word embedding representations and then computes cosine semantic similarity

- **Vector Extrema** (Forgues et al., 2014) builds the sentence-level embedding by taking the most extreme value for each dimension of the embedding representations of all the words composing the sentence. The semantic similarity between hypothesis and reference is estimated by using cosine similarity.

- During **Greedy Matching** (Rus and Lintean, 2012) each word in the reference sentence is greedily matched to a word in the candidate sentence and the cosine similarity between both word embeddings is computed. All resulting word similarity scores are averaged. A similar score is also computed by reversing the roles of the candidate and reference sentences. The final score is the average of the two scores.

With the development of pre-trained language models like BERT (Devlin et al., 2019), embeddings-based metrics transformed into language model based, giving stronger performance and higher correlation with human judgements than traditional methods.

- **MoverScore** (Zhao et al., 2019) measures semantic similarity between reference and hypothesis by aligning semantically similar words and computing the amount of travelling flow between these words using the Word Mover Distance (Kusner et al., 2015).

- **BaryScore** (Colombo et al., 2021) measure the dissimilarity between reference and hypothesis by Wasserstein distance between the barycentric distribution of their contextual representations.

- **BERTScore** (Zhang et al., 2020a) is a pre-trained language model-based metric which computes precision, recall and F1 score between reference and hypothesis based on word-to-word embedding matching with cosine similarity.

As collecting references for each dialogue response generation task is difficult and time-consuming, various *reference-free* automatic evaluation metrics have been proposed for different aspects of dialogue or dialogue turns. They directly compare system outputs to source texts. Intuitively, these metrics learn how to assess the generated responses given dialogue contexts from different aspects.

- **language fluency** can be evaluated using perplexity (Adiwardana et al., 2020a) or a language model, trained with the masked language modelling (MLM) objectives, USR-MLM (Mehri and Eskenazi, 2020b)

- **diversity** can be evaluated through the ratio of distinct uni/bi-grams in generated responses (Li et al., 2016), information gain, the average number of unique tokens per dialogue, and conversation length, the average number of turns in the overall dialogue (Yu et al., 2017)

- **relevance** between the dialogue context and the system response has been measured by a hybrid model RUBER (Tao et al., 2018), consisting of both a referenced metric and an unreferenced metric, its improved version with contextualized word embeddings BERT-RUBER (Ghazarian et al., 2019) or by a dialogue retrieval model USR-DR (Mehri and Eskenazi, 2020b)

- **logical consistency** can be seen as natural language inference task (Dziri et al., 2019) by casting a generated response as the hypothesis and the conversation history as the premise, to better detect contradiction, Nie et al. (2021) proposed a structured utterance-based approach by pairing the utterances in advance

- **context coherence**, i.e. "whether a piece of text is in a consistent and logical manner, as opposed to a random collection of sentences" (Zhang et al., 2021a), can be assessed on a dialogue-level with CoSim (Xu et al., 2018; Zhang et al., 2018a) by averaging the cosine similarities between sentence embeddings of all adjacent utterance pairs within the dialogue, or with S-DiCoh (Mesgar et al., 2020) by modelling a dialogue at both the token and utterance level with a neural network framework consisting of two bidirectional LSTM layers with attention mechanism, or based on entity-grid (Cervone et al., 2018)

- **user engagement** can be measured through *Question Score* based on question words and/or a question mark in bot utterances and *Number of words* in the user response as a proxy of their engagement (Ghandeharioun et al., 2019; Ghazarian et al., 2020)

Graph-based and interactive approaches have also been proposed.
Huang et al. (2020) recently proposed *the GRADE metric*, leveraging graph modelling for turn-level coherence evaluation, trained on DailyDialogue (Li et al., 2017b). GRADE combines turn-level contextualized representations and topic keywords in context-response representations obtained by reasoning over topic-level graphs extended with a common-sense information. Later, Zhang et al. (2021a) introduces a unified graph-based automatic evaluation framework, DynaEval, that performs a turn-dialogue joint evaluation and is being trained on Empathetic Dialogue (Rashkin et al., 2019). In DynaEval, the dialogue is represented through the graph convolutional network with utterances as the graph nodes and dependencies between pairs of utterances as the edges. This model was trained to distinguish well-formed dialogues from synthetic negative ones using contrastive learning.

Finally, interactive methods attempting to model human rating. Ghandeharioun et al. (2019) proposed a *self-play scenario* where the dialogue system chats with itself and a combination of three metrics measuring sentiment, semantic coherence and engagement respectively along the conversation trajectory is computed to approximate dialogue-level quality estimation. Eskénazi et al. (2019) have shown the effectiveness of looking into the next user utterance as a proxy to measure the quality of the chatbot's generated responses. Mehri and Eskenazi (2020a) propose the unsupervised evaluation metric *FED metric*, which evaluates 18 fine-grained qualities of a system utterance in an interactive setting by computing the likelihood of a particular follow-up utterance responded by DialoGPT (Zhang et al., 2020b). Evaluation Dataset FED (Mehri and Eskenazi, 2020a) is a bench-marking dataset useful for both dialogue-level and turn-level evaluation. See and Manning (2021) have shown that predicting next user satisfaction helps to select more relevant system utterances. However, Berlot-Attwell and Rudzicz (2022)'s results demonstrate that relevance-related metrics of FED (i.e. FED-REL and FED-COR) are significantly negatively correlated with human relevance scores. Ghazarian et al. (2022) evaluate system responses by using the next user utterance as a proxy for automatically extracting features, like sentiments.

Yeh et al. (2021) provides a comprehensive assessment of various automatic evaluation metrics for dialogue proposed in recent years (2019 - 2021). Their study has shown that there is not necessarily a one-size-fits-all metric, and that most metrics performed best on the datasets they were originally evaluated on. Taking into account a long dialogue context becomes a struggle for many metrics, especially BERT-based ones. Among presented models in their work, USR and

GRADE were consistently showing higher on the majority of different datasets for evaluating turn-level response generation. Based on their investigation, Yeh et al. (2021) propose to combine different models and metrics to achieve strong results.

Part II

Medical chatbot for collecting patient information

Chapter 1

MedTree and MedBot

Contents

When developing a new dialogue model, a key bottleneck is the lack of adequate training data. Due to privacy issues, dialogue data is even scarcer in the health domain. We propose a novel method for creating dialogue corpora, which we apply to create doctor-patient interaction data. These interactions are often encoded in conversational trees created by experts. However, such dialogue systems may lead to poor user experience and lack of valuable information hidden behind short responses. We use paraphrases techniques to enrich these interactions and learn models that can alternate between following a predefined sequence of dialogue turns and topics addressed by the user. We learn both a generation and a hybrid classification/retrieval model, and find that the generation model consistently outperforms the hybrid model. We show that our data creation method has several advantages. Not only does it allow for the semi-automatic creation of large quantities of training data. It also provides a natural way of guiding learning and a novel method for assessing the quality of human-machine interactions.

1.1 Introduction

Data-driven dialogue models may require large quantities of training data. Because of privacy issues, the situation is even worse in the health domain, where data is particularly scarce. In this work, we propose a novel method for automatically creating the training data necessary to learn a

[0]This chapter was published as Liednikova et al. (2020)

chatbot which can mimic a doctor in doctor-patient interactions. Specifically, we combine expert knowledge provided by physicians with automatic paraphrase extraction techniques. We first ask experts (physicians) to specify typical doctor-patient interactions occurring in the context of clinical studies when talking about the four main topics generally discussed in these studies namely, sleep, mood, anxiety, leisure. Formally, the specification takes the form of a dialogue tree whose nodes are labelled with either an example doctor question or an example patient input. Each node in the tree is associated with a unique identifier, which can be viewed as a simple form of dialogue state.

We then enrich this initial dialogue data by extracting paraphrases for patient turns from an online forum.

This data generation method has several advantages. First, it allows for a straightforward integration of expert knowledge in data generation, model learning and model evaluation as we can use the dialogue turn identifiers both to guide learning and to assess the model (by comparing the sequences it follows with the expert defined sequences). More generally, the association of each dialogue turn with a dialogue turn identifier which reflects its position in the dialogue tree and the consistent use of this identifier during data creation, model learning and model evaluation allows for increased interpretability. Second, this method helps achieve good coverage, as we can ensure that the data does contain all possible dialogue paths. This is not the case with Wizard-of-Oz (WoZ) and crowdsourcing data collections approaches, where the coverage of the possible dialogue paths depends on the crowd-worker decisions and input. Third, by instantiating each dialogue with different paraphrases, we can increase linguistic diversity, i.e., we can create dialogues that have the same structure but different wording.

In sum, our work makes the following contributions. We propose a novel method for creating training data for dialogue models. We apply this method to create training data for a bot mimicking doctor-patient interaction in the context of clinical studies. We use the created data to learn a generation and a hybrid classification/retrieval dialogue model, we show that the generation model generally outperforms the classification model, and we provide a detailed analysis of the models results using automatic metrics, human evaluation and qualitative analysis.

1.2 Related Work

Various methods have been proposed to facilitate the creation of training data for dialogue. Previous work has explored WoZ experiments in which two humans interact based on some pre-defined scenario and the dialogues resulting from these interactions are collected (Green and Wei-Haas, 1985; El Asri et al., 2017a) or crowdsourcing settings where workers provide continuations to incomplete dialogues (Wen et al., 2017). Both approaches are time intensive. Crowdsourcing is also expensive while the human-human dialogues that are collected by both approaches may be very different from the human-machine interactions that should be learned to support efficient human-machine communication where, typically, chat messages are restricted in length. Other work has relied on already available dialogue data or on question/answer pairs extracted from online forums (Wei et al., 2018; Lin et al., 2019a; Xu et al., 2019). In the health domain, however, such data is extremely scarce and difficult to obtain. When obtainable, it also requires extensive pre-processing due to anonymization constraints. Another line of research has been to acquire data through machine-machine simulations (Xu et al., 2019; Majumdar et al., 2019; Shah et al., 2018a). In particular, Majumdar et al. (2019) combines pre-defined dialogue outlines with template-based verbalizations of dialogue turns to automatically create a synthetic dialogue corpus. Our work is similar to Majumdar et al. (2019) but differs from it in two main ways. First, instead of using templates, we use automatically extracted paraphrases to enrich the initial dialogues. Second, we experiment with two dialogue models to investigate how domain knowledge (in the form of dialogue tree positional information) can best be exploited to guide learning and to support error analysis.

1.3 Creating dialogue Corpora

To create training data for the dialogue bot, we start by collecting typical dialogue outlines from an expert. We then extract paraphrases for the patient turns from a Health forum and filter out dialogue interactions with low coherence.

Collecting an Initial Corpus from an Expert Studies have shown that the closed questionnaires traditionally used in the context of clinical studies are ineffective in gathering correct and precise information about the patient status because the patients get used to the questions and routinely input the same answers from one interaction to the next. Our long-term goal is to develop a Human-Machine dialogue system that would complement standard clinical questionnaires by regularly engaging the patient in a dialogue about the questionnaire topics. Since our target users are chronic pain patients, it is more important to keep them engaged for a long period rather than getting all information at the first interaction.

To create our dialogue corpus, we asked a physician to formalize typical patient-doctor interactions occurring in the context of a clinical study in the form of a dialogue tree describing which questions need to be asked and, for each question, which answers are possible. The interactions cover four domains namely, sleep, mood, anxiety and leisure activities and the dialogue tree has 58 nodes. A fragment of the dialogue tree created for the sleep domain is shown in Figure 2.1 on the left, and an example dialogue for the SLEEP domain in the same figure on the right. We call the data collected from the expert D_{init}.

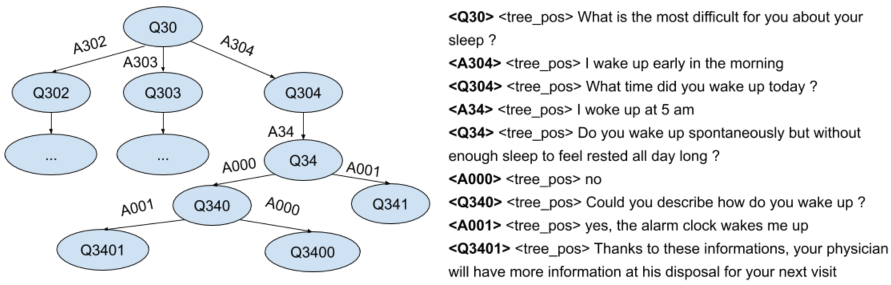


Figure 1.1: Fragment of dialogue tree for the sleep domain and a corresponding dialogue

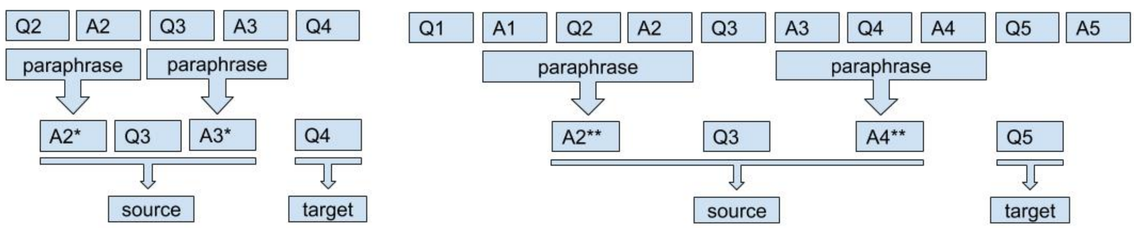


Figure 1.2: Paraphrasing source-target pairs from $INIT_{long}$ (left) and $INIT_{short}$ (right)

Extracting paraphrases We extract paraphrases for the patient turns provided by the expert from the HealthBoard [1] forum in several steps as follows.

As patient turns are mostly assertive responses to the doctor questions, we start by filtering out questions from the forum data to keep only those utterances which are assertions[2]. To this end, we use a binary stacked Bi-LSTM classifier trained on the Switchboard dataset.

We then compare each patient turn in D_{init} together with its context (P, the preceding doctor turn) with the assertive utterances extracted from the forum. For each sequence $D+P$ of contextualized patient turns in D_{init} and each (assertive) utterance U in the forum, we create an S-BERT embedding (Reimers and Gurevych, 2019) (cf. Figure 1.2 left). We then retrieve from the forum all utterances U whose cosine similarity with a contextualized patient turn $D+P$ is higher than 0.70. Finally, we use Maximal Marginal Relevance (MMR) (Goldstein and Carbonell, 1998) to select from this pool of candidates a subset of paraphrases which maximizes both similarity (the paraphrases should be semantically similar to the input turn) and diversity (the resulting set of paraphrases should be maximally diverse[3]). We stop selecting sentences as soon as MMR score becomes negative, as a negative MMR score indicates that adding more paraphrases will not increase diversity.

As illustrated in Figure 1.2, we apply this paraphrase extraction process not only to create paraphrases for a single turn, but also to create paraphrases which summarize 3 consecutive turns. In this way, we can derive compressed versions of the initial dialogues. For instance, we can derive the short dialogue from the longer dialogue interaction shown:

D1: Do you sleep well ?
P1: No
D2: What keeps you awake ?
P2: I have pain in the legs

D1: Do you sleep well ?
P1D2P2: No, I have pain in the legs and that keeps me awake.

We refer to the set of paraphrases that summarize three consecutive turns as SHORT and those that summarize a single turn as LONG.

Filtering Paraphrases We compute cosine and BertScore on the S-BERT embeddings of each pair $\langle C, D\rangle$ of context-doctor interactions (where the context is the string concatenation of the three preceding turns) created in the previous step and keep only those pairs for which both scores are higher than the corresponding scores for the corresponding turn in the initial corpus (INIT).

Training Data Table 1.1 summarizes the training data we created. INIT is the dialogue data collected from the expert; FORUM, the dataset obtained by replacing patient turns in INIT

[1] healthboards.com

[2] As noted by a reviewer, this is a simplification as in fact, users tend to formulate clarification and disambiguation questions. We leave this for future work.

[3] MMR is a measure for quantifying the extent to which a new item is both is similar to those already selected and similar to the target (here the patient turn). It is defined as: $Arg \max_{P_i \in C_U \setminus S} (\lambda\ Sim1(P_i, U) - (1 - \lambda) \max_{P_j \in S} Sim2(P_i, P_j))$ where U is a contextualized user turn, C_U is a pool of candidate paraphrases for U, $P_i; P_j$ are paraphrases in C_U, and S is the set of already selected paraphrases. A high λ value favours similarity. Conversely, a low λ value results in higher diversity. We set this parameter to be 0.5. We use BertScore recall (Zhang et al., 2020a) as function $Sim1$ as this permits checking similarity on a word basis and cosine as function $Sim2$ since we do not need precise comparison between forum sentences, we just want them to be diverse.

	INIT	FORUM	ALL	LONG	SHORT
Nb of src-tgt pairs	388	4 010 696	733 104	373 220	359 884
Nb of distinct turns	483	60 346	28 734	9 761	19 027
Nb of tokens	18 180	204 309 290	37 705 130	19 546 597	18 158 533
Avg Nb of tokens per turn	13.81	12.726	12.858	13.09	12.61
Vocabulary size	426	13169	11314	10 631	9593
Cosine		0.51	0.52	0.59	0.45
BertScore		0.83	0.83	0.85	0.83

Table 1.1: Corpus statistics (INIT: dialogue data collected from the expert; FORUM: extension of INIT with paraphrases; LONG: filtered FORUM dataset with only the single turn paraphrases; SHORT: filtered FORUM dataset with only the three-turn paraphrases; ALL = SHORT+LONG

with their paraphrases, and ALL, the dataset left after filtering. ALL is the combination of LONG and SHORT. As the table shows, the filtered dataset ALL is 5.5 times smaller but has similar coherence (identical or near identical cosine and BertScore scores) while retaining 86% of the vocabulary and 48% of the unique turns present in FORUM. To facilitate learning and reduce training time, we therefore use the filtered datasets in our experiments.

1.4 Health Bot Models

We aim to learn a model which mimics a physician in the kind of doctor-patient interaction that is typical of clinical studies conversations.

As we derive the training data from the dialogue tree, each patient turn and each doctor query is associated with a dialogue state (a node in that dialogue tree). We use this dual information (dialogue turn and dialogue state) to train and compare two models for response generation: a classification model which, given the last three turns of a doctor-patient interaction, predicts a dialogue state and outputs the corresponding doctor query; and a generative sequence-to-sequence model which auto-regressively generates an answer while conditioning on the last three dialogue turns. For both models, we use a pre-training and fine-tuning approach similarly to that presented by Radford (2018).

Classification model Given a dialogue context (3 dialogue turns), the classification model predicts a dialogue state and outputs the corresponding doctor query. Thus, the classification model is a multi-class classifier with 58 target classes, the 58 dialogue states defined by the expert dialogue tree. We use the PyTorch implementation of Radford (2018)'s pre-training and fine-tuning approach provided by Huggingface[4] and the default hyperparameter settings.

The input to the model consists of three turns $\langle p_1 d_1 p_2 \rangle$. We concatenate these three turns, prefixing each turn with its dialogue state identifier and separating them with a delimiter token. Each token is represented by the sum of three embeddings: a word and a position embedding which are learnt in the pre-training phase; and a turn embedding (learned during fine-tuning) indicating whether the token belongs to a patient or to a doctor turn. The input to the model is the sum of all three types – word, position and turn embeddings for each token in the input sequence.

The pre-trained model is the Generative Pre-trained Transformer-based (GPT-2) Language Model trained on the BooksCorpus dataset (7,000 books from different genres including Adventure, Fantasy, and Romance). The parameters are initialized to the smallest version of the GPT-2 model weights open-sourced by Radford (2018).

[4]https://github.com/huggingface/pytorch-openai-transformer-lm

We fine-tune the pretrained language model on our data by passing the input turns through the pre-trained model and feeding the final transformer block's activation to an added linear output layer followed by a softmax to predict a probability distribution over the target classes.

Generative model To generate (rather than retrieve) doctor queries, we use the TransferTransfo[5] model (Wolf et al., 2019), which combines a pretrained language model with a Transformer-based generation model fine-tuned on dialogue data using multi-task learning. Multi-tasking combines a language modelling loss with a next turn classification loss. For the latter, the model is trained to distinguish a correct continuation from one randomly chosen distractor. As for the classification model, we use the GPT-2 language model pretrained on the BooksCorpus. For fine-tuning, we use the same augmented representations as for classification, i.e. each input consists of the three previous turns with a separator and a dialogue state identifier between each turn. From this sequence of input tokens, a sequence of input embeddings for the Transformer is constructed by summing the word and positional embeddings learned during the pre-training phase and the turn embeddings learned during fine-tuning. Multi-task learning is done, as in the TransferTransfo model, by jointly optimizing the language modeling and the next-turn classification loss.

1.5 Experiments

1.5.1 Data and Experimental Setting

We train our models on LONG, SHORT and ALL (cf. Table 1.1) using an 80/20 train/validation ratio.

We created test data for both long and short interactions by manually specifying six distinct paraphrases for each user turn ($TEST_{LONG}$) or 3 turn sequences ($TEST_{SHORT}$) in $INIT$. Paraphrasing the tree user turns permits capturing alternative formulations of the same content, thereby allowing for an evaluation that better takes into account the paraphrasing capacity of natural language. Models trained on the ALL dataset are evaluated on $TEST_{ALL}$ which is a concatenation of $TEST_{LONG}$ and $TEST_{SHORT}$. $TEST_{LONG}$ has 4248 source-target pairs and $TEST_{SHORT}$ 2172.

Both models are 12-layer decoder-only transformer with masked self-attention heads (768 dimensional states and 12 attention heads) a dropout probability of 0.1 on all layers (residual, embedding, and attention). They use learned positional embeddings with supported sequence lengths up to 512 tokens. The input sentences are pre-processed and tokenized using bytepair encoding (BPE) vocabulary with 40,000 merges (Sennrich et al., 2016). Relu activation function is used. CLASSIF is a transformer with a language modelling and a classification head on top, the two heads are two linear layers. The classification head has dropout of 0.1. The model was fine-tuned with a batch size of 8, using OpenAI Adam with a learning rate of 6.25e-5 and a linear learning rate decay schedule with warmup over 0.2% of training. λ was set to 0.5. The GEN model is a transformer with a language modelling and a multiple-choice classification head on top, the two heads are two linear layers. The model was fine-tuned with a batch size of 4, using AdamW with a learning rate of 6.25e-5, $\beta 1 = 0.9$, $\beta 2 = 0.999$, L2 weight decay of 0.01. The learning rate was linearly decayed to zero over the course of the training.

For both models, we trained for 3 epochs using cross entropy loss.

1.5.2 Evaluation

We assess the output of our models using both automatic metrics and human evaluation.

[5] https://github.com/huggingface/transfer-learning-conv-ai

Automatic Metrics In our data, each dialogue turn is associated with a node (or dialogue state) in the initial dialogue tree drawn by the expert. We use this dual information (dialogue turn and dialogue state) for the evaluation. We compute F1 on dialogue state labels to analyse the coherence of the system response with the current dialogue context (For the generative model, if no label was predicted, the score is 0).

We also compute BLEU-4 and BertScore between the model output and the reference turn to assess the similarity of the generated output with the reference.

Human evaluation We ask annotators, coming from the ALIAE company working on health bots and from academia, to interact with a bot which at each new user turn outputs the doctor query suggested by one of our two models. The annotators are instructed to input free-text answers to the chatbot queries, and the interaction stops when the bot repeats a previously output question or when the annotator outputs a closing turn ('Bye!').

To assess the quality of the bots response given the dialogue context, annotators are required to score each system response on a 5 point Likert scale with respect to coherence ('Is the bot question coherent with the dialogue so far ?') where 1 is totally incoherent and 5 is perfectly coherent. For the generation bot, we additionally ask the annotators to rate fluency ('Is the bot response well-formed ?') where 1 is unreadable and 5 is perfectly readable. The annotators are non-native, but their English is fluent. For each model (CLASSIF and GEN trained on LONG), we collect 50 dialogues from 20 annotators. Each annotator interacts at most 5 times with the bot.

We also evaluate the quality of the full dialogues resulting from these human-bot interaction. At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a scale from 1 to 5. In addition, we applied the evaluation protocol proposed by Li et al. (2019). Using the 50 dialogue pairs collected for bot response evaluation, we show the annotators pairs of collected dialogues, one dialogue from the generation model and the other from the classification model and ask them the questions recommended by the protocol: 'Who would you prefer to talk to for a long conversation?' 'If you had to say one of the speakers is interesting and one is boring, who would you say is more interesting?' 'Which speaker sounds more human?' 'Which speaker has more coherent responses in the conversation?'. For this task, we had 16 annotators annotating 50 dialogue pairs. Each pair was rated 3 times except 2 pairs which were only rated twice. Each annotator annotated at most 10 dialogue pairs.

We report the percentage of time one model was chosen over the other. We also compute the average user turn length (number of tokens), the average dialogue length (number of turns) and the proportion of turn sequences of length at least two which occur in the dialogue tree (Sequence Rate). By assessing how often the bot reproduces a sequence of dialogue states that is present in the expert dialogue tree, this latter metrics provides an estimate both of a task success (i.e. how much of the required information has been collected, what proportion of the dialogue tree has been covered) and how much the collected dialogue deviates from the dialogue tree (how many turns are not about the medical topics covered by the dialogue tree).

1.6 Results

We compare the classification and the generation models using both automatic metrics and human evaluation. We present various ablation settings to analyse the impact of dialogue state information on performance. And we display an example dialogue between a human and the generative model in Table 1.7.

Model	F1			BLEU-4			BERTScore		
	L	**S**	**A**	**L**	**S**	**A**	**L**	**S**	**A**
CLASSIF Oracle	0.79	0.43	0.78	0.83	0.42	0.75	0.97	0.91	0.97
CLASSIF	0.63	0.38	0.48	0.65	0.39	0.48	0.95	0.91	0.92
CLASSIF (predict only)	0.63	0.37	0.40	0.66	0.37	0.42	0.95	0.91	0.91
GEN Oracle	0.83	0.68	0.85	0.62	0.52	0.62	0.96	0.95	0.97
GEN	0.66	0.39	0.50	0.49	0.34	0.37	0.95	0.93	0.93
GEN (predict only)	0.61	0.38	0.47	0.46	0.13	0.35	0.95	0.92	0.93
GEN (no d-state)	-	-	-	0.52	0.36	0.40	0.87	0.85	0.80

Table 1.2: **Results** on **L**ong, **S**mall and **A**ll Datasets

1.6.1 Automatic Metrics

Table 1.2 shows results for different versions of the generative and classification models, depending on which dialogue state information is provided in the source and the target, at test and at training time.

In the Oracle setting (Oracle), dialogue state information is provided for all dialogue turns in the input, at training and at test time. This gives an upper bound of how the system would perform given perfect dialogue state information. We compare this Oracle setting with a standard setting (CLASSIF and GEN) in which only the dialogue state associated with the doctor queries are given. At training time, this is the reference dialogue state associated with the doctor query. At test time, it is the dialogue state of the doctor query predicted by the model.

To analyse the impact of dialogue state information on performance, we also execute an ablation study considering models where (i) no dialogue state information is given in the input, but the model is trained to predict the output dialogue state (predict only) and (ii) a model where dialogue state information is not used at all (no dialogue states).

Generation outperforms classification The F1-score is consistently better for the generation models across all datasets, which suggests that learning to generate the system response also helps to predict the correct system dialogue state. As regards similarity with the reference, the generation models also consistently show better BERT score but lower BLEU-4 scores. This is coherent with the specificities of each model. Because the generative model generates the system response rather than select it from the training data (as is the case for the classification model), the similarity in terms of word overlap (as measured by BLEU) with the reference is lower. Nonetheless, the high BERT score indicates a strong semantic similarity between the reference and the generated output.

Predicting the output dialogue states helps For both classification and generation model, dialogue state information helps improve performance. As expected, the improvement is strongest for the Oracle setting. The ablation study further demonstrates that predicting and using predicted dialogue state information (CLASSIF, GEN) yields better results compared to settings where dialogue state information is only predicted (CLASSIF/GEN predict only) or not used at all (GEN no d-state).

Shorter interactions are hard to learn Contrasting the results from Short and Long in Table 1.2, we see that scores for the SHORT dataset are lower across the board – it is harder to handle short interactions. This is because, in that setting, the model needs to handle patient turns which convey multiple information – often from different domains – and, based on this, must decide on the correct response i.e. move to the correct dialogue state. For instance, in

Example (1.3), the model must (i) detect that the patient turn conveys information about both sleep and pain domain and (ii) decide to skip the dialogue state corresponding to D2 in example (1.3).

Domain analysis Table 1.3 shows the results per domain for the generation and the classification models trained on LONG[6]. Unsurprisingly, results are better for domains (Leisure and Anxiety) with a small number of classes (fewer transitions to learn) and when the training data is larger (Anxiety vs. Leisure and Sleep vs. Mood). This suggests two directions for further research: other paraphrasing techniques could be used to create more training data for those domains where the training data is small, and the dialogue tree drawn by the expert could be refined to yield more balanced domain subtrees.

Domain	# D-States	% Tg Data	CLASSIF			GEN		
			F1	BLEU-4	BERTScore	F1	BLEU-4	BERTScore
Mood	18	10.44	0.36	0.55	0.90	0.64	0.54	0.95
Sleep	33	63.40	0.47	0.64	0.92	0.60	0.51	0.95
Leisure	5	4.67	0.55	0.68	0.93	0.53	0.47	0.93
Anxiety	7	21.49	0.89	1.00	0.98	1.00	0.69	0.97

Table 1.3: **Results per Domain** (GEN and CLASSIF models trained on LONG)

1.6.2 Error Analysis

We use the expert dialogue tree to analyse how far off the model's predictions are from the correct predictions and compute the proportion of cases where the predicted dialogue state is the expected one (Correct), the child of this state in the dialogue tree (Child Node) or its parent (Parent Node). We also compute the proportion of cases where the predicted and the expected dialogue state have the same grandparent (Same Gd Parent) and for all remaining cases whether they occur as different leaves of the tree (Diff. Leaves) and are or are not in the same domain (In Domain, Out of Domain). Table 1.4 shows the results.

Most predictions are correct or almost correct We find that together the cases where the prediction is almost correct (Child or Parent Node) covers 13.93% and 13.57% of the cases for the generative and the classification model respectively. This means that the prediction of the dialogue state is correct or almost correct 76.52% and 80.01% of the time for the classification and the generative model, respectively.

Most errors are an artefact of the dialogue tree Most predictions which are very far off the expected dialogue state are transitions associated with the end of the dialogue (Diff. Leaves). This is because although turns concluding a dialogue are similar for all domains and all dialogue paths, they are associated in the dialogue tree with different dialogue state identifiers. This could be fixed by assigning each leaf node the same identifier and restarting the chatbot using a turn from another domain when reaching such a node. More generally, this shows that alternative design choices for representing the expert knowledge might impact performance.

Interestingly, the use of dialogue states derived from the expert dialogue tree increases interpretability and allows for a detailed analysis of the errors made by the models suggesting possible directions for improvement such as, for instance, using the same dialogue state identifier for the end of dialogue transitions in all domains and all dialogue paths (to reduce the proportion of

[6]For the other datasets (SHORT and ALL), we observe the same trends.

Diff. Leaves error) and focusing on identifying these factors which would help better differentiate between turns associated with closely related dialogue states Child or Parent Node.

Error Type	**CLASSIF**	**GEN**
Correct	62.59	65.86
Child Node	4.4	3.28
Parent Node	9.53	10.87
Same Gd Parent	1.69	1.31
Diff. Leaves	15.96	11.51
In Domain	4.28	6.65
Out of Domain	1.53	0.26

Table 1.4: **Error Analysis on Predicted dialogue States** (GEN and CLASSIF models trained on LONG)

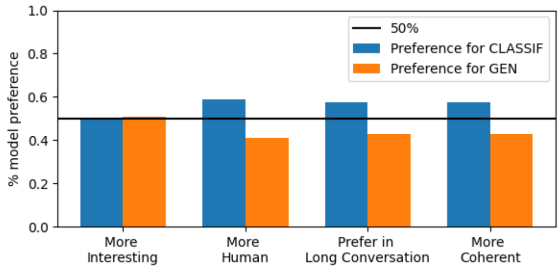


Table 1.5: **Human Comparison of the two Models**

1.6.3 Human Evaluation and Qualitative Analysis

Tables 1.6 and 1.5 show the results of the human evaluation.

Response quality We find that the generative model (GEN, fluency: 4.08) succeeds in generating well-formed responses. Responses that are rated low are often incomplete (e.g., 'in the long run, remaining with such unpleasant thoughts doesn't really seem to me to be 'ten' instead of 'tenable'). This is likely due to the model learning an average sentence length which is below that of longer turns and could be remedied by improved tuning. Both models provide reasonably coherent answers (CLASSIF: 3.14, GEN: 3.32) and while the generative slightly edges out the classification model, the difference (we used a t-test) is not statistically significant at $p < 0.05$.

Dialogue quality Dialogues are quite long, which indicates that the bot succeeds in driving a non-trivial conversation with the user. We also observe that the user turns are much shorter than in our training dataset because annotators often respond to questions by a simple yes or no statement rather than a full sentence. This raises the question of how to encourage the user to be more collaborative and provide more informative responses. We leave this here as an open question for further research.

The sequence rate is around a third for each model. Recall that this metrics is the ratio of bot turns that correspond to sub-sequences in the dialogue tree (with length more than two). A high score indicates that the model is consistent and capable of engaging the user in a conversation

Criteria	CLASSIF	GEN
Responses		
Coherence	3.14 ± 1.43	3.32 ± 1.40
Fluency	-	4.08 ± 1.28
Whole Dialogs		
Satisfaction	3.12 ± 1.19	3.38 ± 1.03
User turn length	4.92 ± 3.93	5.04 ± 4.24
dialogue length	24.42 ± 3.79	28.98 ± 5.63
Sequence rate	0.35 ± 0.13	0.26 ± 0.12

Table 1.6: **Human Evaluation Scores** (differences are not significant)

Model:	I am a bot and I will ask you some questions about your health. What is the most difficult for you about your sleep ?
Human:	I wake up multiple times during the night
Model:	is there a specific event or a change in your activity which could explain this pain?
Human:	I believe it's because of my incoming exams
Model:	are they frequent or infrequent?
Human:	Only twice a year
Model:	even if it is not always easy to treat chronic sleep issues why don't you speak with
Human:	I don't think it's that urgent for now
Model:	okay! do you have any other things to tell me?
Human:	I really feel tired all day long
Model:	would you agree to discuss anxiety and stress?
Human:	Sure!
Model:	if you already feel those troubles for the last days why don't you speak with your physician
Human:	I don't think I really need it for now
Model:	Do you feel like events have been able to affect your mood lately?
Human:	My exams
Model:	do you ever find yourself thinking that it would be best not to wake up the next mornin
Human:	No, it's just my exams!!
Model:	do you sometimes think about ending your life?
Human:	I just told you I'm ok
Model:	in the long run remaining with such unpleasant thoughts doesn't really seem to me to be ten
Human:	I think I should be good after my exams are finished
Model:	Were you rather anxious?
Human:	Yes, I still am, I don't want to fail
Model:	What activities you enjoy participating in?
Human:	I like to play baseball with my university's team
Model:	Okay! Thank you for sharing. So far I have nothing new to tell you.

Table 1.7: **Conversation between Human and Generative Model**.

according to the tree. A low score indicates that the model diverges from the dialogue tree without creating the expected series of questions, but it also indicates that contrary to a finite state dialogue approach where the model is constrained to follow the transitions defined by the finite state automaton, our models can learn new dialogue transitions. The observed sequence rates (0.35 and 0.26) suggest both that the models have correctly learned transitions sequences that were defined as natural sounding by the expert and that they can deviate from those, learning new ways to conduct the dialog. We leave a detailed exploration of how these deviations could be used to create alternative dialogue paths and thereby enrich the model for further research.

The Acute-Eval results are more nuanced. While the satisfaction (Table 1.6) and the interest scores (Table 1.5) are higher for the generative model, the classification model is found more human sounding, more coherent and is preferred for a long conversation. This is in line with previous results (Zhang et al., 2018b; Dinan et al., 2018), where retrieval models (approximated here by our hybrid classification/retrieval model) were found to score very well in dialogue level evaluations because they return human-written utterances from the training set and thus do not suffer from decoding mistakes present in generative models.

1.7 Conclusion

Using paraphrase identification techniques and a dialogue tree to model expert knowledge about doctor-patient interactions, we proposed a novel method to create training data for dialogue models, and we used data created using this method to learn health chatbots that cover the main topics standardly used in the questionnaires of clinical studies. We compared two models, a generative and a hybrid classification/retrieval model, and we showed that the expert knowledge captured by the dialogue tree both helps guide learning and facilitate error analysis.

Results analysis highlights three main directions for future research. First, additional paraphrase techniques could be explored to create a more balanced dataset. As shown in Table 1.3, the quantity of training data available for each domain varies greatly. We are currently exploring whether paraphrase generation (rather than paraphrase extraction) could help address this issue. Second, longer, richer dialogues could be obtained by extending the expert dialogue tree. Here, the American Medical Association Family Medical Guide (Kunz, 1982) may be used to obtain a new dataset with longer and more precise interaction between doctor and patient, giving more advice and information about patient's state.

Third, even in a clinical study context, human dialogues will often mix open-ended chit-chat with targeted health domain interactions. It would be interesting to extend our approach with strategies that engage the user to talk more about his/her problems, e.g., by using ensemble of bots (Papaioannou et al., 2017b).

Chapter 2

ComBot: Ensemble of the bots

Contents

We focus on dialogue models in the context of clinical studies where the goal is to help gather, in addition to the close information collected based on a questionnaire, serendipitous information that is medically relevant.

To promote user engagement and address this dual goal (collecting both a predefined set of data points and more informal information about the state of the patients), we introduce an ensemble model made of three bots: a task-based, a follow-up and a social bot. We introduce a generic method for developing follow-up bots. We compare different ensemble configurations, and we show that the combination of the three bots (i) provides a better basis for collecting information than just the information seeking bot and (ii) collects information in a more user-friendly, more efficient manner than an ensemble model combining the information seeking and the social bot.

[0]This chapter was published as Liednikova et al. (2021)

2.1 Introduction

Current work on Human-Machine interaction focuses on three main types of dialogues: task-based, open domain and question answering conversational dialogues. The goal of task-based models is to gather the information needed for a given task, e.g., gathering the price, location and type of restaurant needed to recommend this restaurant. Usually trained on social media data, open domain conversational models aim to mimic open domain conversation between two humans. Finally, question answering conversational models seek to model dialogues where a series of interconnected questions is asked about a text passage.

In this chapter, we consider dialogue models in the context of clinical studies, i.e., dialogue models which are used to collect the information needed by the medical body to assess the impact of the clinical trial on a cohort of patients (e.g., information about their mood, their activity, their sleeping patterns). In the context of these clinical studies, the goal of the dialogue model is two-fold. A first goal is to collect a set of pre-defined data points, i.e., answers to a set of pre-defined questions specified in a questionnaire. A second goal is to gather relevant serendipitous information, i.e., health related information that is not addressed by the questionnaire, but that is provided by the user during the interaction and which may be relevant to understand the impact of the therapy investigated by the clinical study. This requires keeping the user engaged and prompting him/her with relevant follow-up questions.

To model these three goals (collecting a predefined set of data points, keeping the user engaged and gathering more informal information about the state of the patient), we introduce an ensemble model which combines three bots: a task-based bot (MedBot) whose goal is to collect information about the mood, the daily life, the sleeping pattern, the anxiety level and the leisure activities of the patients; a follow-up bot (FollowUpBot) designed to extend the task-based exchanges with health-related, follow-up questions based on the user input; and an empathy bot (EmpathyBot) whose task is to reinforce the patient engagement by providing empathetic and socially driven feedback.

Our work makes the following contributions.

- We introduce a model where interactions are driven by three main goals: maintaining user engagement, gathering a predefined set of information units and encouraging domain related user input.
- We provide a generic method to create training data for a bot that can follow-up on the user response while remaining in a given domain (in this case, the health domain).
- We show that such a follow-up bot is crucial to support both information gathering and user engagement, and we provide a detailed analysis of how the three bots interact.

2.2 Related Work

Several approaches have explored the use of ensemble models for dialogue. While Song et al. (2016) proposed an ensemble model for human-machine dialogue which combines a generative and a retrieval model, further ensemble models for dialogue have focused on combining agents/bots designed to model different conversation strategies. Yu et al. (2016) focus on open domain conversation and combines three agents, two to improve dialogue coherence (ensuring that pronouns can be resolved and maximizing semantic similarity with the current context) and one to handle topic switch (moving to a new topic when the retrieval confidence score is low). The ALANA ensemble model (Papaioannou et al., 2017b,a), developed for the Amazon Alexa Challenge, i.e., for open domain chitchat, combines domain-specific bots used to provide information from different sources with social bots to smooth the interactions (by asking for clarification, expressing personal views or handling profanities). Similarly, Yu et al. (2017) introduces a dialogue model which interleaves a social and a task-based bot. Conversely, Gunson et al. (2020) showed that success of interleaving depends on the context and that in a public setting, users either prefer purely task-based systems or fail to see a difference between task-based and a richer ensemble model combining task-based and social bots.

Our work differs from these previous approaches in that we combine a standard, task-based model with both a social bot and a domain specific, follow-up bot. This allows both for more natural dialogues (by following up on the user input rather than systematically asking about an item in the predefined set of topics) and for additional relevant, health related information to be gathered.

2.3 ComBot, an ensemble Model for Repeated Task-Based Interactions

We introduce the three bots making up our ensemble model and the ensemble model combining them.

2.3.1 Medical Bot

MedBot is a retrieval model which uses the pre-trained ConveRT dialogue response selection model (Henderson et al., 2019b) to retrieve a query from the MedTree Corpus (Liednikova et al., 2020). It is designed to collect information from the user based on a predefined set of questions contained in a questionnaire.

The MedTree Dataset The MedTree corpus (Liednikova et al., 2020) was developed to train a task-based, information seeking, health bot on five domains: sleep, mood, anxiety, daily tasks and leisure activities. It was derived from a dialogue tree provided by a domain expert (i.e., a physician) and designed to formalize typical patient-doctor interactions occurring in the context of a clinical study. In that tree, each branch captures a sequence of (Doctor Question, Patient Answer) pairs and each domain is modelled by a separate tree with the root introducing the conversation (initial question) and the leaves providing a closing statement. The MedTree corpus is then derived from this tree by extracting from each branch of the tree, all context-question pairs, where the context consists of a sequence of patient-doctor-patient turns present on that branch and the question is the following doctor question. A fragment of the decision tree created for the sleep domain and an example dialogue are shown in Figure 2.1.

There are two versions of the MedTree corpus: one consisting of only the context/question pairs derived from the dialogue tree (INIT) and the other including variants of these pairs based on paraphrases extracted from forum data (ALL). Liednikova et al. (2020) used the ALL corpus to train a generative and a classification model. In our work, we use (a slightly modified version[1] of) the INIT corpus instead, as its small size facilitates retrieval (the number of candidates is small) and preliminary experimentation showed better results when using the INIT corpus.

Model ConveRT is a Transformer-based Encoder-Decoder which is trained on Reddit (727M input-response pairs) to identify the dialogue context most similar to the current context and to retrieve the dialogue turn following this context. To retrieve from the MedTree corpus, the question that best fits the current dialogue context, the MedBot model compares the last three turns of the current dialogue with contexts from the MedTree Corpus. The model identifies the

[1]The modifications consists in shortening the questions, changing all leaves to statements and adding meta-statements about the dialogue to account for cases where the user indicates misunderstanding or agreement

MedTree corpus context with the highest similarity score[2] and outputs the question following that context. If the selected question has already been asked in the dialogue generated so far and provided it is not a question such as “What other things would you like to share with me ?”, we retrieve the next best question that is not a repetition. No fine-tuning is done due to the small amount of data.

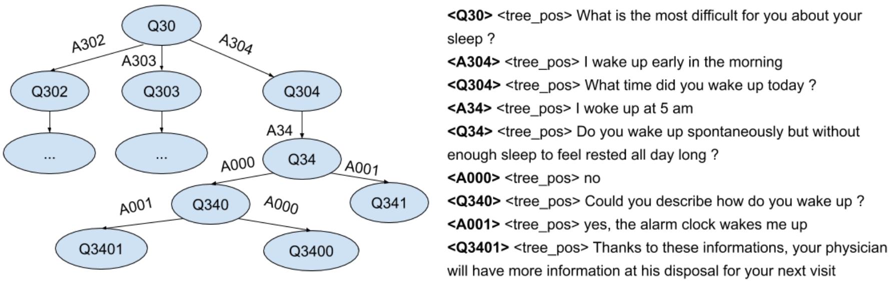


Figure 2.1: **Fragment of decision tree for the sleep domain and a corresponding dialog**

2.3.2 Follow-Up Bot

One main motivation behind the use of a health-bot in clinical studies is to complement the information traditionally gathered through a fixed questionnaire filled in each week by the patients with serendipitous information, i.e., information that is not actively queried by the questionnaire, but that is useful to analyse the cohort results.

The MedBot model introduced in the previous section is constrained to address only those topics which are present in the dialogue tree, in effect, modelling a closed questionnaire. To allow for the collection of serendipitous health information, we develop the FollowUpBot whose function is to generate health-related questions which are not predicted by the dialogue tree but which naturally follow from the user input. Rather than artificially restricting the dialogue to the limited set of topics pre-defined by the dialogue tree, the combined model (MedBot + FollowUpBot) allows for transitions based either on the dialogue tree or on health-related, follow-up questions. In that sense, FollowUpBot allows not only for the collection of health-related serendipitous information, but also for smoother dialogue transitions.

Like MedBot, FollowUpBot used the pre-trained ConveRT model to retrieve context-appropriate queries from a dialogue dataset. In this case, however, the queries are retrieved from the HealthBoard dataset, a new dataset we created to support follow-up questions in the health domain.

The Healthboard Dataset This dataset consists of (s, q) pairs where s is a (health related) statement and q is a follow-up question for that statement. We extract this dataset from the HealthBoard forum[3] as follows. We first select 16 forum categories (listed in Table 2.1) that are relevant to our five domains. In the forum, each category includes multiple conversational threads, each thread consists of multiple posts and each post is a text of several paragraphs that can be split into sentences. In total, we collect 175,789 posts from 31,042 threads with 5.68 posts in average per thread. We then segment each post into sentences using the default NLTK sentence segmenter. We label each sentence with a dialogue act classifier in order to distinguish statements ("sd" label) from questions ("qo" label). For this labelling, we fine-tune the

[2]Both contexts are encoded using ConveRT as average of embeddings of the last turn and concatenation of preceding ones. The inner product is used to compute similarity.

[3]https://www.healthboards.com/

Category	Threads	Posts	Avg
anxiety	6852	38523	5.63
anxiety tips	42	71	1.69
chronic fatigue	670	3856	5.77
chronic pain	646	4893	7.59
depression	5327	32998	6.21
depression tips	27	51	1.89
exercise fitness	1583	8142	5.16
general health	7279	29858	4.11
healthy lifestyle	104	621	5.97
pain management	4985	38738	7.79
panic disorders	1314	8376	6.39
share your anxiety story	42	42	1
share your depression story	55	71	1.29
share your pain story	28	42	1.50
sleep disorders	1671	7656	4.59
stress	415	1973	4.76

Table 2.1: Forum Categories used for the Creation of the HealthBoard Dataset

DistilBERT Transformer-based classification model [4] on the Switchboard Corpus (Stolcke et al., 2000) using 6 classes "qo" (Open-Question), "sd" (Statement-non-opinion), "ft" (Thanking), "aa" (Agree/Accept), "%" (Uninterpretable) and "ba" (Appreciation). For each question q (i.e., sentence labelled "qo") in each thread T, we gather all statements (i.e., all sentences labelled as "sd") which precede q in T into a pool of candidate statements[5]. As dialogue turns in bots should remain short, we filter sentences that have more than 100 tokens. For each candidate statement, we calculate its similarity with the question using the dot product of their ConveRT embeddings. We filter out all candidate statements whose score with the question is less than 0.6. If, after filtering, the resulting pool contains at least one candidate, we select the top-ranked statement and add the statement-question pair to the dataset. The resulting dataset contains 3,181 (statement, question) pairs.

Model Similar to the MedBot model, the FollowUpBot model used the pre-trained ConveRT model to compare the current dialogue context (the preceding three turns) with the statements contained in the HealthBoard dataset using the inner product.

The top-20 candidates are then retrieved and filtered using Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) with $\lambda = 0.5$ to control for repetitions[6]. Next, we compute the similarity between the remaining selected questions and the questions included in the current dialogue context (all preceding dialogue turns) and we exclude candidates with similarity score 0.8 or higher. After filtering, the top ranking candidate is selected and the associated follow-up question is output.

[4] https://huggingface.co/distilbert-base-uncased

[5] We do not restrict the set of candidates at that stage, i.e., we consider all posts that precede the question within the question thread and all statements in these posts, no matter how far away the statement is from the question. In practice, the set of such statements has limited size and distance does not seem to matter too much, although an investigation of that factor would be interesting. We leave this question open for further research, as it is not central to our work.

[6] MMR is a measure for quantifying the extent to which a new item is both dissimilar to those already selected and similar to the target (here a selected question). A λ value of 0.5 favours similarity and diversity equally, both matter equally.

2.3.3 Empathy Bot

As the name suggests, the role of the EMPATHYBOT is to engage the user by showing empathy. For this bot, we use Roller et al. (2021)'s generative model which was pre-trained on a variant of Reddit discussion (Baumgartner et al., 2020) and fine-tuned on the ConvAI2 (Zhang et al., 2018b), Wizard of Wikipedia (Dinan et al., 2018), Empathetic Dialogues (Rashkin et al., 2019), and Blended Skill Talk datasets (BST) (Smith et al., 2020) to optimize engagigness and humanness in open-domain conversation.

2.3.4 Ensemble Model (ComBot)

Each bot provides a single candidate. To rank them, we encode the whole current dialogue context and each candidate response using the ConveRT encoder, we calculate similarity (dot product) for each candidate/context pair, and we select the candidate with the highest similarity score. In case all candidates scores are less than 0.1, we consider that there is no good response, and we end the conversation.

2.4 Experiments

2.4.1 Data

Table 2.2 shows some statistics for the corpora used for pretraining (ConveRT, Blender) and for retrieval (INIT, HealthBoard). For MEDBOT and FOLLOWUPBOT, we use the ConveRT model from PolyAI[7]. For EMPATHYBOT, we use the Blender model with 90M parameters from the ParlAI library[8].

One benefit of the ensemble approach is that several models can be combined, each modelling different types of dialogue requirements. We compare different configurations of our three bots: COMBOBOT (which combines the three bots), MEDBOT (using only the task-based bot) and MED+EMPATHYBOT an ensemble model which combines the task-based (MEDBOT) and the social bot (EMPATHYBOT). In this way, we can compare a model (COMBOBOT) which allows for follow-up questions (i) with one that doesn't but allows for some social interactions (MED+EMPATHYBOT) and (ii) with one (MEDBOT) which is purely task-based.

We first use automatic metrics and global satisfaction scores to compare the three models. As the MED+EMPATHYBOT model generally under-performs regarding the other two models, we restrict the Acute-Eval, human-based model comparison to COMBOBOT and MEDBOT.

	Reddit	ConvAI2	WoW	EmpaDial	BSD	INIT	HealthBoard
Nb of context-question pairs		211803	83011	76673	27018	168	3181
Nb of distinct turns	1.50B	267945	165213	88757	53335	154	73140
Nb of tokens	568B	3791971	2720426	2625338	912857	3688	202389
Nb of tokens per turn		8.95	16.39	17.12	16.89	6.92	11.5
Vocabulary size		20707	95590	59438	52561	306	7321

Table 2.2: Corpus statistics (Reddit: pre-training corpus for ConveRT and the Empathy bot. ConvAI2, WoW, EmpaDial and BSD: Datasets used to fine-tune the Empathy Bot. INIT: used for the MedBot retrieval step. HealthBoard: for FollowUp Bot Fine-Tuning and Retrieval.)

[7] https://github.com/connorbrinton/polyai-models/releases/tag/v1.0
[8] https://parl.ai/projects/recipes/

2.4.2 Evaluation

As there does not exist a dataset of well-formed health-related dialogues whose aim is both to answer a clinical study questionnaire and to allow for serendipitous interactions, we have no test set on which to compare the output of our dialogue models. Moreover, as has been repeatedly argued, reference-based, automatic metrics such as BLEU or METEOR, fail to do justice to the fact that a dialogue context usually has many possible continuations. We therefore use reference-free automatic metrics and human assessment for evaluation.

Human evaluation We use the MTurk platform to collect human-bot dialogues for our three models (ComboBot, MedBot and Med+EmpathyBot) and ask the crowdworkers to provide a satisfaction rate at the end of their interaction with the bot. We then run a second MTurk crowdsourcing task to grade and compare dialogues pairs produced by different models.

To collect dialogues, we ask participants to interact with the bot for as long as they want. The conversation starts randomly with one of the initial questions of MedBot. The interaction stops either when all candidates scores are less than 0.1 (cf. Section 2.3.4) or when the user ends the conversation. For each model, we collect 50 dialogues. Each annotator interacts at most once with a bot.

At the end of each human-bot conversation, the annotator is asked to rate satisfaction on a 1-5 Likert scale (a higher score indicates more satisfaction).

Assigning a satisfaction score to a single dialogue is a highly subjective task, however, with scores suffering from different bias and variance per annotator (Kulikov et al., 2019). As Li et al. (2019) argued, comparing two dialogues, each produced by different models, and deciding on which dialogue is best with respect to a predefined set of questions, helps support a more objective evaluation. We therefore use the Acute-Eval human evaluation framework to compare the dialogues collected using different bots. Since the automatic evaluation (cf. Section 2.5.1) shows that ComboBot and MedBot are the best systems, we compare only these two systems asking annotators to read pairs of dialogues created by these two bots and to then answer the pre-defined set of questions recommended by Li et al. (2019)'s evaluation protocol namely:

- Who would you prefer to talk to for a long conversation?
- If you had to say one of the speakers is interesting and one is boring, who would you say is more interesting?
- Which speaker sounds more human?
- Which speaker has more coherent responses in the conversation?

We report the percentage of time one model was chosen over the other.

For this comparison, we consider 50 dialogue pairs (one dialogue produced by ComboBot, the other by MedBot) and for each Acute-Eval question, collected 50 judgments, one per dialogue pair. We had ten annotators, each annotating at most 5 dialogue pairs. To maximize similarity between the dialogues being compared, we create the dialogue pairs by computing euclidean distance between context embeddings of MedBot and ComboBot dialogue sets. Then we composed a pair of two closest items and excluded them from the choice in the next iteration.

Automatic Metrics After collecting dialogues, we perform their automatic evaluation. All scores are computed on the 50 bot-human dialogues collected for a given model. Table 2.3 shows the result scores averaged over 50 dialogues.

To measure *coherence*, we exploit the unsupervised model CoSim (Mesgar et al., 2019; Xu et al., 2018; Zhang et al., 2018a). This model measures the coherence of a dialogue as the average of the cosine similarities between ConveRT embedding vectors of its adjacent turns.

To assess *task success*, we count the number of unique medical entities (Slots) mentioned. We do this using the clinical NER-model from the Stanza library (Zhang et al., 2020c)[9], a model trained on the 2010 i2b2/VA dataset (Uzuner et al., 2011) to extract named entities denoting a medical problem, test or treatment.

We report the average number of medical entities, both per dialogue and in the user turns (to assess how much medical information comes from the user).

Model	Satisf.	CoSim	Slots	InfoGain	ConvLen	UserQ
MedBot	3.92	0.26	6.40 (1.72)	106.82 (3.89)	27.46	0.08 (4)
Med+EmpathyBot	3.74	0.31	4.00 (1)	170.50 (2.00)	85.00	0
ComboBot	3.82	0.36	7.22 (2.78)	126.66 (5.93)	21.34	0.45 (23)

Table 2.3: Satisfaction Scores (Satisf.) and Results of the Automatic Evaluation. CoSim: Average Cosine Similarity between adjacent turns. Slots: Average Number of Medical Entities per dialogue (in brackets: average number in the user turns). ConvLen: Average Number of turns per dialogue. InfoGain: Average number of unique tokens per dialogue (in brackets: normalized by dialogue length). UserQ: number of questions asked by Human (in bracket: total number for 50 dialogues). All metrics are averaged over the 50 Human-Bot dialogues collected for each model.

Following Yu et al. (2017), we also calculate *Information gain (InfoGain)*, the average number of unique tokens per dialogue and *Conversation Length (ConvLen)*, the average number of turns in the overall dialogue.

Finally, we compute the number of questions asked by the user (*UserQ)* as an indication of the user trust and engagement. We compute both the total number of questions presents in the 50 dialogue collected for a given model and the average number of question per dialogue.

2.5 Results and Discussion

We compare three models using automatic metric and absolute satisfaction scores. Based on this first evaluation, we compare two of these models using the Acute-Eval human evaluation framework. We display an example dialogue and discuss the respective use of each bot in the ComboBot model.

2.5.1 Automatic Evaluation and Absolute Satisfaction Scores

Table 2.3 shows the absolute satisfaction scores (i.e., scores provided based on a single dialogue rather than by comparing dialogues produced by different models) and the results of the automatic evaluation for the three models mentioned above.

ComBot provides a better basis for collecting information than MedBot The automatic scores show that ComboBot consistently outperforms MedBot on informativity (Slots, InfoGrain) and user engagement (UserQ) while allowing for shorter dialogues (ConvLen). In other words, ComboBot allows for a larger range of informational units (words and medical named entities) to be discussed in fewer turns.

[9] http://stanza.run/bio

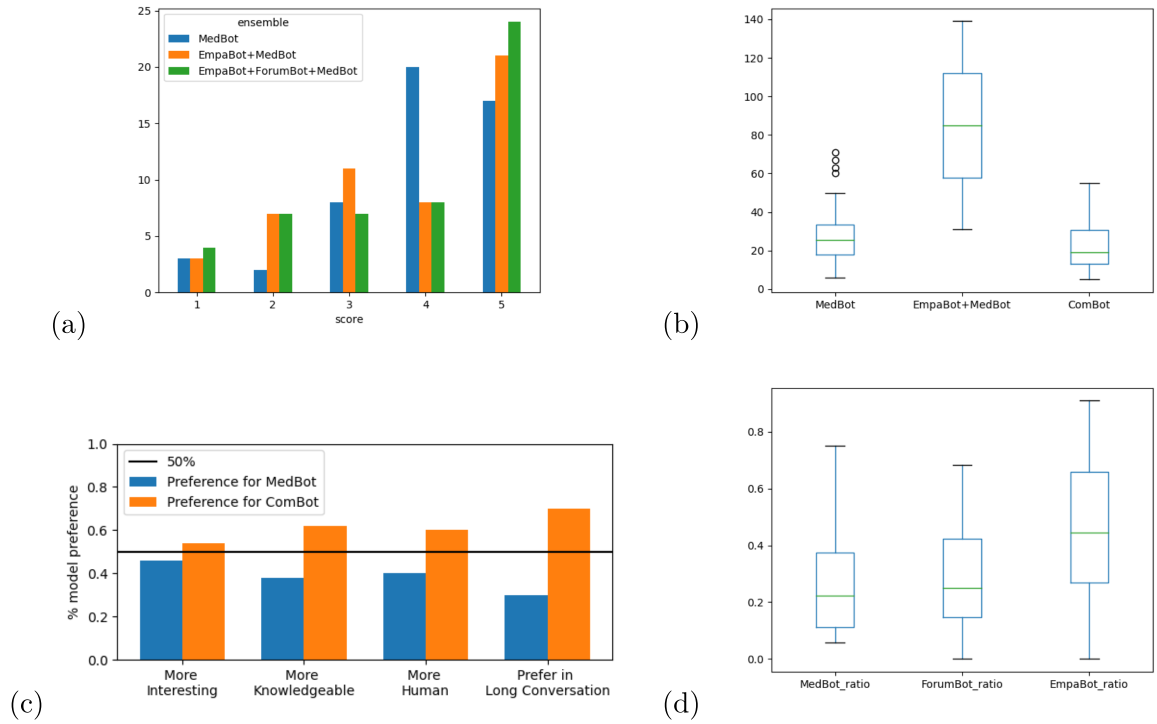


Figure 2.2: (a) Distribution of the Satisfaction Scores for each configuration, (b) Conversation length distribution for MedBot and ComBot, (c) Acute-Eval results for both systems, (d) Majority bot ratio

ComBot collects information in a more user-friendly, more efficient manner than Med+EmpathyBot While the InfoGain scores are higher for MED+EMPATHYBOT than for COMBOBOT (InfoGain: 170.5 vs. 126.66), this is achieved at the cost of much longer dialogues (ConvLen: 85 vs. 21.34; cf. also Figure 2.2b) and lower user engagement (UserQ: 0 vs. 0.45). In fact, when normalising InfoGain by the number of dialogue turns (ConvLen), we see that in average, a turn in COMBOBOT dialogues contains a much higher number (5.93) of unique tokens (i.e., is more informative) than both MEDBOT (3.89) and MED+EMPATHYBOT (2.00).

ComBot allows for more coherent dialogues In terms of quality, the differences in satisfaction scores between the three models is not statistically significant ($p < 0.05$, T-test). For dialogue coherence (measured by CoSim), however, COMBOBOT scores highest (0.36) and the difference with MEDBOT is statistically significant ($p < 0.05$, T-test). This suggests that follow-up questions help support smoother transitions between dialogue turns.

2.5.2 Comparative Human Evaluation

The results of the comparative human evaluation are presented in Figure 2.2.

ComBot is judged more knowledgeable, more interesting, more human and better for long conversations COMBOBOT outperforms MEDBOT on all Acute-Eval questions (Figure 2.2c).

In particular, users find COMBOBOT more knowledgeable by a large margin. This is in line with the automatic metrics results (higher COMBOBOT values for Slots and InfoGain) and is likely due to the fact that the COMBOBOT model supports the use of health-related, follow-up questions, which in turn allows for a wider range of medical issues to be discussed than just those present in the MedBot corpus.

Users also show a clear preference for ComboBot in long conversations (Figure 2.2a). While this seems to contradict the fact that both models have similar satisfaction score, we conjecture that the high MedBot satisfaction score is an artefact of the MedBot model. Since the MedBot coverage is restricted, the users have low expectations and correspondingly give high satisfaction scores (they are easily satisfied because their expectations are low). An indication of these low user expectations is given by the number of questions asked : when users feel that the system they interact with is unrestricted, they will feel comfortable asking questions and will start to do so. Conversely, if they feel the model is restricted, they will refrain from asking questions. The results show a much higher number of questions for users interacting with ComboBot (Table 2.3).

2.5.3 Component analysis

BOT: : *[opening question]* What would you like to share today?
USER: : i feel soo tired
BOT: : *[FollowUpBot]* Is something making you hyper at night?
USER: : maybe i am so stressed that i cannot sleep
BOT: : *[MedBot]* What wakes you up at night ?
USER: : nothing, i just cannot fall asleep
BOT: : *[FollowUpBot]* Are your sleep issues anxiety related?
USER: : i think so
BOT: : *[FollowUpBot]* Also do you snore when you do sleep possibly sleep apnea?
USER: : i do not know
BOT: : *[MedBot]* Do you feel anxiety or stress during nights awakenings ?
USER: : yes, a bit
BOT: : *[MedBot]* Anxiety and sleep may be connected, it is time to speak with your physician
USER: : okay
BOT: : *[FollowUpBot]* What about getting a second opinion from another doctor about your situation?.
USER: : why not
BOT: : *[FollowUpBot]* What does your doctor say about coming off of them and then trying to go back on.
USER: : i do not know
BOT: : *[MedBot]* Are you not able to fall asleep regularly or punctually ?
USER: : regularly
BOT: : *[EmpathyBot]* i hope things get better for you , and that you are able to enjoy your time with your family .
USER: : thank you
BOT: : *[EmpathyBot]* you are welcome , and i wish you the best of luck with the rest of your life . :)

Figure 2.3: Example of Human-ComBot dialogue

Figure 2.3 displays an example Human-Bot dialogue using the ComboBot model which illustrates the interactions between the three composing bots: the EmpathyBot closes the conversation with social chit-chat, the FollowUpBot response to the user turn and MedBot asks questions from the dialogue tree whenever suitable.

The proportion of turns generated by each bot (cf. Figure 2.2d) varies from one dialogue to another, illustrating the capacity of the ensemble model to adapt to various dialogue users and contexts. We find that in 55% of the collected dialogues, a majority of turns (i.e., more than 33% of the turns) is generated by the EmpathyBot model; in 29% of the cases by the FollowUpBot and in 16% of the cases by the MedBot[10]

[10]Since a ComboBot dialogue has an average of 21 turns and only half of those are generated by the bot,

We also observe interesting dependencies and correlations. MedBot is triggered twice more often after FollowUpBot (30 cases) than after EmpathyBot (12 cases) – this indicates that follow-up questions help to bring the user back to the questions contained in the dialogue tree. There is, furthermore, a weak positive correlation between the number of questions raised by the user (UserQ) and the ratio of FollowUpBot turns in the conversation (Pearson's r = 0.206) and conversely, a weak negative correlation between UserQ and the ratio of both MedBot (r= -0.112) and EmpathyBot (r = -0.068) turns in the dialogue. This is in line with the hypothesis that the FollowUpBot not only helps to gather health related information, but also increases user engagement.

2.6 Conclusion

A qualitative analysis of the collected dialogues indicates several directions for further research.

Negation is often not recognized, leading to interactions in which the model continues discussing a topic which was declared as irrelevant by the user.

Another difficulty knows when to end the conversation. Long ones are good to complete the task, but bad for people who are ready to finish the conversation but feel forced to continue.

To improve user engagement, a possibility would be to explore whether the information provided by sentiment analysers could be exploited to help maintain a positive interaction. By detecting polarity, it could also help improve negation handling.

Another key issue concerns the emotional impact of the dialogue on the user. An interaction with the bot might highlight a health issue the user was not aware of, resulting in increased user stress. In such a situation, a good policy would be to provide the user with some notion of solution, some piece of information or advice which can help her face the situation and if possible, incite her to act to improve her health. Indeed, some dialogues collected with ComboBot show that users sometimes ask for help. Here, a knowledge-based agent could be useful either to provide facts that are related to the topic at hand or to highlight the connections between facts that have been mentioned in the dialogue.

More generally, it would be interesting to investigate to what extent Prochaska and Velicer (1997); Petrocelli (2002)'s transtheoretical model for therapeutic change could be encoded in a dialogue model, i.e., whether we can develop bots which help the patient transition from a Contemplation to a Preparation stage which would help them to act to improve their health.

this means that for 55% of the collected dialogues, the dialogue contains more than 3 "social" dialogue turns (turns generated by EmpathyBot). Similarly, 29% of the collected dialogues contain more than 3 follow-up turns (FollowUpBot) and 16% more than 3 task-based turns (MedBot).

Part III

Questionnaire filling based on patient-bot interactions

Chapter 1

Exploring models for zero-shot application

Contents

In clinical studies, chatbots mimicking doctor-patient interactions are used for collecting information about the patient's health state. Later, this information needs to be processed and structured for the doctor. One way to organize it is by automatically filling the questionnaires from the human-bot conversation. It would help the doctor to spot the possible issues. Since there is no such dataset available for this task and its collection is costly and sensitive, we explore the capacities of state-of-the-art zero-shot models for question answering, textual inference, and text classification. We provide with a detailed analysis of the results and propose further directions for clinical questionnaire filling.

1.1 Introduction

Chatbots in healthcare can be used to collect information about the user with different purposes: treatment adherence, monitoring, patient support program, patient education, behaviour change, diagnosis (Car et al., 2020a).

Considering monitoring, patient-doctor conversations have mainly been used for the automation of medical records' creation through the extraction of clinical entities such as symptoms,

[0]This chapter was published as Ghassemi Toudeshki et al. (2021)

medications, and their properties (Du et al., 2019), generating reports (Finley et al., 2018) and summaries (Zhang et al., 2018c). Surprisingly, the task of filling clinical questionnaires received less attention.

Patients fill standard questionnaires during each medical visit, which frequency is usually several weeks. Performing this task in an automated way based on serendipitous talk (through a chatbot) opens the opportunity to get updated information more regularly and then monitor the patient more closely and in a seamless way.

Ren et al. (2020) were the first to introduce a questionnaire filling task as a classification problem, with the targets in the form of symptom phrases. Though, there are plenty of questionnaires that consist of full meaningful assertions or questions. The difference between questionnaire filling and slot filling is in the complexity of the questions that require machine reading comprehension (MRC) and question answering (QA).

The goal of a typical MRC task is to process a (set of) text passage(s) and then to answer questions about the passage(s). Though usually, multi-choice answer options are semantically different, in the case of questionnaires, the answers are often on the same scale (agree-disagree, often-rare).

More concretely, we make the following contributions:

- a clinical questionnaires' categorization based on questions and answers types
- data collection schema for filling questionnaires for 5 question types: open questions (OQ), closed questions (CQ), agreement Likert-scale (ALS), frequency Likert-scale (FLS) and visual analogue scale (VAS)
- analysis of question answering (QA), natural language inference (NLI), and zero-shot text classification (ZeroShot-TC) state-of-the-art models performance for the mentioned questions types

1.2 Related work

Four formats are commonly used for posing questions and answering them: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). UnifiedQA (Khashabi et al., 2020) is a single pre-trained QA model, that performs well across 20 QA datasets spanning 4 diverse formats.

Demszky et al. (2018) proposed a sentence transformation model which converts question-answer pairs into their declarative forms that allow to solve the task with NLI models. On the other hand, Yin et al. (2019) shows how ZeroShot-TC models may be successfully used as NLI models, being partially inspired by the way Obamuyide and Vlachos (2018) transformed relation classification task into NLI problem.

Questions in multiple-choice RACE (Lai et al., 2017) require holistic reasoning about the text, involving people's mental states and attitudes and the way they express them. Mishra et al. (2020) described reasoning categories present in RACE and showed that if the given passage is a dialogue, NLI model performs better than QA model.

Ren et al. (2020) presented a new medical Information Extraction task which uses questionnaires to convert medical text data into structured data. Their work is based on neural network classifier which makes selection among given options for closed-end type questions to fill out one complete questionnaire using only one model.

To our best knowledge, it is a first work to address the problem of filling in clinical questionnaires based on dialogue history. In Section 1.3 we are going to discuss the most common question types in questionnaires, including Likert scale which wasn't addressed before in clinical natural language processing. Also in Section 1.6 we show state-of-the-art models' ability to solve this task.

The task of filling questionnaires from user-bot dialogue history is a very specific subfield of MRC in NLP and as a result, reduces the chance of availability of such data for training/fine-tuning current models or later for evaluation, specially in medical domain. Both dialogues and answered questionnaires based on them are required that makes it difficult to collect such data on a large scale. This low resource setting is conducive to zero-shot approaches.

1.3 Question types

Question type	**Questionnaire**
Open Question (OQ)	Morin (Morin, 1993) QCD (R. C. Daut, 1983) MOS-SS (Sherbourne and Stewart, 1991)
Closed Question (CQ)	QPC (C. Thomas-Antérion, 2004) EPICES (Bihan et al., 2005)
Agreement Likert-scale (ALS)	TSK (D.D. Kori, 1990) PBPI (Williams and Thorn, 1989) TAS-20 (Bagby et al., 1994) LOT-R (Scheier et al., 1994) IEQ-CF (R. Sullivan, 2008) JCK (KARASEK, 1985)
Frequency Likert-scale (FLS)	MOS-SS (Sherbourne and Stewart, 1991) MBI (Maslach et al., 1997) PCL-S (Weathers et al., 1993) HAD (Zigmond and Snaith, 1983) SF-12 (Ware Jr et al., 1996)
Visual Analogue Scale (VAS)	QCD (R. C. Daut, 1983) Dallas (Lawlis et al., 1989) LEEDS (Parrott and Hindmarch, 2004) FABQ-W (Waddell et al., 1993)

Table 1.1: Clinical questionnaires and their types of questions

Khashabi et al. (2020) categorized existing QA datasets into four categories based on the format of answers: Extractive (EX), Abstractive (AB), Multiple-Choice (MC), and Yes/No (YN). In the following section we would like to introduce also Likert scale and Visual analogue scales (VAS) type of answers which is very common for clinical questionnaires. Sinha et al. (2017) came to the conclusion in their review that Visual analogue scale (VAS) and numerical rating scale (NRS) were the best adapted pain scales for pain measurement in endometriosis. Also, VAS is often used in epidemiologic and clinical research to measure the intensity or frequency of various symptoms (Paul-Dauphin, 1999).

For each question type present in the medical questionnaire, we provide examples of questionnaires and question in Table 1.1.

1.4 Task and Data

1.4.1 Task

Given a human-healthbot dialogue history D and a set of questions q_i extracted from a questionnaire Q, the task is to determine the correct answer a_i to each question q_i, including the 'not mentioned' option, i.e. the possibility that the dialogue does not address that question.

1.4.2 Chatbot

To create a chatbot for interactions, we followed the ComBot ensemble (Liednikova et al., 2021). The conversation always starts with the opening 'What is the most difficult for you about your sleep ?'. We made sure that there is no intersection between questions of the bot and questionnaire questions, due to the point of the research for answering questions implicitly addressed in the dialogue. Since creating a dialogue following the questionnaire topics is a very resource consuming task, we decided to mitigate the risk of the system proposing undesired direction of the conversation with a rejection function. If a bot reply isn't consistent with the dialogue history, the user can reject it and receive the next best candidate to continue the conversation.

1.4.3 Questionnaires

For experiments, we have chosen three questionnaires that are semantically close to the topics of the chatbot model: Morin (OQ), PBPI (ALS, VAS, CQ), Mos-ss (FLS). For PBPI questionnaire we ask annotators to give the answers in three format types at the same time: CQ, ALS, VAS. We provide statistics in Table 1.2. The full list of the questions could be found in Appendix, the answer options could be found in Tables 1.5 and 1.6.

1.4.4 Data collection

We asked 10 annotators to interact with the chatbot one time for each of the three questionnaires. So in total, we collected 30 dialogues and their corresponding question answers.

The annotators were first asked to read the questionnaire so that they could guide the interaction with the health bot, maximizing the number of questions addressed during the dialogue.

Then the annotators were asked to fill in the questionnaire based on their conversation and to select 'Not mentioned' (NA) option if the current question couldn't be answered from the dialogue history. For PBPI questionnaire we ask annotators to give the answers in three format types at the same time: CQ, ALS, VAS. You can find screenshot of interfaces in Appendix and answer options in Tables 1.5 and 1.6.

To ensure the reliability of collected data, we conducted a double annotation with adjudication. We ask two people to fill in the questionnaires based on the collected dialogues. So, totally for each question of each dialogue, we have three answers (one from the author of the dialogue and two from other annotators). Table 1.4 demonstrates final agreement between two annotators engaged in double annotation, as well as between two annotators and initial participants. In case of disagreement between annotators, the third person (adjudicator) decides the final label. These ground truth labels are used for evaluation of the models later.

During annotation, we came with some definition of classes that helped the annotators to come to agreement. We consider that the user would totally agree with the questionnaire statement if this statement or its paraphrase is explicitly mentioned in text, otherwise if there are phrases that fully or partially support this statement we annotate it with agree label. The same rules are applied to give disagreement and total disagreement label for contradictory statements.

Due to the high level of complexity of the task, we enrolled high-educated volunteers from our professional network. We made sure that participants were well aware of the context of the work, and they were all properly compensated. On average, performing a complete conversation with chatbot took 20 minutes from participants, and about 15 minutes for answering a questionnaire from a dialogue history.

Since the data and questionnaires were initially in French, we have translated them with DeepL [1] to English and to run the models.

Q Type	Questionnaire	Nb. of Q	Nb. of A
OQ	Morin	22	inf
CQ	PBPI	16	3
ALS	PBPI	16	5
FLS	Mos-ss	10	7
VAS	PBPI	16	11

Table 1.2: Questionnaires statistics: number of questions and number of answer options

	MOS-SS	Morin	PBPI
# turn	24.4	34.5	23.5
# tokens	259.1	403.2	269.2
# nb of tokens / turn	11.2	11.9	12.2
# Q answered	76%	59.3%	85.6%
# uniq tokens	140.8	201	147.6

Table 1.3: Statistics of dialogues for each questionnaire

OQ	**CQ**	**ALS**	**FLS**	**VAS**
two annotators				
0.81 (BertScore) 0.72 (Rouge-1)	0.81	0.70	0.67	0.31
two annotators + user				
0.75 (BertScore) 0.67 (Rouge-1)	0.67	0.47	0.56	0.20

Table 1.4: Inter-annotator agreement, using Kappa score for closed-ended questions and F1 score for Open question type

1.5 Models

In this section, we present the models that can be used in zero-shot setting with selected question types.

QA model UnifiedQA (Khashabi et al., 2020) [2] is a single pre-trained QA model, that performs well across 20 QA datasets spanning 4 diverse formats. Fine-tuning this pre-trained QA

[1] https://www.deepl.com/fr/translator
[2] https://github.com/allenai/unifiedqa

question type	choice	ratio
CQ	yes	55%
	no	30.6%
	NA	14.4%
ALS	totally agree	35%
	agree	19.4%
	rather disagree	11.8%
	totally disagree	19.4%
	NA	14.4%
FLS	all the time	5.6%
	most of the time	11.9%
	a good part of the time	5%
	sometimes	12.5%
	rarely	7.5%
	never	5%
	NA	15%
VAS	10	34.4%
	9	1.2%
	8	3.1%
	7	11.9%
	6	3.8%
	5	0%
	4	0.6%
	3	3.1%
	2	8.1%
	1	1.2%
	0	18.1%
	NA	14.4%

Table 1.5: Statistics of number of different choices for each question type

model into specialized models results in a new state of the art on 10 factoid and common-sense QA datasets, establishing UnifiedQA as a strong starting point for building QA systems. In our experiments, we used UnifiedQA-t5-3b version.

Applied to the following question types:

- **OQ** The dialogue history and question are provided in format of NarrativeQA dataset (Kočiský et al., 2018)
- **CQ and ALS** To fit format of MCTest dataset (Richardson et al., 2013), we transformed questionnaire statements to question form, by changing fist pronoun to second and adding "Do you agree with that" at the beginning. Otherwise, the model tends to choose NA option.
- **FLS** The dialogue history and question are provided in format of MC Test dataset (Richardson et al., 2013)

NLI model We use DeBERTa V2 xlarge model (He et al., 2020)[3] fine-tuned with MNLI dataset (Williams et al., 2018) for NLI task. We pass to the model concatenated dialogue history as premise and question in declarative form as hypothesis. The output is probabilities for three classes: Entailment, Contradiction, Neutral.

[3]https://github.com/microsoft/DeBERTa

Applied to the following question types:

- **CQ** A question is transformed into a statement and treated as hypothesis. Entering the premise and hypothesis to the model, we choose the class with the highest score as the final answer. Entailment class is considered as 'yes', contradiction as 'no' and neutral as 'NA'.

- **ALS** Premise is a dialogue history, hypothesis is a questionnaire statement. If probability for neutral class was higher than contradiction and entailment class, we consider that the dialogue doesn't contain relevant information to the question and give NA label. Otherwise, we take probability of Contradiction with negative sign and sum up it with probability of Entailment with positive sign. The resulting score lies in the interval of (-1,1). We uniformly divide this interval into N segments, where N is a number of options on Linkert-scale (usually 4 or 5).

- **FLS** For each question, we enter the model with dialogue as the premise and concatenation of one of the frequency scales with statement format of question (hypothesis = freq + question_statement_format) as hypothesis. We treat transformed question-answer combinations as distinct hypotheses, as presented in (Trivedi et al., 2019). Among frequency choices, we choose the one which has the highest entailment score. If none of them has the entailment score higher than 50%, we consider that the dialogue doesn't contain relevant information to the question and give NA label. Otherwise, we select the frequency scale with the highest entailment score.

- **VAS** If score for neutral is highest, then the selected output would be NA. Otherwise, we subtract the entailment score from contradiction. The result would be in range (-1,1). To map the result in range (0,10), we add value 1 to the subtraction result and then multiply it with 5 (shown in equation 1.1).

$$value = (ent - cont + 1) * 5 \tag{1.1}$$

where *ent* and *cont* are the predicted probabilities of entailment and contradiction classes.

ZeroShot-TC model We use Bart-large model (Lewis et al., 2020)[4] for zero-shot text classification trained on MNLI corpus (Williams et al., 2018). In this setting, we pass concatenated dialogue history as a context and formulate target labels as filled templates, so that the input data could be closer to entailment format.

Then we add one more target label 'NA' to consider the situation when the answer is not mentioned in the dialogue. The model provides probability scores for each candidate, and the candidate with the highest probability would be chosen as the final answer.

Applied to the following question types:

- **CQ** We transform a question into the statement beginning with "I agree that" for "yes" choice or with "I disagree that" for "no" choice.

- **ALS** The candidates are formed following the template "I *agreement_option* that *questionnaire statement*", where agreement options are totally disagree, disagree, agree, totally agree.

- **FLS** We transform a question into the statement and add in the beginning the frequency option to form the candidates.

[4] `https://huggingface.co/facebook/bart-large-mnli`

- **VAS** Candidates are defined as the same as ones in CQ. Model outputs probability scores for each class. By having these scores for each label and map them to entailment/contradiction/neutral labels, we continue transforming the results to VAS scale by doing the same process explained previously for VAS using NLI model.

Q type	Answers	NLI		ZeroShot-TC	
CQ	yes, no, NA	**Premise**	dialogue history	**Context**	dialogue history
		Hypothesis	There are times when I don't have pain.	**Targets**	***I agree that*** there are times when I don't have pain, ***I disagree that*** there are times when I don't have pain., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	**Model output**	I agree that there are times when I don't have pain.
		Final output	yes	**Final output**	yes
ALS	totally disagree, rather disagree, agree, totally agree, NA	**Premise**	dialogue history	**Context**	dialogue history
		Hypothesis	There are times when I don't have pain.	**Targets**	***I totally disagree that*** there are times when I don't have pain., ***I rather disagree that*** there are times when I don't have pain.,.., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	**Model output**	I totally disagree that there are times when I don't have pain.
		Final output	agree	**Final output**	totally disagree
FLS	all the time, most of the time, a good part of the time, sometimes, rarely, never, NA	**Premise**	dialogue history	**Context**	dialogue history
		Hypothesis	{freq_scale} I got the amount of sleep I needed.	**Targets**	***all the time*** I got the amount of sleep I needed., ***most of the time*** I got the amount of sleep I needed., .., NA
		Model output	entailment score for each freq. scale	**Model output**	rarely I got the amount of sleep I needed.
		Final output	choosing freq. scale which has the highest entailment score	**Final output**	rarely
VAS	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, NA	**Premise**	dialogue history	**Context**	dialogue history
		Hypothesis	I got the amount of sleep I needed.	**Targets**	***I agree that*** there are times when I don't have pain, ***I disagree that*** there are times when I don't have pain., NA
		Model output	entailment(0.5), contradiction(0.3), neutral(0.2)	**Model output**	I agree that there are times when I don't have pain.(0.5), I disagree that there are times when I don't have pain.(0.3), NA(0.2)
		Final output	6	**Final output**	6

Table 1.6: Examples of input and output for textual inference with Deberta and zero-shot classification with Bart-large for closed, agreement, frequency and VAS scale questions

1.6 Evaluation

Since for each question type we had different output, we used different evaluation scores.

For evaluating the performance of UnifiedQA on open-question types, we have used two common evaluation metrics: ROUGE (Lin, 2004) and BERTscore (Zhang et al., 2020a). ROUGE counts how many n-grams in the generated response matches the n-grams in the reference answer. Since we used abstractive mode of UnifiedQA, generated outputs might have different n-grams from reference ones. Therefore, using BERTscore might show a more realistic evaluation perspective since it computes the semantic similarity between generated and referenced answers based on embeddings.

We evaluate closed questions (CQ), agreement Likert scale (ALS), frequency Likert scale (FLS) and visual analogue scale (VAS) with macro and weighted F1 score.

Metric	**All**	**Answered**
ROUGE	0.38	0.63
BERT	0.55	0.93

Table 1.7: Scores for zero-shot evaluation of OQ type

The results are shown in the Table 1.7 for open-ended question type and in Table 1.8 for closed-ended ones. Because UnifiedQA generates answer even for unmentioned questions, we report the results for both all questions (mentioned and not mentioned questions) and for just mentioned ones.

ModelQuestion type	metric	CQ	ALS	FLS	VAS
Random (Baseline)		0.33	0.25	0.14	0.09
UnifiedQA-t5-3b	macro F1	0.44	0.13	**0.29**	
	weighted F1	0.58	0.12	**0.32**	
deberta-v2-xlarge-mnli	macro F1	0.417	**0.240**	0.158	**0.064**
	weighted F1	0.470	**0.262**	0.192	**0.104**
facebook/bart-large-mnli	macro F1	**0.484**	0.166	0.220	0.04
	weighted F1	**0.575**	0.136	0.262	0.03

Table 1.8: Scores for zero-shot evaluation for question types: CQ - closed question, ALS - agreement Likert-scale, FLS - frequency Likert-scale, VAS - Visual Analogue Scale

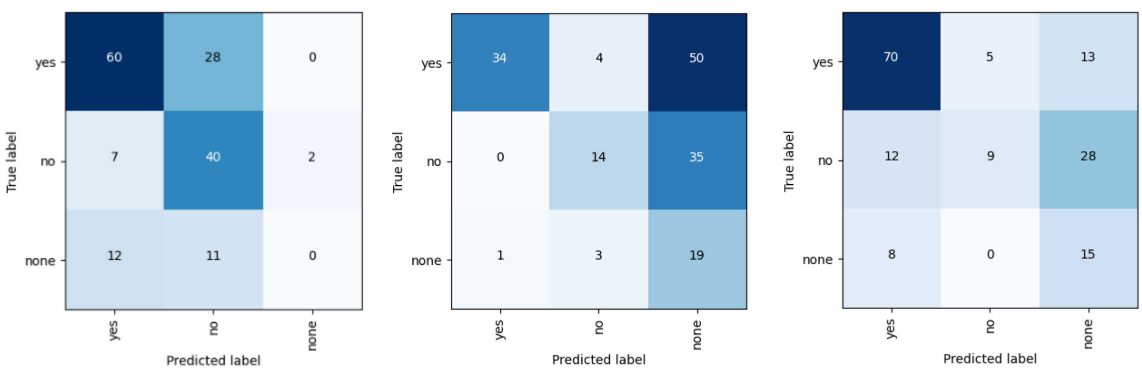


Figure 1.1: Confusion matrix for CQ using QA (left), NLI (center) and ZeroShot-TC (right) models

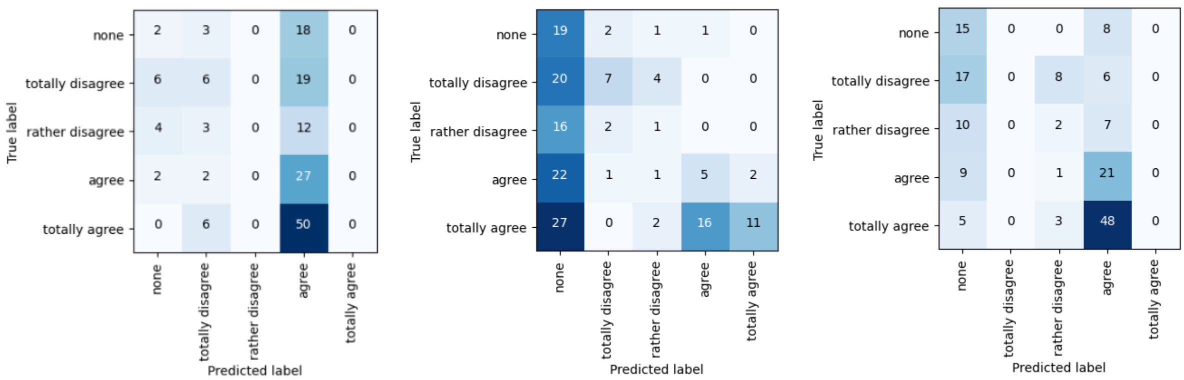


Figure 1.2: Confusion matrix for ALS using QA (left), NLI (center) and ZeroShot-TC (right) models

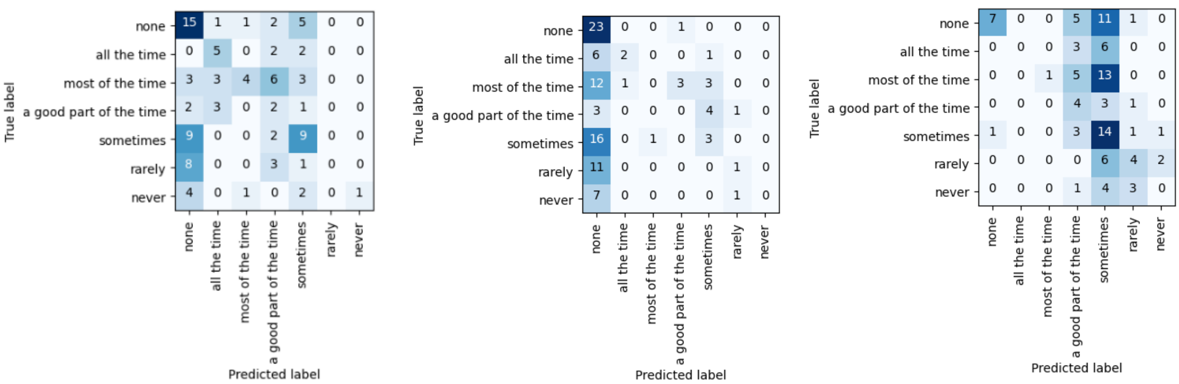


Figure 1.3: Confusion matrix for FLS using QA (left), NLI (center) and ZeroShot-TC (right) models

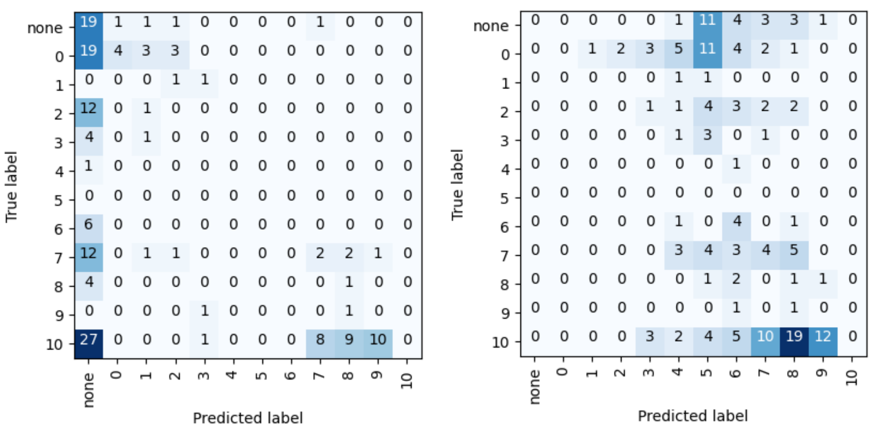


Figure 1.4: Confusion matrix for VAS using NLI (left) and ZeroShot-TC (right) model

1.7 Results and discussion

Detecting unanswerable questions Considering all open questions, we can see from Table 1.7 that the scores are considerably low, 0.38 for ROUGE and 0.55 for BERT. On the other hand, if we calculate these scores only for open questions being mentioned in the dialogue (according to the user answers), the scores are almost 2 times better. Results indicate the high performance of UnifiedQA model for answering mentioned questions and its lack to distinguish given the context if the question is answerable or not.

Impact of number of choices Table 1.8 shows the performance of SOTA NLI and ZeroShot-TC models for answering closed-ended question types (CQ, ALS, FLS, VAS). Comparing results for different question types can tell us that number of multiple-choices in each question type has a great impact on final results. Closed question type with 3 choices has the highest results and on the other hand, VAS with 11 choices has the lowest performance.

The table 1.8 also indicates the superiority of ZeroShot-TC for CQ and FLS questions types than NLI. After comparing confusion matrices provided for the NLI and ZeroShot-TC models, we can observe that the NLI model has a high tendency to give NA (neutral) class as output, while this is not the case for the ZeroShot-TC model.

On the other hand, the inability of ZeroShot-TC model to correctly predict extreme choices (totally agree/totally disagree) for ALS question type (Figure 1.2) has led to the lower performance of this model in comparison with NLI.

Measuring agreement is the most challenging task From Table 1.8 and Figures 1.2 and 1.4 we can see that predicting answers for agreement scale is the most challenging task. Since the answer option isn't semantically different enough to facilitate models choice, the probabilities for target classes don't help with selecting the correct level of agreement. In future work, we would like to explore other approaches, such as multi-hop reasoning and argument mining.

Importance of text input From our experiments, we can derive that the models are sensitive to the input text format and noise. Also, the different models are sensitive to the different data preprocessing technics. They may include using speaker tags, punctuation cleaning, selecting a subset of input text on some criteria. Such experiments should show what kind of preprocessing may boost performance without unnecessary data collection and training. These results may be a contribution to green and sustainable NLP.

1.8 Conclusion

In this chapter, we introduced the approaches for clinical questionnaire filling in a zero-shot setting for the five question types most used in clinical studies. Also, we described the data collection process for their evaluation. With the results, we show that this task is not easy to solve.

There is a type of questions which are more well-known in multi-hop reading comprehension. For such type of questions, there is a need to properly integrate multiple pieces of evidence to answer them. Song et al. (2018) investigates graph convolutional network and graph recurrent network to perform evidence integration.

As future work, we plan to take advantage of graph convolutional networks to improve the textual entailment for questionnaire filling. It would be possible due to enriching the model with knowledge graphs both in the open domain and close domain (medical).

Another direction to explore is text transformations. For example, we plan to transform statement-question into cloze form by masking the most important word and providing it as one option of the answers, use common-sense knowledge and graph structure for better reasoning.

Chapter 2

Exploring the Influence of dialogue Input Format for Unsupervised Clinical Questionnaire Filling

Contents

In the medical field, we have seen the emergence of health-bots that interact with patients to gather data and track their state. One of the downstream application is automatic questionnaire filling, where the content of the dialogue is used to automatically fill a pre-defined medical questionnaire. Previous work has shown that answering questions from the dialogue context can successfully be cast as a Natural Language Inference (NLI) task and therefore benefit from current pre-trained NLI models. However, NLI models have mostly been trained on text rather than dialogues, which may have an influence on their performance. In this chapter, we study the influence of content transformation and content selection on the questionnaire filling task. Our results demonstrate that dialogue pre-processing can significantly improve the performance of zero-shot questionnaire filling models which use health-bots dialogues as input.

[0]This chapter was published as Ghassemi Toudeshki et al. (2022)

dialogue
bot: What is the most difficult for you about your sleep ? **patient:** I have back pain that prevents me from sleeping. **bot:** I'm sorry to hear that. How long have you had back pain? **patient:** Since I've been working out, I've had constant back pain at night. **bot:** Do you think pain can last for long? **patient:** I think it will stop once I stop playing sports. **bot:** Should we let time fix the pain? **patient:** My doctor thinks that I need to get used to doing sports and that the pain will disappear after a while.
Questionnaire
(1) My pain is a temporary problem in my life. CQ: No **Yes** NA ALS: Totally disagree Rather disagree Agree **Totally agree** NA

Figure 2.1: An example of a dialogue and a question from the PBPI Questionnaire, answered in CQ and ALS format

2.1 Introduction

Work on Question Answering (QA) and Machine Reading Comprehension (MRC) mostly focuses on wh-questions of arbitrary types (who, what, where etc.) whose answer can be found in text. The answer can be extractive where a short span of the text is identified as the answer (Pearce et al., 2021) or it can be abstractive where a free-form answer is generated from the question and some support document (Bauer et al., 2018).

Here, we focus instead on a QA setting where questions are restricted to polar (yes/no) and Agreement Likert Scale (ALS) questions and where answers are contained in a dialogue rather than a paragraph text. As illustrated in Figure 2.1, this setting is useful for automatic questionnaire filling (AQF) in the medical field. Given a dialogue between a patient and a health bot, the goal of automatic questionnaire filling is to answer a set of predefined questions from a medical questionnaire (here the Pain Beliefs and Perceptions Inventory (PBPI) questionnaire (Williams and Thorn, 1989), based on the dialogue content.

In previous work, Ghassemi Toudeshki et al. (2021) compared three ways of deriving answers to questions from dialogues (Natural Language Inference, Question Answering and Text Classification) and showed that for automatic questionnaire filling in a medical setting, Natural Language Inference (NLI) models performed better on average on polar and ALS question types. One possible limitation of their approach however is that they apply NLI models to dialogues while NLI models are trained on non-dialogue text.

In this chapter, we propose different ways of transforming and selecting dialogue content before applying NLI to answer questions, and we analyse the impact of these operations on NLI-based questionnaire filling. Our hypothesis is that transforming the input dialogue into a format closer to the text format, on which NLI models are trained, should help these models perform better. Our experimental results confirm this hypothesis by demonstrating that, in a zero-shot setting, transforming and selecting dialogue content yields significant improvements over a baseline which takes the full dialogue content as input.

2.2 Related work

We briefly situate our work with respect to three tasks which have similarities with Automatic Questionnaire Filling, namely, Machine Reading Comprehension and Question Answering on the one hand and Aspect-Based Sentiment Analysis (ABSA) on the other.

MRC/QA Given a text and a question, MRC and QA models aim to derive the answer to that question from some input document (Zeng et al., 2020).

Similar to our approach, Ren et al. (2020) focus on filling in medical questionnaires consisting of polar questions about medical terms. However, in their case, the input to the model is text (patient records) rather than dialogue. Furthermore, QA is modelled as a classification task, which restricts the approach to a limited set of possible questions and answers. Finally, the questions are restricted to polar questions about terms, whereas we consider polar and ALS questions about full sentences.

Recently, some work has focused on answering questions from dialogues rather than text. A simple approach for modelling a multi-turn dialogue is to concatenate all turns (Zhang et al., 2020b; Adiwardana et al., 2020b). However, for retrieval-based response selection, Zhang et al. (2018d); Yuan et al. (2019) showed that turns-aware aggregation methods can achieve a better understanding of dialogues compared to considering all turns equally. Similarly, for MRC on dialogues, turns-aware approach have been proposed which select turns in the conversation that are related to the input question: Zhang et al. (2021b) uses embedding-based similarity to select such turns while Li et al. (2020) uses a pre-trained language model fine-tuned on NLI tasks. Their results showed that eliminating irrelevant turns effectively improves results. Our work extends on this work, showing that both content selection and content transformation help improve MRC on dialogues.

Aspect-Based Sentiment Analysis Aspect based sentiment analysis (ABSA) is the process of determining sentiment polarity for a specific aspect in a given context. An aspect term is generally a word or a phrase which describes an aspect of an entity (Jiang et al., 2019). For instance, Jang et al. (2021); Sun (2022) investigate aspect-based sentiment analysis on user tweets related to COVID-19. While AQF could be viewed as an ABSA task where each item should be labelled with one of three (polar question) or five (ALS question) sentiment value (agree, disagree, etc.), two key differences between ABSA and AQF is that (i) labels apply to sentences rather than aspect terms and (ii) contrary to these terms, the questions used in medical questionnaire can be very similar semantically (e.g., “Is your pain constant?” “Is your pain a temporary problem?”) making it harder to extract the correct answer from the input dialogue.

Closest to our work, Ghassemi Toudeshki et al. (2021) showed that pre-trained NLI models can be used to fill in questionnaires from dialogues in a zero-shot setting. We depart from their work in that we propose different ways of transforming and selecting dialogue content and investigate how this impact zero-shot, dialogue-based, automatic questionnaire filling.

2.3 Automatic Questionnaire Filling (AQF)

Task Given a dialogue D and a questionnaire Q, the Automatic Questionnaire Filling task consists in providing an answer a_i for each question $q_i \in Q$.

We address the task in a zero-shot setting (no training data). For evaluation, we provide a test set consisting of 100 dialogues and their associated questions and answers.

Questionnaire We consider two types of questions: Closed Questions (CQ) and Agreement Likert Scale (ALS) questions. CQ have three possible answers (yes, no or Not Applicable, i.e.

the dialogue does not address the question) and ALS has five (totally disagree, rather disagree, agree, totally agree, NA). As illustrated in Figure 2.1, questions are reformulated as declarative statements with multiple choice answers. With the emergence of health-bots, AQF can help

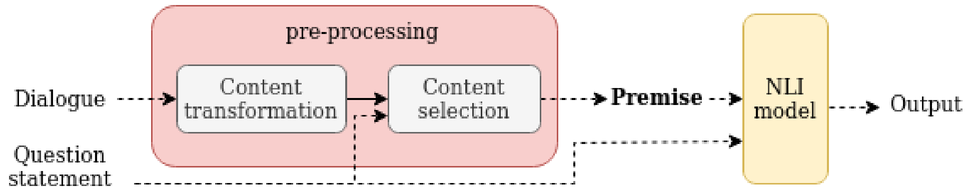


Figure 2.2: dialogue pre-processing schema

transform human-bot dialogues into structured data which can be used by physicians to track patients condition. In particular, it can be used to fill in questionnaires such as the Pain Beliefs and Perceptions Inventory (PBPI, (Williams and Thorn, 1989)) questionnaire, which includes 16 questions and is standardly used in the context of clinical studies. Collecting dialogues that include information for all of these questions is a difficult task, however. To facilitate data collection for the creation of the test set, we therefore decrease the number of questions by selecting five questions out of sixteen. Because the questions in the PBPI are often very similar, and knowing the answer to one of them allows deriving the answer to others, we chose questions that are semantically distinct from one another. The list of all PBPI questions is given in Appendix A and the five selected questions are indicated in bold.

Test Data To evaluate our approach, we create a test set of 100 dialogues and their associated question/answer pairs.

The creation of the test data involves first, collecting human-bot dialogues and second, extracting answers to the PBPI questions from the collected dialogues.

Collecting dialogues. We collect the dialogues using the Amazon Mechanical Turk platform and asking Turkers to interact with the ComBot health bot (Liednikova et al., 2021), while behaving as if they had chronic pain issues. To avoid Turkers introducing the PBPI questions verbatim in the dialogue, they were given a list of topics to be mentioned rather than the questions themselves (See details in Appendix C). In this way, we ensure that the collected dialogues address the questions to be answered while encouraging their diversified paraphrasing during the conversation. Turkers received bonuses each time they mention a key point. Turkers were also given the ability to modify the bot utterance in order to redirect the conversation more easily: they could reject the current candidate, in which case, the turn with the next highest confidences score would be displayed by the bot. We gathered dialogues for experiments using Amazon Mechanical Turk. Because of the task's difficulty and estimated completion time, we set the initial reward at 1$. We assigned 0.5$ bonus for each key point mentioned by the user during the dialogueue. If the user was successful in mentioning all five key points, he was awarded a bonus of 2.5$ in total. Details of the instructions given to the Turkers and a screenshot of the annotation interface are given in the Appendix C.

Identifying Question Answers. Two annotators with good English proficiency were asked to select the correct answer for each of the five selected questions based on each of the 100 collected dialogues. We computed agreement between the two annotators on all Q/A pairs and all 100 dialogues. The Kappa score is 0.94 for CQ and 0.86 for ALS question type. Thereafter, we used adjudication to decide on the final answer for all cases where the two annotators disagreed. The annotators were the authors of this work.

The final test corpus consists of 100 dialogues, each associated with 10 questions (5 yes/no questions and 5 ALS questions) and their answers. dialogue length varies from 4 to 70 turns and from 47 to 593 tokens, with 17.1 turns and 218.7 tokens on average.

2.4 Approach

Following Ghassemi Toudeshki et al. (2021), we model question answering as an NLI task where the premise is derived from the dialogue, the hypothesis from the question and the answer from the NLI result. Given a question and a dialogue, our model, illustrated in Figure 2.2, answers the question in three steps as follows.

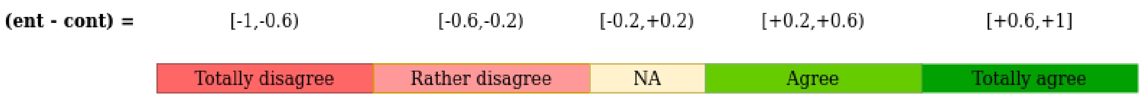


Figure 2.3: Map NLI scores to ALS answer types

Deriving an NLI Premise from the dialogue The NLI premise is derived from the input dialogue using first, Content Transformation and second, Content Selection. As detailed in Section 2.5, we experiment with different ways of transforming and selecting content.

Deriving an NLI hypothesis from a question To derive an NLI hypothesis from a question, we simply represent questions as statements (E.g., "I have pain regularly" instead of "Do you have pain regularly?"). Since the PBPI questionnaire questions are already in the form of a statement, we did not make any changes to them and used them as they are.

Deriving the answer We use RoBERTa large (Liu et al., 2019a)[1] fine-tuned on the MNLI dataset (Williams et al., 2018) to determine the entailment relation. We then derive the answer from the entailment relation between dialogue and question as follows.

For Close Questions, we set the answer to "Yes" if NLI returns an entailment, "No" if it returns a contradiction and "NA" if it returns "neutral".

For ALS questions, we map the NLI result to agreement choices as follows. If "neutral" has the highest score, the answer is "NA". Else, the contradiction score is subtracted from the entailment score. The subtraction result lies in a range of (-1,1) which is uniformly divided into 5 segments corresponding to the 5 ALS answer types, as shown in figure 2.3.

2.5 NLI-oriented dialogue Pre-processing

We consider different ways of transforming and selecting dialogue content.

We also study the impact of the NLI model used, comparing DeBERTa, the model used in Ghassemi Toudeshki et al. (2021), with RoBERTa (Liu et al., 2019a), the model used in our approach.

The DeBERTa model (He et al., 2020)[2] extends the BERT architecture with two innovative techniques: disentangled attention mechanism and an enhanced mask decoder. We compare AQF models with and without pre-processing and based on RoBERTa vs. DeBERTa, and find that whereas, when no pre-processing is applied, a DeBERTa model generally outperforms a RoBERTa-based model, the reverse is true when pre-processing is applied. This shows that

[1] https://huggingface.co/roberta-large-mnli

[2] https://github.com/microsoft/DeBERTa

while the improved DeBERTa-based, NLI model helps bridge the gap between dialogue and text, explicit pre-processing still yields better results.

2.5.1 Content transformation

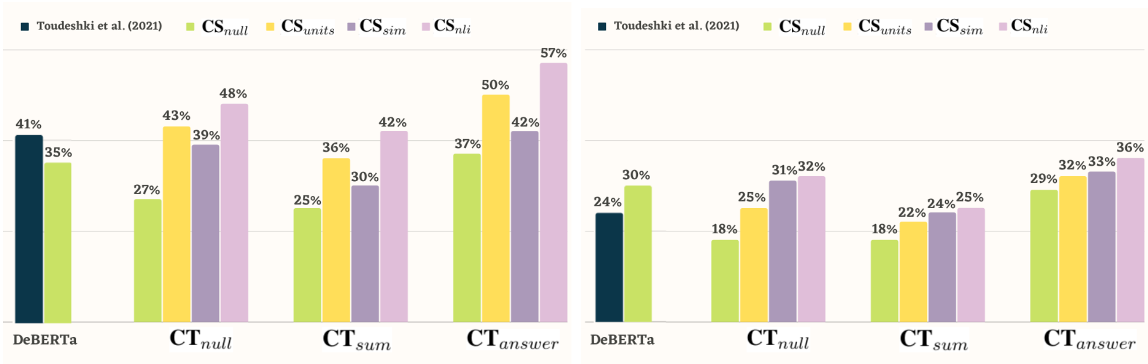


Figure 2.4: F1 macro average for Close Questions (on the left) and ALS questions (on the right) for the RoBERTa variant of our model. The two most left columns indicate the performance of Ghassemi Toudeshki et al. (2021)'s model on their (dark blue) and our (light green) test set. The best results are obtained by the CT_{answer}, CS_{nli} model.

Null Transformation (CT_{null}) A null transformation baseline where we simply concatenate the turns of the input dialogue. To encode the speaker information in each turn, the utterance is accompanied by the speaker role (patient/bot) at the beginning.

Summary (CT_{sum}) Pairs of adjacent turns are summarized, and the resulting summaries are concatenated. In this way, the input dialogue is transformed into a sequence of two-turn summaries. We also tried summarizing the whole dialogue in one go, but found that applying summarization on each two turns rather than on the whole dialogue gives better results. We use the **BART-large** model[3] (Lewis et al., 2020) fine-tuned on the News summarization corpus XSUM (Narayan et al., 2018) and on the dialogue summarization corpus SAMSum (Gliwa et al., 2019). The model achieves ROGUE-L score of 0.44 on SAMSum test set[4].

Long Answers (CT_{answer}) In information seeking dialogue, adjacent turns often are question-answer pairs. Based on this observation, we map each pair of adjacent turns in the dialogue into a single declarative sentence assuming that the first turn is a question (e.g., "Which drug did you take?"), the second is a short answer to that question (e.g., "Doliprane") and the sentence derived from the mapping is a long answer to the question (e.g., "I took Doliprane"). To learn this mapping, we fine-tune T5 (Raffel et al., 2019), a pre-trained encoder-decoder model, on two datasets of (question, incomplete answer, full answer) triples, one for wh- and one for yes-no (YN) questions. For wh-questions, we use 3,300 entries of the dataset consisting of (question, answer, declarative answer sentence) triples gathered by Demszky et al. (2018) using Amazon Mechanical Turk workers. For YN questions, we used SAMSum corpus (Gliwa et al., 2019), which contains short dialogues in chit-chat format. We created 1,100 (question, answer, full answer) triples by automatically extracting YN (question, answer) pairs from this corpus and manually associating them with the corresponding declarative answer. Data was split into train and test (9:1) and the fine-tuned model could achieve 0.90 ROUGE-L score on the test set.

[3] https://huggingface.co/Salesforce/bart-large-xsum-samsum

[4] https://paperswithcode.com/sota/abstractive-text-summarization-on-samsum

This fine-tuned model was applied to each two subsequent turns of the input dialogues, and the resulting declarative sentences were then concatenated to form the declarative transform of the whole dialogue.

2.5.2 Content selection

The transformation operations described in the previous section yield sequences of dialogue turns, two-turn summaries or full answers. We call these "input units" and consider three ways of pre-selecting the input units that will be used as premise when testing for entailment.

Null Content Selection (CS_{null}) A null content selection baseline where the premise is the concatenation of all the input units produced by the content transformation operations (dialogue turns, sequence of two turn summaries, sequence of full form answers).

Unit-Based (CS_{units}) Each question is assessed against each input item. Given an input sequence I_n of length n, the answer a_i to a question q is then determined by aggregating the resulting entailment probabilities as follows:

- $a_i = NA$ if for all input items $i \in I_n$, the NA probability is highest.
- $a_i = Yes$ (resp. $a_i = No$) if for at least one item $i \in I_n$, the Yes (resp. No) probability is highest and the highest Yes (resp. No) probability is higher than the highest No (resp. Yes) probability.

Similarity (CS_{sim}) For each question q, we select a subset of input units that are semantically similar to q. We encode question and input units using SBERT[5] (Reimers and Gurevych, 2019) and compute *cosine similarity* for each (q, input unit) pair. We then select items whose similarity score is higher than 0.5, concatenate them, and use the result as the NLI premise.

NLI (CS_{nli}) For each question q in the questionnaire, we select the input units that are related to q using the NLI model (RoBERTa-Large). Specifically, we select sentences which have an entailment or contradiction score higher than 0.5. All selected sentences are then concatenated to form the NLI premise.

2.5.3 Baseline and Comparison

Our baseline is the null method (CT_{null}+CS_{null}) i.e., the approach where question answering applies to the untransformed, unfiltered dialogue. To compare our approach with Ghassemi Toudeshki et al. (2021), we also report the performance of their model on both their test set (10 dialogues) and on ours (100 dialogues).

2.6 Results

The experiments were conducted with a laptop having Intel® Core™ i7-10610U CPU @ 1.80GHz * 8 and NVIDIA Quadro P520. We evaluate our approach using macro and weighted F1 score.

2.6.1 How much does pre-processing help improve performance ?

Figure 2.4 shows the results for all combinations of our content transformation and selection methods[6].

[5] https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v2

[6] We first focus on the results of our RoBERTa based model and delay the comparison with DeBERTa based models to Section 2.6.4.

Two turns
bot: do you feel anxiety or stress during nights awakenings ? **patient:** I feel stressed during night awakenings although I am not feeling guilty about being in pain.
Generated summary Patient feels stressed during night awakenings although **he's not in pain**.

Figure 2.5: An example of the summarization model performance on two subsequent turns, showing missing and **inconsistent** information in the output summary

Improvement over the baseline Comparing our best model (CT_{answer}, CS_{nli}) with the no-preprocessing CT_{null}, CS_{null} baseline, we see (Figure 2.4) that pre-processing can multiply the macro and weighted F1 scores by two. The best pre-processing method combines a question+answer to sentence transformation (CT_{answer}) with the entailment-based content selection method (CS_{nli}).

Content transformation The CT_{answer} question+answer transform, which merges pairs of adjacent dialogue turns into declarative statements, consistently yields the best results. A possible explanation is that this transform yields an input, a declarative sentence, which is consistent with the format of the training data used for NLI models.

Conversely, summarization (CT_{sum}) has the lowest performance. This could be due to errors such as hallucinations or omissions known to be produced by summarization systems (Zhao et al., 2020). Figure 2.5 shows an example of such errors when applying the CS_{sum} transformation.

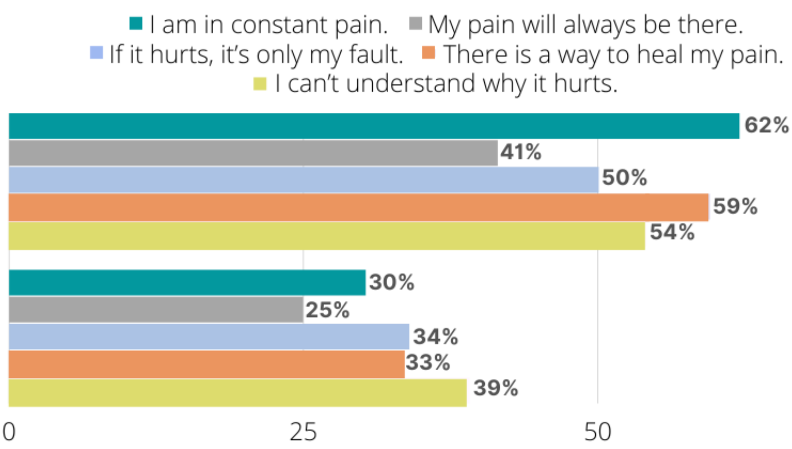


Figure 2.6: Break down of F1 macro average scores for each question based on out-performed model ($CT_{answer} + CS_{nli}$) results

Content selection The NLI-based content selection method (CS_{nli}) consistently outperforms other content selection approaches. This is consistent with Ghassemi Toudeshki et al. (2021)'s findings that for automatic questionnaire filling in a medical setting, NLI models performed better on average on polar and ALS question types.

We also see that the second best performing content selection method varies depending on the question type. As CS_{unit} first filters question/item pairs with the highest probability, the

	CQ					ALS						
	NA	YES	NO	macro	weighted	NA	TD	RD	A	TA	macro	weighted
support	142	228	130			142	54	79	115	110		
CT_{null}												
CS_{null}	0.39	0.15	0.27	0.27	0.25	0.28	0.11	0.26	0.07	0.16	0.18	0.18
CS_{units}	0.33	0.48	0.46	0.43	0.43	0.33	0.25	0.02	0.07	0.58	0.25	0.25
CS_{sim}	0.52	0.55	0.10	0.39	0.42	0.54	0.07	0.09	**0.23**	0.60	0.31	0.36
CS_{nli}	0.34	0.60	0.48	0.48	0.50	0.34	0.29	0.08	0.21	**0.67**	0.32	0.34
CT_{sum}												
CS_{null}	0.41	0.11	0.23	0.25	0.23	0.36	0.12	0.21	0.11	0.10	0.18	0.20
CS_{units}	0.32	0.33	0.43	0.36	0.35	0.32	0.23	0.06	0.02	0.44	0.22	0.23
CS_{sim}	0.49	0.40	0.02	0.30	0.32	0.51	0.00	0.05	0.21	0.46	0.24	0.30
CS_{nli}	0.37	0.43	0.46	0.42	0.42	0.31	0.28	0.10	0.09	0.48	0.25	0.26
CT_{answer}												
CS_{null}	0.45	0.28	0.37	0.37	0.35	0.41	0.27	**0.27**	0.17	0.33	0.29	0.30
CS_{units}	0.40	0.59	0.51	0.50	0.52	0.41	0.29	0.17	0.16	0.57	0.32	0.33
CS_{sim}	**0.53**	0.60	0.13	0.42	0.46	**0.55**	0.10	0.20	**0.23**	0.59	0.33	**0.38**
CS_{nli}	0.45	**0.70**	**0.57**	**0.57**	**0.59**	0.42	**0.35**	0.16	**0.23**	0.65	**0.36**	**0.38**

Table 2.1: F1-Scores for RoBERTa for closed (CQ) and agreement Likert scale (ALS) question types; TD - totally disagree, RD - rather disagree, A - agree, TA - totally agree. CT: content transformation, CS: content selection.

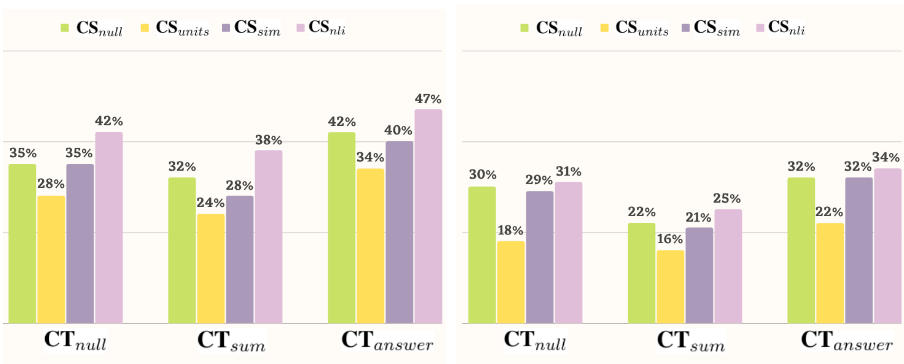


Figure 2.7: F1 macro average for the DeBERTa variant of our model on Closed Questions (CQ) on the left and Agreement Likert Scale (ALS) on the right. Test set of 100 dialogues with 10 questions each (5 yes/no questions and 5 ALS questions).

method works well on CQ questions but struggles to handle more nuanced ALS questions, which leads to an overall drop in performance on ALS questions.

2.6.2 Impact of pre-processing on different question/answer types

Table 2.1 shows the results for all combinations of pre-processing steps for each question/answer type.

Agreement answers (Yes, Totally agree) have the highest accuracy (about 70% in the best case) in both CQ and ALS questions, which suggests that the NLI model is better at confirming rather than rejecting a statement.

On *CQ questions*, various content selection methods have different impacts on each answer type. CS_{sim} shows much lower (3-4 times lower) performance on 'No' class than on 'NA' or

	CQ					ALS						
	NA	YES	NO	macro	weighted	NA	TD	RD	A	TA	macro	weighted
support	142	228	130			142	54	79	115	110		
CT_{null}												
CS_{null}	0.43	0.33	0.31	0.35	0.35	0.41	0.23	0.19	0.22	0.47	0.30	0.32
CS_{units}	0.15	0.29	0.40	0.28	0.28	0.15	0.21	0.07	0.00	0.45	0.18	0.17
CS_{sim}	0.51	0.45	0.09	0.35	0.37	**0.54**	0.03	0.05	**0.24**	0.60	0.29	0.35
CS_{nli}	0.34	0.51	0.40	0.42	0.43	0.29	0.23	0.17	0.21	**0.63**	0.31	0.32
CT_{sum}												
CS_{null}	0.40	0.29	0.26	0.32	0.31	0.37	0.16	0.11	0.16	0.31	0.22	0.25
CS_{units}	0.20	0.18	0.33	0.24	0.23	0.20	0.17	0.11	0.05	0.26	0.16	0.16
CS_{sim}	0.48	0.34	0.01	0.28	0.29	0.49	0.00	0.00	0.17	0.40	0.21	0.27
CS_{nli}	0.39	0.39	0.35	0.38	0.38	0.37	0.20	0.07	0.15	0.45	0.25	0.27
CT_{answer}												
CS_{null}	0.44	0.48	0.35	0.42	0.43	0.43	0.26	0.14	0.19	0.57	0.32	0.34
CS_{units}	0.19	0.45	0.39	0.34	0.36	0.19	0.20	0.07	0.10	0.55	0.22	0.23
CS_{sim}	**0.52**	0.51	0.16	0.40	0.42	0.53	0.09	0.15	0.20	0.61	0.32	**0.37**
CS_{nli}	0.40	**0.60**	**0.42**	**0.47**	**0.50**	0.41	**0.30**	**0.21**	0.16	**0.63**	**0.34**	0.36

Table 2.2: F1-Scores for DeBERTa for closed (CQ) and agreement Likert scale (ALS) question types; TD - totally disagree, RD - rather disagree, A - agree, TA - totally agree. CT: content transformation, CS: content selection.

'Yes', CS_{null} has higher accuracy for the 'NA' class than for 'Yes' or 'No' classes and CS_{nli} performs better on 'Yes' and 'No' answers than on 'NA'. Both CS_{nli} and CS_{units} gives the most balanced F1 distribution across classes.

For *ALS questions*, CS_{nli} and CS_{sim} show the best results. While the CS_{nli} model is best at identifying 'Totally agree' and 'Totally disagree' classes, CS_{sim} distinguishes well whether the answer is absent ('NA') or whether it belongs to the 'Totally agree' class.

Performance on ALS questions is always lower. This can be explained by choice of threshold that distinguishes classes 'Totally agree' and 'Agree' as well as 'Totally disagree' and 'Rather disagree'. As mentioned above, CS_{units} favourites the extreme classes, which leads to a higher performance drop in comparison with CS_{sim} on ALS.

2.6.3 Break down of results for each question

Figure 2.6 presents the results of our best model (CT_{answer}+CS_{nli}) for each PBPI question separately.

The question "I am in constant pain." obtains the highest score in CQ, while it performs poorly in ALS, demonstrating that the model is effective at detecting the presence of consistent pain but bad at predicting the level of agreement. The same behaviour can be seen for the question "There is a way to heal my pain". On the other hand, for question "My pain will always be there" gets lowest score for both question types. The presence of the term "always" in the question turns it into a strong statement, and consequently the model mostly rejects the statement unless it has been explicitly mentioned in the dialogue.

2.6.4 Comparison with previous work and a different classifier (RoBERTa vs. DeBERTa)

Our model differs from previous work by Ghassemi Toudeshki et al. (2021) in two ways: it includes a pre-processing phase and uses the RoBERTa classifier, whereas Ghassemi Toudeshki et al. (2021) applies DeBERTa to the whole input dialogue. We compare our model with (i) the same model using DeBERTa and (ii) Ghassemi Toudeshki et al. (2021)'s model both on their and our test set.

Comparison with previous work In Figure 2.4, the two columns on the far left show the performance of Ghassemi Toudeshki et al. (2021)'s model on two test sets: the test set they used (10 instances and 16 questions) and our test set (100 instances and 5 questions).

Unsurprisingly, Ghassemi Toudeshki et al. (2021)'s results vary with the test set: while they report F1 score of 41 for CQ and 24 for ALS questions on their test set, these change to 35 and 30 on ours.

We also see that Ghassemi Toudeshki et al. (2021)'s DeBERTa-based, no pre-processing model out-performs our RoBERTa-based, null-preprocessing model (CT_{null}, CS_{null}) on both test sets. We conjecture that this difference can be explained by DeBERTa's improved attention mechanism, which selects relevant information in the input dialogue with respect to the hypothesis.

However, our best model outperforms Ghassemi Toudeshki et al. (2021)'s approach by 22 points F1 for CQ questions and 6 points for ALS questions, which indicates that pre-processing better helps bridge the gap between dialogue and NLI-based QA.

DeBERTa vs. RoBERTa figure 2.7 and Table 2.2 show the result of our model when using DeBERTa instead of RoBERTa.

When using pre-processing, we see that the best RoBERTa model (CT_{answer}, CS_{nli}) outperforms the best DeBERTa model by 10 points F1 for CQ questions and 2 points for ALS questions.

Conversely, when no pre-processing is applied, the DeBERTa variant of our model outperforms the RoBERTa variant, which is consistent with the results discussed in the previous paragraph. For the DeBERTa variant, we observe that the CS_{null} baseline is no longer the lowest performing content selection approach, while the performance of CS_{units} and CS_{sim} becomes lower than the baseline (CS_{null}). This highlights the fact that the DeBERTa model performs better without weak content selection approaches. On the other hand, it can be seen that the impact of content selection and transformation approaches is significant in RoBERTa, although using a weaker classifier, and our model outperforms previous work. This shows that the proposed select-and-transform pre-processing approach improves results in both RoBERTa and DeBERTa, though this improvement is more significant in RoBERTa, suggesting that this latter model is more sensitive to the form and size of the input content.

2.7 Conclusion

In this chapter, we studied how dialogue pre-processing can impact the task of filling medical questionnaires based on patient-bot interactions. Our experimental results show that converting pairs of adjacent turns to declarative sentences and selecting input units based on their entailment relation with the question can significantly enhance performance.

Conclusion

The research presented in this thesis makes a first step towards developing a solution that can complement visit and questionnaire routine between doctor and chronic pain patient – this is a main mission of ALIAE company. It consists of two main parts: a dialogue system and automatic questionnaire filling. A dialogue system is dedicated to maintain a contact and conversation with a patient between the visits, while collecting important information about their health. Later, this dialogue history may be used to automatically fill the questionnaire without explicitly asking the same questions over and over again. In such a way, natural and serendipitous talk with a conversational agent helps to escape from routine when the patient is getting used to give repetitive answers to the same questionnaires.

First, as started, we faced the data creation problem that is a common bottleneck for developing a health bot. Using paraphrase identification techniques and a medical dialogue tree of doctor-patient interactions created by the experts, we proposed a novel method to create training data for dialogue models. Later, we used this data to learn health chatbots that cover the main topics standardly used in the questionnaires of clinical studies. We wanted the model to follow a pre-defined sequence as long as possible, while giving an opportunity to switch to a topic raised by the user. We compared two models, a generative and a hybrid classification/retrieval model, and we showed that the expert knowledge captured by the dialogue tree both helps guide learning and facilitate error analysis.

However, during human evaluation, we found out that the users, given freedom to express themselves, tend to be social and deviate from presented topics. A common approach to address this issue is to develop a dialogue system based on ensemble of task-oriented and chit-chat model. Our experiments showed, that such combination improved engagement, but decreased efficiency of the health bot. The natural solution to this problem was to introduce a new type of bot that combines medical topics and small talk by asking follow-up questions based on context-response dataset. Ablation studies showed that an ensemble of three bots gives the best performance and balance between bot efficiency and engagement.

Finally, we introduced the approaches for clinical questionnaire filling in a zero-shot setting for the five question types most used in clinical studies: open, closed, agreement Likert scale, frequency Likert scale and visual analogue scale. Also, we describe the data collection process for dialogues and questionnaires for model evaluation. Experimental results showed us that this task is not an easy one to solve. Later, we focused our research on closed and agreement scale question type and studied how dialogue pre-processing can impact the performance of natural language inference model that previously showed the best results on these question types. Experiments demonstrated that converting pairs of adjacent turns to declarative sentences and selecting input units based on their entailment relation with the question can significantly enhance performance of filling the questionnaire with closed or agreement scale questions.

The results of research presented in this thesis have been implemented as prototypes and were the basis for product development by ALIAE company for different projects.

Future work

Since this thesis propose the first steps in development of prototype, the future work can be concerned about improving each part of the system or the whole system in general.

Based on results presented in Chapter 1, we derived three main directions for future research for improving dialogue agent based on medical decision tree. First, additional paraphrase techniques could be explored to create a more balanced dataset, for example paraphrase generation rather than paraphrase extraction. Second, longer, richer dialogues could be obtained by extending the expert dialogue tree or using existing ones, like on the American Medical Association

Family Medical Guide (Kunz, 1982). Third, existing medical tree can be expanded with more advices and descriptions of patient state that might result in a new dataset with longer and more precise interaction between doctor and patient for further model training.

A qualitative analysis of the collected dialogues with ensemble of the bots in Chapter 2 indicates multiple problems that should be addressed in the future. Negation is often not recognized, leading to interactions in which the model continues discussing a topic which was declared as irrelevant by the user. Another difficulty is to know a right time to end the conversation. Long ones are good to complete the task, but bad for people who are ready to finish the conversation but feel forced to continue. One of the possible ways to improve user engagement would be to explore whether the information provided by sentiment analysers could be exploited to help maintain a positive interaction as well as to improve negation handling. Another key issue concerns the emotional impact of the dialogue on the user. An interaction with the bot might highlight a health issue the user was not aware of, resulting in increased user stress level. In such a situation, a good policy would be to provide the user with some notion of solution, some piece of information or advice which can help her to face the situation and if possible, incite her to act to improve her health. Indeed, some dialogues collected with ComboBot show that users sometimes do ask for help. Here, a knowledge-based agent could be useful either to provide facts that are related to the topic at hand or to highlight the connections between facts that have been mentioned in the dialogue.

Automated questionnaire filling results presented in Chapters 1 and 2 made us encounter the need for multi-hop reading comprehension, when multiple pieces of evidence can be properly integrated to answer scale type of the questions. Song et al. (2018) investigates graph convolutional network and graph recurrent network to perform evidence integration. Also, it is possible to take advantage of graph convolutional networks to improve the textual entailment for questionnaire filling. It would be possible due to enriching the model with knowledge graphs both in the open domain and close domain (medical).

If we take the dialogue system as a whole, one of the important future directions would be to move from one-session dialogue to multi-session one. It means, that we need to handle how information of patient is dynamically stored and used for moving conversation forward. Another, important task to address, transforming results from automated questionnaire filling to conversation agenda for the next session. It is highly probable that the questionnaire will not be filled from the first session, since we do not want to force the user, so we want to guide the next conversation in such a way that will improve our chances to get all the answers. Though, the optimal solution would be to change direction of the dialogue on flow, it is very challenging due to resource-intensive calculation, that will decrease dramatically response time which is crucial for production environment.

Till now, we were talking about a stable system that can be used by all users in more or less the same manner. Dialogue system can be developed as a model-based reinforcement learning framework, that starts from a base model and later adapts to the user behaviour. It can be seen in several dimensions. First, we can train a model to learn the best sequence of questions to arrive to a filled questionnaire with minimum effort and maximum user satisfaction. Secondly, we can train a model to adapt to user style and needs following, for example, so-called Big Five personality dimensions (de Raad, 2000). One of the future directions in this topic mentioned by Moilanen et al. (2022) is to see how the user seeking or considering mental health help using self-help methods perceives these personalities changes.

Looking really ahead, it would be also interesting to investigate to what extent Prochaska and Velicer (1997); Petrocelli (2002)'s transtheoretical model (TTM) for therapeutic change could be encoded into a dialogue model. The main idea of TTM is that there is a series of stages rather than a single event to be accomplished before the behaviour change occurs. Intention to take action is developed during the first three stages of introspective cognitions: precontemplation, contemplation, action. During the next two ones, preparation and maintenance, the intent to change is fully actualized and behaviour changed can be observed. We claim that by language it could be possible to identify the stage of the patient concerning a particular belief. In this case, it could be possible to develop a dialogue system which help the patient to transition from a Contemplation to a Preparation stage. This can result in helping them to better understand their state and to seek doctor help in advance in order to overcome dysfunctional behaviour and improve quality of life with chronic conditions.

Appendix A

Questionnaire

PBPI questionnaire statements are provided in table A.1.

Id	Question
1	No one is able to tell me why it hurts.
2	I thought my pain could be healed, but now I'm not so sure.
3	There are times when it doesn't hurt.
4	My pain is difficult for me to understand.
5	**My pain will always be there.**
6	**I am in constant pain.**
7	**If it hurts, it's only my fault.**
8	I don't have enough information about my pain.
9	My pain is a temporary problem in my life.
10	I feel like I wake up with pain and fall asleep with it.
11	I am the cause of my pain.
12	**There is a way to heal my pain.**
13	I blame myself when it hurts.
14	**I can't understand why it hurts.**
15	One day, again, I won't have any pain at all.
16	My pain varies in intensity but it is always present with me.

Table A.1: List of questions in PBPI questionnaire

nb.	Question
1	What time do you usually go to bed on weekday evenings?
2	What time is your final wake-up call in the morning?
3	What time do you usually get up during the week?
4	What time do you go to bed on your days off?
5	What time do you get up on your days off?
6	How soon do you fall asleep after turning off the light?
7	How many times a night do you wake up on average?
8	How long do you spend waking up between your first sleep and your final awakening?
9	How many hours per night do you sleep on average?
10	Why do you wake up at night? (pain, noise, child, nightmare, spontaneous awakening, others)
11	What medicine are you taking or were you taking and at what dose?
12	How many nights per week do you currently take the medicine?
13	When did you start taking the medicine?
14	When was the last time you took the medicine?
15	How long have you suffered from insomnia?
16	When was the first time you had trouble sleeping?
17	Did your insomnia start gradually or suddenly?
18	Were there any stressful events that could be linked to the onset of your insomnia? (death, divorce, retirement, family or professional problem, ..)
19	How many times a week do you exercise?
20	How much coffee, tea or Coca-Cola do you consume per day?
21	How many cigarettes do you smoke per day?
22	How many glasses of beer, wine or alcohol do you drink per day?

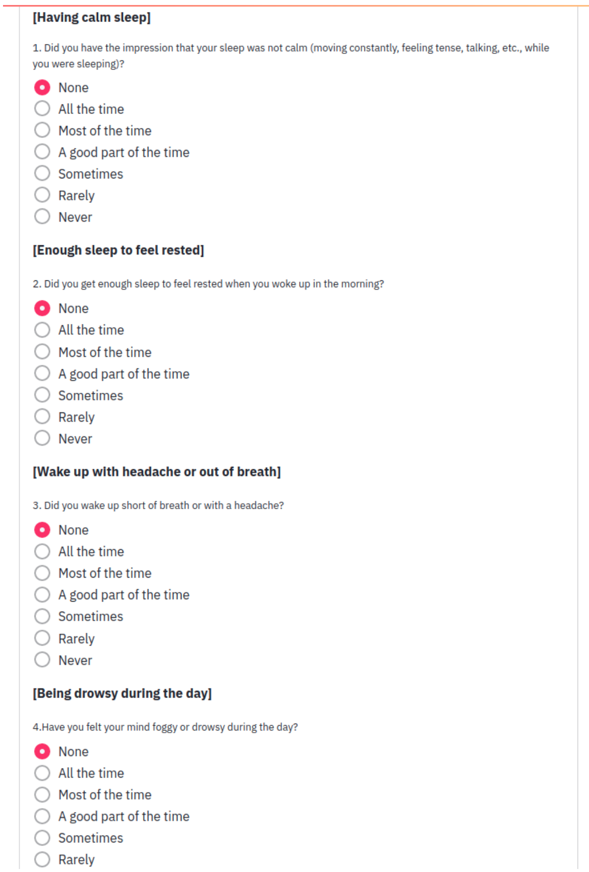
Table A.2: List of questions in Morin questionnaire

nb.	Question
1	No one is able to tell me why it hurts.
2	I thought my pain could be healed, but now I'm not so sure.
3	There are times when it doesn't hurt.
4	My pain is difficult for me to understand.
5	My pain will always be there.
6	I am in constant pain.
7	If it hurts, it's only my fault.
8	I don't have enough information about my pain.
9	My pain is a temporary problem in my life.
10	I feel like I wake up with pain and fall asleep with it.
11	I am the cause of my pain.
12	There is a way to heal my pain.
13	I blame myself when it hurts.
14	I can't understand why it hurts.
15	One day, again, I won't have any pain at all.
16	My pain varies in intensity but it is always present with me.

Table A.3: List of questions in PBPI questionnaire

Appendix B

User interface for annotation in Chaper 3. 1



[Having calm sleep]

1. Did you have the impression that your sleep was not calm (moving constantly, feeling tense, talking, etc., while you were sleeping)?

- None
- All the time
- Most of the time
- A good part of the time
- Sometimes
- Rarely
- Never

[Enough sleep to feel rested]

2. Did you get enough sleep to feel rested when you woke up in the morning?

- None
- All the time
- Most of the time
- A good part of the time
- Sometimes
- Rarely
- Never

[Wake up with headache or out of breath]

3. Did you wake up short of breath or with a headache?

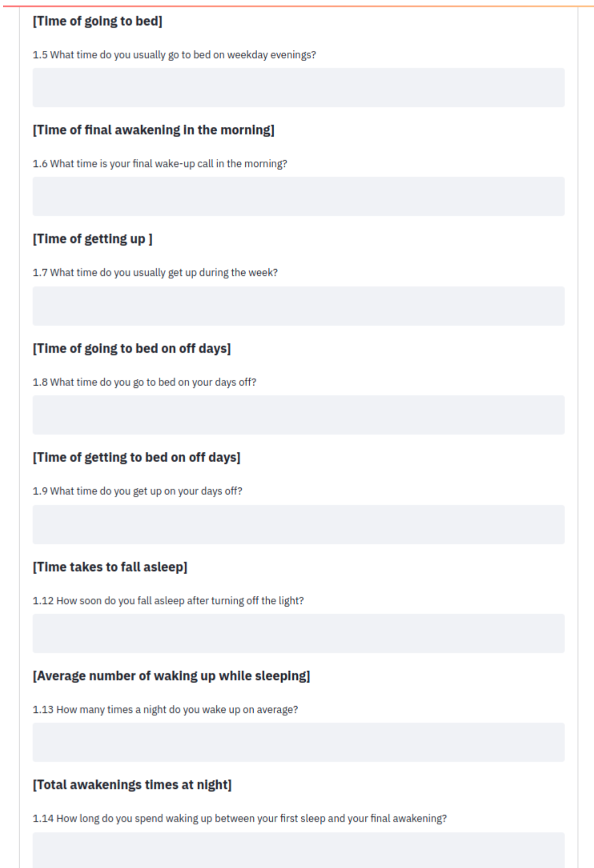
- None
- All the time
- Most of the time
- A good part of the time
- Sometimes
- Rarely
- Never

[Being drowsy during the day]

4.Have you felt your mind foggy or drowsy during the day?

- None
- All the time
- Most of the time
- A good part of the time
- Sometimes
- Rarely

Figure B.1: MOS-SS questionnaire



[Time of going to bed]

1.5 What time do you usually go to bed on weekday evenings?

[Time of final awakening in the morning]

1.6 What time is your final wake-up call in the morning?

[Time of getting up]

1.7 What time do you usually get up during the week?

[Time of going to bed on off days]

1.8 What time do you go to bed on your days off?

[Time of getting to bed on off days]

1.9 What time do you get up on your days off?

[Time takes to fall asleep]

1.12 How soon do you fall asleep after turning off the light?

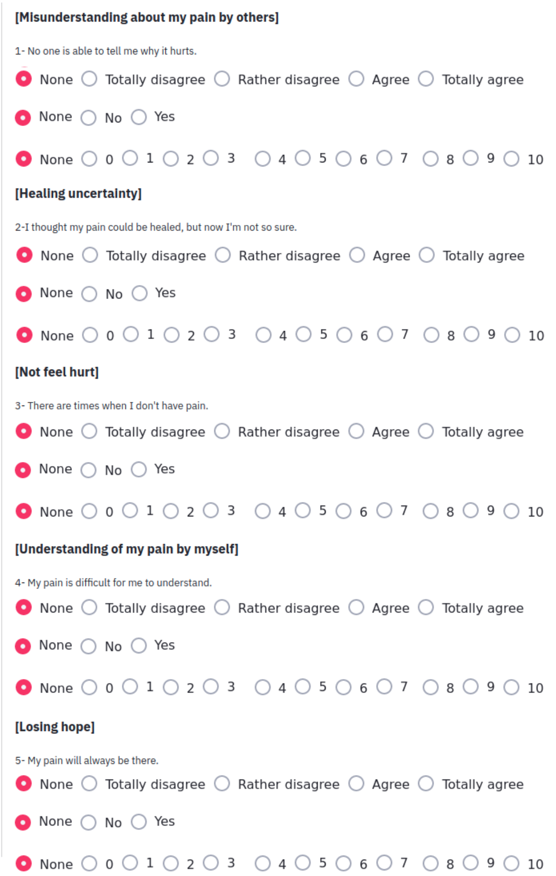
[Average number of waking up while sleeping]

1.13 How many times a night do you wake up on average?

[Total awakenings times at night]

1.14 How long do you spend waking up between your first sleep and your final awakening?

Figure B.2: Morin questionnaire



[Misunderstanding about my pain by others]

1- No one is able to tell me why it hurts.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Healing uncertainty]

2-I thought my pain could be healed, but now I'm not so sure.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Not feel hurt]

3- There are times when I don't have pain.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Understanding of my pain by myself]

4- My pain is difficult for me to understand.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

[Losing hope]

5- My pain will always be there.

None Totally disagree Rather disagree Agree Totally agree

None No Yes

None 0 1 2 3 4 5 6 7 8 9 10

Figure B.3: PBPI questionnaire

Appendix C

Data Collection for Chapter 3.2

Instructions used for data collection in Amazon Mechanical Turk and the interface are shown in figures C.1, C.2 and C.3.

We requested the Turkers to converse with the heath-bot for at least 10 turns in total.

Task Description

In this task, you are going to talk to a chatbot about health and quality of your life.

You are supposed to **play the role of a chronic pain patient**, and **share your pain with the bot**.
What is chronic pain? Doctors often define chronic pain as any pain that lasts for 3 to 6 months or more. Chronic pain can have real effects on day-to-day life and mental health.

It is very important that you **mention about all follwing key points during your conversation** (in a seamless way):

1. (Constantly/Temporarly) in pain
2. (Having/Losing) hope for getting healed
3. (Feeling/Not feeling) guiltiness that the pain is your fault
4. (Possibility/Impossibility) of healing
5. (Understanding/Not understanding) the reason of having pain

Try to **give an implicit and seamless reference to these keypoints** in the dialogue (with considering the flow of conversation). **Prevent using the same wording** in your messages.

**** BONUS ****

Playing the role of a chronic pain patient and mentioning each keypoint will get 0.5 $ bonus. By metioning all keypoints you will get 2.5 $ bonus (do not use the same wording).

note1: It is an information seeking conversation and **you are not expected to ask questions from the bot**.
note2: Wait for the bot message to be appeared completely, and then reply.
note3: When it is your time to reply, **reply only once**.

Figure C.1: Instructions (part 1)

Lead the conversation

The chatbot is not developed to ask you explicitly about these key points. Therefore, you have to mention them creatively during the dialog flow.
To make it easier for you, we have given you the authority of **controling chatbot messages**. You can direct the conversation by **changing chatbot reply**. To do that, you can **click on the "next" button** (below the bot message) to change the chatbot utterance and if you found it good enough you can just continue the conversation.

Annotating each user reply

After you entered your answer, you will notice **5 checkpoints apear below your answer**. Each check point refers to each of the keypoints. **If one or mutliple keypoints have been mentioned in your answer (implicitly or explicitly) choose the related checkpoints**.
Please keep in my mind that you have to **fill check points before entering your next response**. They would be disabeled afterwards.

End the conversation

To end the conversation, you can click on **green "Submit" button**. But before that, **wiat for the bot message to be appeared completely, and then press the submit button**.
If you click on the button before reaching to the minimum number of turns (5 messages each user), you will receive an alert error message and be taken back to the conversation to complete the task.

Figure C.2: Instructions (part 2)

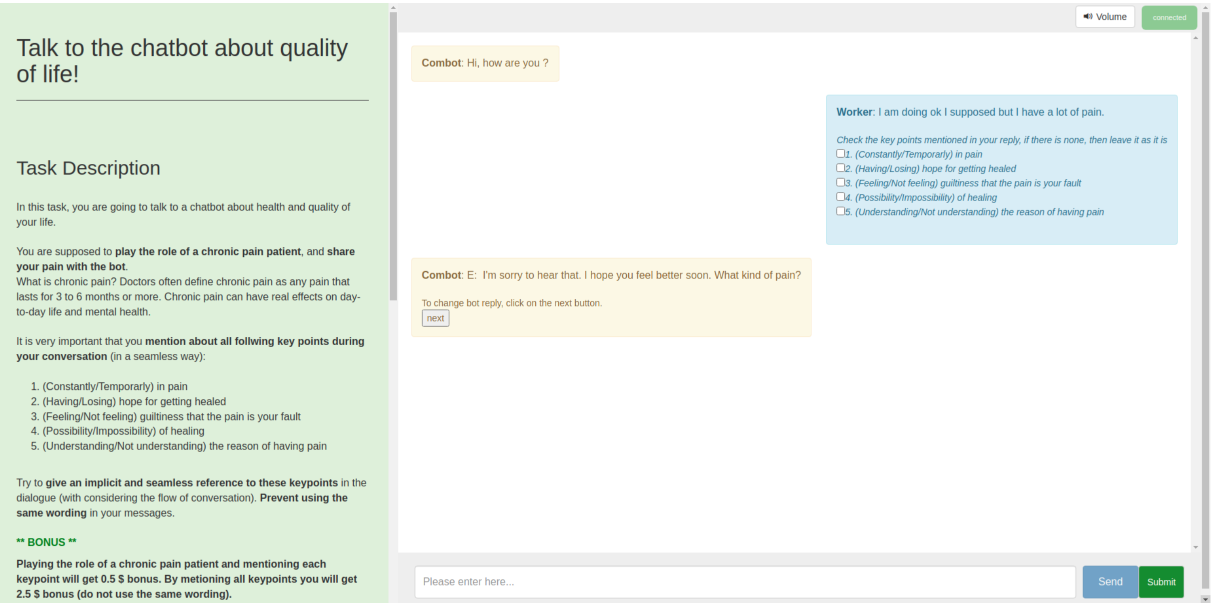


Figure C.3: Interface

Bibliography

Afia Fairoose Abedin, Amirul Islam Al Mamun, Rownak Jahan Nowrin, Amitabha Chakrabarty, Moin Mostakim, and Sudip Kumar Naskar. 2021. A deep learning approach to integrate human-level understanding in a chatbot. *ArXiv*, abs/2201.02735.

Bacic JD Walsh KE Shanahan CW Paasche-Orlow MK Adams WG, Phillips BD. 2014. Automated conversation system before pediatric primary care visits: a randomized trial. *Pediatrics*, 134.

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020a. Towards a human-like open-domain chatbot.

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

Monica Agrawal, Janette L. Cheng, and Caelin Tran. 2017. What ' s up , doc ? a medical diagnosis bot.

Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *ArXiv*, abs/1606.00372.

Vinsen Marselino Andreas, Genta Indra Winata, and Ayu Purwarianti. 2021. A comparative study on language models for task-oriented dialogue systems. *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5.

Modafar Ati, Wail Omar, and Asmaa S Hussein. 2015. Integration framework of chronic disease management system and a recommender system in the united arab emirates. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 570–574. IEEE.

R.michael Bagby, James D.a. Parker, and Graeme J. Taylor. 1994. The twenty-item toronto alexithymia scale—i. item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, 38(1):23–32.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *ICWSM*.

Ian Berlot-Attwell and Frank Rudzicz. 2022. Relevance in dialogue: Is less more? an empirical comparison of existing metrics, and a novel simple metric. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 166–183, Dublin, Ireland. Association for Computational Linguistics.

Hélène Bihan, Silvana Laurent, Catherine Sass, Gérard Nguyen, Caroline Huot, Jean Jacques Moulin, René Guegen, Philippe Le Toumelin, Hervé Le Clésiau, Emilio La Rosa, Gérard Reach, and Régis Cohen. 2005. Association among individual deprivation, glycemic control, and diabetes complications. *Diabetes Care*, 28(11):2680–2685.

Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *ArXiv*, abs/1605.07683.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

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

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

S. Honoré-Masson G. Berne P.H. Ruel B. Laurent C. Thomas-Antérion, C. Ribas. 2004. Le questionnaire de plainte cognitive (qpc) : outil de dépistage de la plainte des sujets présentant une maladie d'alzheimer ou un mci. *Revue Neurologique*, 1027(4502):5–75.

Leonardo Campillos Llanos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum, and Sophie Rosset. 2015. Description of the PatientGenesys dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–440, Prague, Czech Republic. Association for Computational Linguistics.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin Leng Theng, and Rifat Atun. 2020a. Conversational agents in health care: Scoping review and conceptual analysis.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq R. Joty, Yin Leng Theng, and Rifat A. Atun. 2020b. Conversational agents in health care: Scoping review and conceptual analysis. *Journal of Medical Internet Research*, 22.

Jaime Carbonell and Jade Goldstein. 1998. Use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, pages 335–336, New York, New York, USA. ACM Press.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alessandra Cervone, Evgeny A. Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *INTERSPEECH*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017a. A survey on dialogue systems: Recent advances and new frontiers. *ArXiv*, abs/1711.01731.

Qian Chen and Wen Wang. 2019. Sequential attention-based network for noetic end-to-end response selection. *ArXiv*, abs/1901.02609.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stanislav Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? *ArXiv*, abs/2110.06918.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555.

Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artif. Intell.*, 2:1–25.

Pierre Colombo, Guillaume Staerman, Chloé Clavel, and Pablo Piantanida. 2021. Automatic text evaluation through the lens of Wasserstein barycenters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10466, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miller R.P. Todd D.D. Kori, S.H. 1990. Kinesiophobia: A new view of chronic pain behavior. *Pain Management*, (3):35–43.

Andrea DeCapua and Joan Findlay Dunham. 1993. Strategies in the discourse of advice. *Journal of Pragmatics*, 20(6):519–531.

Kees van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Squibs and discussions: Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755 – 810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephanie J Dimitroff, Omid Kardan, Elizabeth A. Necka, Jean Decety, Marc G. Berman, and Greg J Norman. 2017. Physiological dynamics of stress contagion. *Scientific Reports*, 7.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ICLR*.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 915–925, Florence, Italy. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. 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)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

L. F. D'Haro, Koichiro Yoshino, Chiori Hori, Tim K. Marks, Lazaros Polymenakos, Jonathan K. Kummerfeld, Michel Galley, and Xiang Gao. 2020. Overview of the seventh dialog system technology challenge: Dstc7. *Comput. Speech Lang.*, 62:101068.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017a. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017b. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. *Proceedings of the 24th International Conference on World Wide Web*.

Maxine Eskénazi, Shikib Mehri, Evgeniia Razumovskaia, and Tiancheng Zhao. 2019. Beyond turing: Intelligent agents centered on the user. *ArXiv*, abs/1901.06613.

Robert G Farrell, Catalina M Danis, Sreeram Ramakrishnan, and Wendy A Kellogg. 2012. Intrapersonal retrospective recommendation: lifestyle change recommendations using stable patterns of personal behavior. In *Proceedings of the First International Workshop on Recommendation Technologies for Lifestyle Change (LIFESTYLE 2012), Dublin, Ireland*, page 24. Citeseer.

Denis G. Fedorenko, Nikita Smetanin, and Artem Rodichev. 2017. Avoiding echo-responses in a retrieval-based conversation system. *ArXiv*, abs/1712.05626.

Greg P. Finley, Erik Edwards, Amanda Robinson, Najmeh Sadoughi, James Fone, Mark Miller, David Suendermann-Oeft, Michael Brenndoerfer, and Nico Axtmann. 2018. An automated assistant for medical scribes. In *INTERSPEECH*.

Michael H. Fischer and Monica S. Lam. 2016. From books to bots: Using medical literature to create a chat bot. *Proceedings of the First Workshop on IoT-enabled Healthcare and Wellness Technologies and Systems*.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Michel Galley, Eric Fosler-Lussier, and Alexandros Potamianos. 2001. Hybrid natural language generation for spoken dialogue systems. In *INTERSPEECH*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. 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)*, pages 1229–1238, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah J. Jones, Àgata Lapedriza, and Rosalind W. Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *NeurIPS*.

Farnaz Ghassemi Toudeshki, Philippe Jolivet, Alexandre Durand-Salmon, and Anna Liednikova. 2021. Zero-shot clinical questionnaire filling from human-machine interactions. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Farnaz Ghassemi Toudeshki, Anna Liednikova, Philippe Jolivet, and Claire Gardent. 2022. Exploring the influence of dialog input format for unsupervised clinical questionnaire filling. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sarik Ghazarian, Behnam Hedayatnia, Alexandros Papangelis, Yang Liu, and Dilek Hakkani-Tur. 2022. What is wrong with you?: Leveraging user sentiment for automatic dialog evaluation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4194–4204, Dublin, Ireland. Association for Computational Linguistics.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarik Ghazarian, Ralph M. Weischedel, A. G. Galstyan, and Nanyun Peng. 2020. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *AAAI*.

Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: Facilitating user symptom check using a personalised chatbot-oriented dialogue system. *Studies in health technology and informatics*, 252:51–56.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. 1996. A form-based dialogue manager for spoken language applications. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 701–704 vol.2.

Jade Goldstein and Jaime Carbonell. 1998. Summarization: (1) using MMR for diversity-based reranking and (2) evaluating summaries. In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 181–195, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, Florence, Italy. Association for Computational Linguistics.

Enying Gong, Shaira Baptista, Anthony W. Russell, Paul A. Scuffham, Michaela A. Riddell, Jane Speight, Dominique Bird, Emily D Williams, Mojtaba Lotfaliany, and Brian Oldenburg. 2020. My diabetes coach, a mobile app–based interactive conversational agent to support type 2 diabetes self-management: Randomized effectiveness-implementation trial. *Journal of Medical Internet Research*, 22.

Allen L. Gorin, Giuseppe Riccardi, and Jeremy H. Wright. 1997. How may i help you? *Speech Commun.*, 23:113–127.

Paul Green and Lisa Wei-Haas. 1985. The rapid development of user interfaces: Experience with the wizard of oz method. *Proceedings of the Human Factors Society Annual Meeting*, 29(5):470–474.

Jia-Chen Gu, Tianda Li, Quan Liu, Xiaodan Zhu, Zhenhua Ling, Zhiming Su, and Si Wei. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *AAAI*.

Nancie Gunson, Weronika Sieińska, Christopher Walsh, Christian Dondrup, and Oliver Lemon. 2020. It's good to chat?: Evaluation and design guidelines for combining open-domain social conversation with task-based dialogue in intelligent buildings. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*.

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

Stefan Hahn, Marco Dinarelli, C. Raymond, Fabrice Lefèvre, Patrick Lehnen, Renato De Mori, Alessandro Moschitti, Hermann Ney, and Giuseppe Riccardi. 2011. Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:1569–1583.

Dilek Hakkani-Tur. 2018. Announcing the 2018 alexa prize winner: University of california, davis.

Nidhin Harilal, Rushil Shah, Saumitra Sharma, and Vedanta Bhutani. 2020. CARO: An empathetic health conversational chatbot for people with major depression. In *ACM International Conference Proceeding Series*, pages 349–350, New York, NY, USA. ACM.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019b. Convert: Efficient and accurate conversational representations from transformers.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*.

Richard David Jacques, Asbjørn Følstad, Elizabeth Gerber, Jonathan T. Grudin, Ewa Luger, A. Monroy-Hernández, and Dakuo Wang. 2019. Conversational agents: Acting on the wave of research and development. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*.

Raheleh Jafari, Sina Razvarz, and Alexander E. Gegov. 2020. End-to-end memory networks: A survey. In *SAI*.

Sina Jafarpour. 2010. Filter, rank, and transfer the knowledge: Learning to chat.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*.

H Jang, E Rempel, D Roth, G Carenini, and NZ Janjua. 2021. Tracking covid-19 discourse on twitter in north america: Topic modeling and aspect-based sentiment analysis. *Journal of Medical Internet Research*.

Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *ArXiv*, abs/1408.6988.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

Dan Jurafsky and Elizabeth Shriberg. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual.

Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayan, and A. Sheth. 2019. kbot: Knowledge-enabled personalized chatbot for asthma self-management. *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143.

Rudolf Kadlec, Martin Schmid, and Jan Kleindienst. 2015. Improved deep learning baselines for ubuntu corpus dialogs. *ArXiv*, abs/1510.03753.

R KARASEK. 1985. Job content questionnaire user's guide. *Department of Work Environemnt*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal, Dilek Z. Hakkani-Tür, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *ArXiv*, abs/1812.10757.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Ilia Kulikov, Alexander H. Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling.

Jeffrey R. M. Kunz. 1982. *The American Medical Association Family Medical Guide*. New York: Random House.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, and Vincent S Tseng. 2015. An intelligent recommender system based on short-term risk prediction for heart disease patients. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 102–105. IEEE.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

G FRANK Lawlis, RAMON Cuencas, DAVID Selby, and CE McCoy. 1989. The development of the dallas pain questionnaire. an assessment of the impact of spinal pain on behavior. *Spine*, 14(5):511–516.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324.

Alan Lee, Rashmi Prasad, Aravind K. Joshi, Nikhil Dinesh, and Bonnie Lynn Webber. 2006. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax?

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yi Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, i Feel You": Encouraging Deep Self-disclosure through a Chatbot. *Conference on Human Factors in Computing Systems - Proceedings*, (April).

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2020. Knowledgeable dialogue reading comprehension on key turns. *arXiv preprint arXiv:2004.13988*.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017a. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning health-bots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 638–648, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2021. Gathering information and engaging the user ComBot: A task-based, serendipitous dialog model for patient-doctor interactions. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 21–29, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019a. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2019b. CAiRE: An End-to-End Empathetic Chatbot.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R. Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, Wai Leng Chow, and Nancy F. Chen. 2019b. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Gale M. Lucas, Albert A. Rizzo, J. Gratch, Stefan Scherer, Giota Stratou, Jill Boberg, and Louis-Philippe Morency. 2017. Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers. *Frontiers Robotics AI*, 4:51.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

Sourabh Majumdar, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Generating challenge datasets for task-oriented conversational agents through self-play. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 693–702, Varna, Bulgaria. INCOMA Ltd.

Sourab Mangrulkar, Suhani Shrivastava, Veena Thenkanidiyoor, and Dileep Aroor Dinesh. 2018. A context-aware convolutional natural language generation model for dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 191–200, Melbourne, Australia. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Christina Maslach, Susan E Jackson, and Michael P Leiter. 1997. *Maslach burnout inventory*. Scarecrow Education.

Robert Meadows, Christine Hine, and Eleanor Suddaby. 2020. Conversational agents and the making of mental health recovery. *DIGITAL HEALTH*, 6:2055207620966170. PMID: 33282335.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3836–3845, Florence, Italy. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2017. Coherent dialogue with attention-based language models. In *AAAI*.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450, Online. Association for Computational Linguistics.

Mohsen Mesgar, Sebastian B ¨ Ucker, and Iryna Gurevych. 2019. A Neural Model for Dialogue Coherence Assessment. Technical report.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Z. Hakkani-Tür, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:530–539.

Katherine Middleton, Mobasher Butt, Nils Y. Hammerla, Steven Hamblin, Karan Mehta, and Ali Davod Parsa. 2016. Sorting out symptoms: design and evaluation of the 'babylon check' automated triage system. *ArXiv*, abs/1606.02041.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Li, Pavan Kapanipathi, and Kartik Talamadupula. 2020. Reading Comprehension as Natural Language Inference: A Semantic Analysis.

Teruhisa Misu, Kallirroi Georgila, Anton Leuski, and David Traum. 2012. Reinforcement learning of question-answering dialogue policies for virtual museum guides. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 84–93, Seoul, South Korea. Association for Computational Linguistics.

Rashmi Mohan and Lihong Li. 2009. Reinforcement learning for dialog management using least-squares policy iteration and fast feature selection. In *Proceedings of INTERSPEECH*.

Joonas Moilanen, Aku Visuri, Elina Kuosmanen, Andy Alorwu, and Simo Johannes Hosio. 2022. Designing personalities for mental health conversational agents. In *IUI Workshops*, pages 139–148.

Charles M Morin. 1993. *Insomnia: Psychological assessment and management.* Guilford press.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 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)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Tomás Nestorovic. 2009. Towards flexible dialogue management using frames. In *TSD*.

Tomás Nestorovic. 2011. Task-oriented dialogue agent architecture. *J. Digit. Inf. Manag.*, 9:1–8.

Anh Nguyen and Wayne Wobcke. 2005. An agent-based approach to dialogue management in personal assistants. In *IUI '05*.

Lin Ni, Chenhao Lu, Niu Liu, and J. Liu. 2017. Mandy: Towards a smart primary care chatbot application.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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)*, pages 1699–1713, Online. Association for Computational Linguistics.

Rodrigo Nogueira. 2019. From doc2query to docTTTTTquery.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Olabiyi Oluwatobi and Erik Mueller. 2020. DLGNet: A transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 54–62, Online. Association for Computational Linguistics.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of Medical Internet Research*, 21.

Ioannis Papaioannou, Amanda Cercas Curry, Jose Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondrej Dusek, Verena Rieser, and Oliver Lemon. 2017a. Alana: Social Dialogue using an Ensemble Model and a Ranker trained on User Feedback. In *2017 Alexa Prize Proceedings*.

Ioannis Papaioannou, Amanda Cercas Curry, Jose L Part, Igor Shalyminov, Xinnuo Xu, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. 2017b. An ensemble model with ranking for social dialogue. *arXiv preprint arXiv:1712.07558*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

A. Parrott and I. Hindmarch. 2004. The leeds sleep evaluation questionnaire in psychopharmacological investigations—a review. *Psychopharmacology*, 71:173–179.

Puntip Pattaraintakorn, Gregory M Zaverucha, and Nick Cercone. 2007. Web based health recommender system using rough sets, survival analysis and rule-based expert systems. In *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pages 491–499. Springer.

Guillemin F. Virion J. M. Briançon S. Paul-Dauphin, A. 1999. Bias and precision in visual analogue scales: a randomized controlled trial. *American journal of epidemiology*, 150(10):1117–1127.

Kate Pearce, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. 2021. A comparative study of transformer-based language models on extractive question answering. *arXiv preprint arXiv:2110.03142*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Juanan Pereira and Oscar Díaz. 2019. Using health chatbots for behavior change: A mapping study. *Journal of Medical Systems*, 43:1–13.

John Petrocelli. 2002. Processes and stages of change: Counseling with the transtheoretical model of change. *Journal of Counseling & Development*, 80.

James Prochaska and Wayne Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion : AJHP*, 12:38–48.

Simon Provoost, Ho Ming Lau, Jeroen Ruwaard, and Heleen Riper. 2017. Embodied conversational agents in clinical psychology: A scoping review. *Journal of Medical Internet Research*, 19.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2021. Valuenet: A new dataset for human value driven dialogue system.

Cleeland C. S. Flanery R. C. Daut, R. L. 1983. Development of the wisconsin brief pain questionnaire to assess pain in cancer and other diseases. *Pain*, (17):197–210.

Adams H. Horan S. Mahar D. Boland D. Gross R. Sullivan, M.J.L. 2008. The role of perceived injustice in the experience of chronic pain and disability: Scale development and validation. *Journal of Occupational Rehabilitation*, (18):249–61.

Boele de Raad. 2000. The big five personality factors : the psycholexical approach to personality.

A. Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize. *ArXiv*, abs/1801.03604.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ehud Reiter and R. Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3:57–87.

Jiangtao Ren, Naiyin Liu, and Xiaojing Wu. 2020. Clinical questionnaire filling based on question answering framework. *International Journal of Medical Informatics*, 141:104225.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

A Rivero-Rodriguez, S Th Konstantinidis, Carlos Luis Sánchez-Bocanegra, and Luis Fernández-Luque. 2013. A health information recommender system: Enriching youtube health videos with medline plus information by the use of snomedct terms. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 257–261. IEEE.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Jost Schatzmann, Blaise Thomson, Karl Weilhammer, Hui Ye, and Steve Young. 2007. Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152, Rochester, New York. Association for Computational Linguistics.

Jost Schatzmann, Karl Weilhammer, Matthew N. Stuttle, and Steve J. Young. 2006. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *The Knowledge Engineering Review*, 21:97 – 126.

Michael F. Scheier, Charles S. Carver, and Michael W. Bridges. 1994. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, 67(6):1063–1078.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks : the official journal of the International Neural Network Society*, 61:85–117.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See and Christopher Manning. 2021. Understanding and predicting user dissatisfaction in a neural generative chatbot. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Iulian Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, A. P. Sarath Chandar, Nan Rosemary Ke, Sai Mudumba, Alexandre de Brébisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. 2017. A deep reinforcement learning chatbot. *ArXiv*, abs/1709.02349.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *ArXiv*, abs/1507.04808.

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

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018a. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Pararth Shah, Dilek Z. Hakkani-Tür, Gökhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak Kennard, and Larry Heck. 2018b. Building a conversational agent overnight with dialogue self-play. *ArXiv*, abs/1801.04871.

Yong Shan, Jinchao Zhang, Zekang Li, Yang Feng, and Jie Zhou. 2022. Mental health assessment for the chatbots.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Bayan Shawar and Eric Atwell. 2007. Chatbots: Are they really useful? *LDV Forum*, 22:29–49.

Cathy Donald Sherbourne and Anita L. Stewart. 1991. The mos social support survey. *Social Science Medicine*, 32(6):705–714.

Anshumali Shrivastava and Ping Li. 2014. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *ArXiv*, abs/1405.5869.

Sarthak Sinha, Amanda J. Schreiner, Jeff Biernaskie, Duncan Nickerson, and Vincent A. Gabriel. 2017. Treating pain on skin graft donor sites: Review and clinical recommendations. *Journal of Trauma and Acute Care Surgery*, 83(5):954–964.

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

David R. So, Chen Liang, and Quoc V. Le. 2019. The evolved transformer. *ArXiv*, abs/1901.11117.

Insu Song and Nigel V Marsh. 2012. Anonymous indexing of health conditions for a similarity measure. *IEEE Transactions on Information Technology in Biomedicine*, 16(4):737–744.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are Better than One: An Ensemble of Retrieval- and Generation-Based Dialog Systems.

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

Sebastian Spänig, Agnes Emberger-Klein, Jan-Peter Sowa, Ali Canbay, Klaus Menrad, and Dominik Heider. 2019. The virtual doctor: An interactive artificial intelligence based on deep learning for non-invasive prediction of diabetes. *Artificial intelligence in medicine*, 100.

Published by Statista Research Department and Mar 17. 2022. Global intelligent assistant market share 2017-2020.

Amanda Stent and Srinivas Bangalore. 2014. Natural language generation in interactive systems: Introduction.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Pei-Hao Su, David Vandyke, Milica Gasic, Dongho Kim, Nikola Mrksic, Tsung-Hsien Wen, and Steve J. Young. 2015. Learning from real users: rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *INTERSPEECH*.

Sainbayar Sukhbaatar, Arthur D. Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. *ArXiv*, abs/2010.12757.

Mary Sun. 2022. *Natural Language Processing for Health System Messages: Deep Transfer Learning Approach to Aspect-Based Sentiment Analysis of COVID-19 Content*. Ph.D. thesis, Harvard University.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Stephen Sutton, Ronald A. Cole, Jacques de Villiers, Johan Schalkwyk, Pieter J. E. Vermeulen, Michael W. Macon, Yonghong Yan, Edward C. Kaiser, Brian Rundle, Khaldoun Shobaki, John-Paul Hosom, Alexander Kain, Johan Wouters, Dominic W. Massaro, and Michael M. Cohen. 1998. Universal speech tools: the cslu toolkit. In *ICSLP*.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*.

Nguyen Tho Thong et al. 2015. Hifcf: An effective hybrid model between picture fuzzy clustering and intuitionistic fuzzy recommender systems for medical diagnosis. *Expert Systems with Applications*, 42(7):3682–3701.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

Jörg Tiedemann. 2009. News from opus — a collection of multilingual parallel corpora with tools and interfaces.

David R. Traum and Peter A. Heeman. 1996. Utterance units in spoken dialogue. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems*.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. 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)*, pages 2948–2958, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Murray Turoff and Starr Roxanne Hiltz. 2008. The future of professional communities of practice. In *Workshop on E-Business*, pages 144–158. Springer.

Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *ArXiv*, abs/1506.05869.

Gordon Waddell, Mary Newton, Iain Henderson, Douglas Somerville, and Chris J Main. 1993. A fear-avoidance beliefs questionnaire (fabq) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain*, 52(2):157–168.

Mattias Wahde and M. Virgolin. 2022. Conversational agents: Theory and applications. *ArXiv*, abs/2202.03164.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2015. Syntax-based deep matching of short texts. *ArXiv*, abs/1503.02427.

John E Ware Jr, Mark Kosinski, and Susan D Keller. 1996. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Medical care*, pages 220–233.

Frank W Weathers, Brett T Litz, Debra S Herman, Jennifer A Huska, Terence M Keane, et al. 1993. The ptsd checklist (pcl): Reliability, validity, and diagnostic utility. In *annual convention of the international society for traumatic stress studies, San Antonio, TX*, volume 462. San Antonio, TX;.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9:36–45.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory networks. *CoRR*, abs/1410.3916.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and Heuiseok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *INTERSPEECH*.

Martin Wiesner and Daniel Pfeifer. 2014. Health recommender systems: concepts, requirements, technical basics and challenges. *International journal of environmental research and public health*, 11(3):2580–2607.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

David A. Williams and Beverly E. Thorn. 1989. An empirical assessment of pain beliefs. *Pain*, 36(3):351–358.

J. Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue Discourse*, 7:4–33.

WoeBot. 2022. [link].

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents.

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

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *ArXiv*, abs/2007.00808.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of AAAI*, volume 33.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, William Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Steve J. Young. 2002. Talking to machines (statistically speaking). In *INTERSPEECH*.

Steve J. Young, Milica Gasic, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Comput. Speech Lang.*, 24:150–174.

Zhou Yu, Alan W. Black, and Alexander I. Rudnicky. 2017. Learning conversational systems that interleave task and non-task content. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4214–4220. AAAI Press.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 404–412, Los Angeles. Association for Computational Linguistics.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 111–120.

Wlodek Zadrozny, Malgorzata Budzikowska, Joyce Yue Chai, Nanda Kambhatla, Sylvie Levesque, and Nicolas Nicolov. 2000. Natural language dialogue for personalized interaction. *Commun. ACM*, 43:116–120.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. DynaEval: Unifying turn and dialogue level evaluation. 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)*, pages 5676–5689, Online. Association for Computational Linguistics.

Hainan Zhang, Yanyan Lan, J. Guo, Jun Xu, and Xueqi Cheng. 2018a. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018c. Learning to summarize radiology findings. In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*.

Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2020c. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018d. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*.

Zhuosheng Zhang, Junlong Li, and Hai Zhao. 2021b. Multi-turn dialogue reading comprehension with pivot turns and knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1161–1173.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–10, Los Angeles. Association for Computational Linguistics.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017a. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 27–36, Saarbrücken, Germany. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017b. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Xinyan Zhao, Bin He, Yasheng Wang, Yitong Li, Fei Mi, Yajiao Liu, Xin Jiang, Qun Liu, and Huanhuan Chen. 2022. UniDS: A unified dialogue system for chit-chat and task-oriented dialogues. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 13–22, Dublin, Ireland. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

Anthony S Zigmond and R Philip Snaith. 1983. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, 67(6):361–370.