Perspectivas Interdisciplinares entre Modelos de Fundação para Inteligência Artificial e o Interacionismo Simbólico no Design de Interação [1]

Everaldo PEREIRA[2]
Ana Paula Scabello MELLO[3]
Flávia Janine Rosante BEO[4]
Rafael Angel BORDENABE[5]
Eduardo Takase SAWADA[6]
Daniel Djinishian de BRIQUEZ[7]
Instituto Mauá de Tecnologia, São Caetano do Sul, SP

RESUMO

Com a rápida evolução da tecnologia de reconhecimento de voz e o aumento da popularidade de dispositivos controlados por voz, como assistentes virtuais e sistemas de navegação embutidos em dispositivos IoT, o design de interação tornou-se uma área de estudo importante para garantir interações mais eficientes e naturais entre humanos e dispositivos. Entender os princípios fundamentais do design de interação, os desafios enfrentados por comunicadores, designers e programadores, permite melhores práticas para criar interfaces de voz intuitivas. Uma vez que um dos propósitos do design de interação é o desenvolvimento de interfaces que sejam fáceis de usar e compreendam as intenções e necessidades dos usuários, o Interacionismo Simbólico (IS) tem uma forte contribuição nesse contexto, tendo em vista que esta corrente teórica se concentra na compreensão da sociedade por meio da análise das interações simbólicas entre indivíduos. Blumer (1969), um dos principais expoentes do IS, nota que as pessoas atribuem significado aos objetos, eventos e situações por meio de processos interpretativos e comunicativos. A teoria sugere que a realidade social é construída por meio de uma interação simbólica, na qual as pessoas interpretam e atribuem significado às ações, palavras umas das outras. Compreende-se que o mundo simbólico se constrói por meio da interação entre duas ou mais pessoas, sendo assim, o simbolismo não é resultado de interação de uma pessoa consigo e, mesmo com objetos (BLUMER, 1969). Sociólogos que trabalham nessa linha pesquisaram uma extensiva gama de tópicos utilizando uma variedade de métodos de investigação. Entretanto, a maioria das pesquisas de interacionistas utiliza métodos de pesquisa qualitativa, como observação participante, para estudar aspectos da interação social e *self* (individualidade). O Interacionismo Simbólico originou-se no trabalho de George Herbert Mead, que demonstrou que os egos das pessoas são produtos sociais, sem deixar de terem propósitos e de serem criativos. Outros precursores na área foram Herbert Blumer e Charles Cooley. Blumer, um

[1] Trabalho apresentado no GP Tecnologias e Culturas Digitais, evento do 46º Congresso Brasileiro de Ciências da Comunicação realizado de 4 a 8 de setembro de 2023.

[2] Doutor em Comunicação Social pela Universidade Metodista de São Paulo (Umesp). Docente no curso de Design do Instituto Mauá de Tecnologia. E-mail: everaldo@maua.br

[3] Doutora em Design pela Faculdade de Arquitetura e Urbanismo da Universidade de São Paulo. Docente no curso de Design do Instituto Mauá de Tecnologia. E-mail: anamello@maua.br

[4] Mestre em Ciência da Computação. Docente no curso de Ciências da Computação do Instituto Mauá de Tecnologia. E-mail: flavia.beo@maua.br

[5] Graduando no curso de Design do Instituto Mauá de Tecnologia. Aluno de Iniciação Científica. E-mail:

[6] Graduando no curso de Ciências da Computação do Instituto Mauá de Tecnologia. Aluno de Iniciação Científica. E-mail: sawadaeduardo@gmail.com

[7] Graduando no curso de Ciências da Computação do Instituto Mauá de Tecnologia. Aluno de Iniciação Científica. E-mail: 22.00251-0@maua.br

estudioso de Mead, e criador do termo "Interacionismo Simbólico", pôs em evidência as principais perspectivas dessa abordagem: as pessoas agem em relação às coisas baseando-se no significado que essas coisas tenham para elas; e esses significados são resultantes da sua interação social e modificados por sua interpretação. Blumer seguiu Mead e apresentou a proposição de que as pessoas interagem umas com as outras por meio de interpretação mútua das ações e definição um do outro, em vez de somente reagir às ações um do outro. Suas respostas não são dadas diretamente às ações um do outro, mas baseadas no significado que eles atribuem a tais ações. Assim, interação humana é mediada pelo uso de símbolos e significados, através de interpretação, ou determinação do significado das ações um do outro. Blumer comparou esse processo, que ele designou "interação simbólica", com as explicações behavioristas do comportamento humano que não consideram a interpretação entre estímulo e resposta, como em Schram (1949), entre outros. Chamamos a atenção que atualmente muitos objetos usam a interação por inteligência artificial e que, portanto, *a priori* poderiam conduzir conversações com significados específicos. Naturalmente essas conversações são programadas por outros indivíduos que, mesmo dentro de suas corporações, podem gerar campos de significações. Assim, observamos no campo computacional, que os Modelos de Fundação (BOMMASANI *et al*., 2021) podem ser usados para treinar inteligências artificiais e melhorar a capacidade de entender e responder a comandos de voz em design de interação e têm demonstrado habilidades avançadas de Processamento de Linguagem Natural (PLN), permitindo que eles entendam e gerem conteúdo de maneira mais sofisticada. Desse ponto de vista, vale lembrar que a linguagem é um instinto inato e universalmente presente na espécie humana. Steven Pinker (2002) argumenta que a linguagem é uma adaptação biológica que surge naturalmente em nossa espécie e é moldada por processos evolutivos, independentemente das bases culturais, mas por essas modificada. Pinker entende que há uma gramática universal, que sugere que as pessoas têm uma predisposição inata para adquirir a linguagem, exemplificando como as crianças aprendem de forma rápida e eficiente, mesmo sem instrução formal. O conceito de gramática universal e a aquisição da linguagem pode ajudar a orientar o desenvolvimento de algoritmos e modelos na PLN. Neste recorte, procuramos avaliar as perspectivas de Modelos de Fundação associados ao Interacionismo Simbólico, analisando a capacidade de entender melhor comandos de voz complexos, responder a perguntas com mais precisão ou gerar respostas mais naturais e contextualmente relevantes. Além disso, é possível que esta associação possa ser usada para melhorar a capacidade das interfaces de voz em realizar tarefas específicas, como fazer doações, fazer reservas ou fornecer informações detalhadas sobre determinado assunto. Uma vez que os Modelos de Fundação são treinados em grandes quantidades de texto que contêm informações simbólicas e significados culturais e aprendem a compreender e gerar linguagem, incluindo símbolos e significados associados a palavras e frases é relevante observá-los por uma perspectiva teórica que se concentra na forma como os seres humanos desenvolvem o sentido e o significado por meio da interação social. O design de interação por voz requer uma compreensão profunda das capacidades e limitações da voz, por isso é essencial considerar elementos como a compreensão da linguagem natural, a adaptação a diferentes contextos e a personalização da experiência do usuário, como observado por Deibel e Evanhoé (2021). No entanto, o ambiente IoT atual está se configurando em um emaranhado de dispositivos distintos e de difícil compartilhamento de ativos gráficos e sonoros, o que torna o desenvolvimento de soluções de design em IoT um grande desafio. São necessários projetos específicos para cada plataforma e com graus de complexidade distintos e que demandam muitas adaptações. A interação necessita ser natural e intuitiva,

isso envolve o desenvolvimento de sistemas de reconhecimento de voz precisos, *feedback* adequado e concepção de interações de voz e imagem intuitivas o que é possível a partir da PLN que está relacionada ao desenvolvimento de algoritmos e sistemas capazes de compreender e processar a linguagem humana em seus diferentes aspectos, incluindo a compreensão de significado, gramática, contexto e intenção. (JURAFSKY e MARTIN, 2000). Portanto, a questão de pesquisa neste recorte diz respeito a como o Interacionismo Simbólico pode ser usado como suporte teórico para ajudar IAs generativas a treinarem modelos que são aplicados em sistemas de interação. Estudar a interação entre pessoas e IoT como mídia do ponto de vista da comunicação e do design requer uma metodologia interpretativa e proposicional que resulte em um artefato, ou seja, em um modelo de prescrições, recomendações e diretrizes, que possam aprofundar o estado da arte nesse campo . Os resultados podem ajudar organizações, governos e pessoas a interpretar a internet das coisas como mídia de forma abrangente.

PALAVRAS-CHAVE: design de interação, Interacionismo Simbólico, modelos de fundação, design de voz, internet das coisas

INTRODUÇÃO

Com a rápida evolução da tecnologia de reconhecimento de voz e o aumento da popularidade de dispositivos controlados por voz, como assistentes virtuais e sistemas de navegação embutidos em dispositivos de internet das coisas (IoT), a interface de voz tornou-se um componente de estudo importante na Comunicação para garantir interações mais eficientes e naturais entre humanos e computadores. Entender os princípios fundamentais de interface de voz, o design os desafios enfrentados pelos comunicadores, designers e programadores, permite melhores práticas para criar relacionamentos mais intuitivos. Nesse sentido, compreendemos que o Interacionismo Simbólico, a partir dos conceitos de Mead e Blummer pode criar chaves de entendimento melhores para a interação entre pessoas e dispositivos IoT.

A questão de pesquisa neste recorte diz respeito a como pode ser a experiência de interação e a relação entre comunicação, design e computação em IoT, e como a inteligência artificial (IA) generativa pode ajudar a treinar modelos que são aplicados a esses sistemas de interação por voz e avalia-se a introdução do Interacionismo Simbólico como suporte teórico para treinar IAs em significação e contexto para interfaces de voz e imagem do usuário (IVIU).

Estudar a interação entre pessoas e IoT, compreender IoT como mídia, requer uma metodologia interpretativa e proposicional que resulte em um artefato, ou seja, em um modelo de prescrições, recomendações e diretrizes, que possam aprofundar o estado da

arte nesse campo. Os resultados podem ajudar organizações, governos e pessoas a interpretar a internet das coisas como mídia de forma abrangente. Entendemos que a abordagem da proposta é inovadora e interdisciplinar, pois une, por meio da Design Science, a compreensão da Comunicação, a prática do Design e a ciência da Computação, construindo pesquisas não paradigmáticas.

Em pesquisas anteriores (PEREIRA e MELLO, 2023) sobre anúncios interativos existentes, bem como testes de interação usando o software de prototipagem gráfica digital Adobe XD, foram realizadas entrevistas em profundidade com usuários de dispositivos Echo, com o sistema de atendimento por voz Alexa. É possível desenvolver "habilidades" para Alexa, incluindo a personalização de intenções e a construção do modelo de interação. Um conjunto de intenções representa ações que os usuários podem realizar com suas habilidades. Essas intenções representam a funcionalidade principal da habilidade. As intenções personalizadas atendem às funções exclusivas de uma habilidade e cada uma possui um conjunto de asserções.

Neste trabalho analisamos perspectivas de interface de voz e imagem a partir do entendimento da corrente teórica do Interacionismo Simbólico que compreende que a realidade social é construída por meio de uma interação simbólica, na qual as pessoas interpretam e atribuem significado às ações, palavras umas das outras e, procuramos desvelar o sentido de que também poderia ocorrer com em conversações de pessoas e dispositivos IoTs com interação por voz.

No entanto, procuramos demonstrar que Modelos de Fundação (MF), originários das ciências da Computação, podem ser usados para melhorar a capacidade de entender e responder a comandos de voz em IVIU, incluindo a possibilidade de incorporar conceitos do Interacionismo Simbólico. Os Modelos de Fundação têm demonstrado habilidades avançadas de Processamento de Linguagem Natural (PLN), permitindo que eles entendam e gerem texto de maneira mais sofisticada. Isso pode ser aplicado para aprimorar a compreensão da fala e a resposta das interfaces de voz. Como um desenvolvimento do modelo PLN requer grandes quantidades de dados, algo que tem ajudado a treinar esses modelos é o uso de Geração de Linguagem Natural, que utiliza MF para aumentar e gerar conjuntos de dados de forma que possam ser utilizados para um modelo PLN desenvolvimento.

Perspectivas para design de interface de voz e imagem (IVIU) a partir do Interacionismo Simbólico

As interfaces de voz têm ganhado destaque como uma forma alternativa e natural de interação com dispositivos eletrônicos (ver PEREIRA et al, 2022). O design de interface de voz requer uma compreensão profunda das capacidades e limitações da interação por voz, por isso é essencial considerar elementos como a compreensão da linguagem natural, a adaptação a diferentes contextos e a personalização da experiência do usuário, como observado por Deibel e Evanhoé (2021). Além disso, o feedback auditivo adequado e a capacidade de corrigir erros são aspectos decisivos do design de interface de voz. Embora este campo seja centrado na interação e comunicação por meio da fala, a interface visual começa a desempenhar um papel importante em melhorar e complementar essa interação, uma vez que dispositivos de áudio com assistentes de voz começam a ganhar telas, como é o caso do Echo Show 15, ao mesmo tempo que interfaces visuais começam a ganhar assistentes de voz, como o Google Assistente. A interface visual, nesse sentido, fornece contexto e clareza para o usuário e, embora a voz possa fruir informações, a interface visual pode fornecer elementos visuais adicionais, como imagens, ícones, botões, menus e gráficos, para ajudar o usuário a construir sentidos melhores na interação.

Embora o design para voz e o design visual se concentrem em diferentes aspectos da experiência do usuário em um sistema IoT, e cada um tenha suas próprias características e considerações específicas, há uma série de pontos de interseção entre esses dois campos. Ambos se preocupam com a usabilidade, a facilidade de aprendizado e a satisfação do usuário. Tanto o design de voz quanto o design visual devem fornecer *feedback* adequado e orientação ao usuário, por meio de *prompts* de áudio e respostas verbais no sentido de informar e guiar os usuários, assim como utilizando elementos visuais como ícones, animações e *feedback* visual. Uma interface de voz pode ser combinada com uma interface visual para fornecer suporte adicional aos usuários, para, por exemplo, exibir informações complementares na tela enquanto a interação ocorre por voz. Elementos visuais, como indicadores de progresso, animações e dicas visuais, podem ajudar os usuários a entenderem o estado atual da interação e o que esperar a seguir. Isso cria uma experiência mais orientada e confiável. Por exemplo, durante uma interação de compra em um assistente de voz, a interface visual pode exibir imagens dos produtos, detalhes do preço e avaliações dos clientes. Essas informações visuais adicionais enriquecem a

experiência e facilitam a tomada de decisões do usuário. A interface visual ainda desempenha um papel importante na acessibilidade, porque nem todos os usuários podem ou preferem interagir apenas por voz. Isso permite que os usuários tenham opções adicionais de interação, como toques na tela, seleção por toque ou navegação visual, o que garante que pessoas com deficiência auditiva ou aqueles em ambientes barulhentos possam acessar e interagir com o sistema.

No entanto, o ambiente IoT atual está se configurando em um emaranhado de dispositivos distintos e de difícil compartilhamento de ativos gráficos e sonoros, o que torna o desenvolvimento de soluções de design em IoT um grande desafio (AKSU et al, 2018). São necessários projetos específicos para cada plataforma e com graus de complexidade distintos e que demandam muitas adaptações. Um desafio atual é criar um *framework* de produção que possa unir o grande leque de plataformas, dispositivos e sistemas diferentes em um processo coerente e de fácil compreensão para designers e demais profissionais interessados.

Uma plataforma como a Alexa, por exemplo, que conta com dispositivos unicamente de voz, como os Echo Dots, e com dispositivos com voz e imagem, como os Echo Shows, necessita de soluções de design responsivo e flexível para diferentes aplicações. Nos dispositivos exclusivamente com voz, a falta de contexto visual e a necessidade de projetar uma comunicação mais linear e sem tantas ambiguidades são alguns dos desafios que os profissionais enfrentam, como vemos em Deibel e Evanhoé (*ibidem*). Além disso, um reconhecimento de voz que seja preciso e uma capacidade de lidar com sotaques e variações linguísticas sem o apoio de telas são questões críticas a serem abordadas, pensando em dispositivos múltiplos, como observamos na pesquisa anterior já citada.

A interação necessita ser natural e intuitiva, tanto por voz como por imagem. Isso envolve o desenvolvimento de sistemas de reconhecimento de voz precisos, *feedback* adequado e concepção de interações de voz e imagem intuitivas. O objetivo é projetar interfaces que entendam e interpretem corretamente a linguagem falada pelos usuários, proporcionando uma experiência de interação em áudio e vídeo fluida e eficiente.

Para tanto, compreendemos que é necessário um alinhamento entre a comunicação, o design de voz e imagem, e a PLN. Esta está relacionada ao desenvolvimento de algoritmos e sistemas capazes de compreender e processar a linguagem humana em seus diferentes aspectos, incluindo a compreensão de significado, gramática, contexto e intenção. A PLN permite que os computadores entendam e interpretem comandos e

consultas em linguagem humana, facilitando a interação entre humanos e máquinas (JURAFSKY e MARTIN, 2000). A interseção entre IVIU e a PLN é aqui sugerida para criar sistemas de interação nos quais a PLN fornece uma base tecnológica necessária para reconhecer, interpretar e processar a fala dos usuários, enquanto que a IVIU compreenda e responda adequadamente às interações e solicitações feitas pelos usuários.

Além disso, a NPL desempenha um papel importante na personalização e adaptação das interfaces de voz às preferências individuais dos usuários. Segundo Jurafsky e Martin (2000), com técnicas avançadas de PLN, é possível criar sistemas de interface de voz que sejam capazes de aprender e se adaptar ao estilo de comunicação e preferências linguísticas de cada usuário, proporcionando uma experiência mais personalizada e agradável. Uma IVIU precisa lidar com questões de reconhecimento de fala de modo preciso, assim necessita compreender nuances linguísticas e tratamento de erros e ambiguidades, diferenciando, inclusive, pessoas de diferentes idades e tons de voz no mesmo grupo que utiliza o dispositivo.

Uma vez que um dos propósitos da comunicação entre pessoas e dispositivos é o desenvolvimento de interfaces que sejam fáceis de usar e compreendam as intenções e necessidades dos usuários, o Interacionismo Simbólico[8] (IS) tem uma forte contribuição nesse contexto, uma vez que esta corrente teórica se concentra na compreensão da sociedade por meio da análise das interações simbólicas entre pessoas. Blumer (1969), um dos principais expoentes do IS, nota que as pessoas atribuem significado aos objetos, eventos e situações por meio de processos interpretativos e comunicativos. A teoria sugere que a realidade social é construída por meio de uma interação simbólica, na qual as pessoas interpretam e atribuem significado às ações, palavras umas das outras. Há um certo consenso na Comunicação que o mundo simbólico só se constrói por meio da interação entre duas ou mais pessoas, sendo assim, o simbolismo não é resultado de interação de uma pessoa consigo e – não seria - mesmo com objetos[9]. Chamamos a atenção que atualmente muitos objetos usam a interação por inteligência artificial e que, portanto, *a priori* poderiam conduzir conversações com significados específicos. Naturalmente essas conversações são passíveis da ambiguidade dos algoritmos e dos contextos nas quais as IAs geram campos de significações. Este contexto ainda carece de

[8] Corrente teórica sociológica surgida na década de 1930, a partir de estudos de Mead (1934) e posteriormente Blumer (1969), também reconhecida como Escola de Chicago.
[9] Aqui sugerimos uma reinterpretação, que carece de aprofundamento, uma vez que as IVIU permitem uma interação similar às pessoas.

aprofundamento, como bem observado em trabalhos sobre o estudo do algoritmo, principalmente na publicidade (TRINDADE, PEREZ e TEIXEIRA FILHO, 2020).

O Interacionismo Simbólico originou-se no trabalho de George Herbert Mead, que demonstrou que os egos das pessoas são produtos sociais, sem deixar de terem propósitos e de serem criativos. Outros precursores na área foram Herbert Blumer e Charles Cooley. Blumer, um estudioso de Mead, e criador do termo "Interacionismo Simbólico", pôs em evidência as principais perspectivas dessa abordagem: as pessoas agem em relação às coisas baseando-se no significado que essas coisas tenham para elas; e esses significados são resultantes da sua interação social e modificados por sua interpretação.

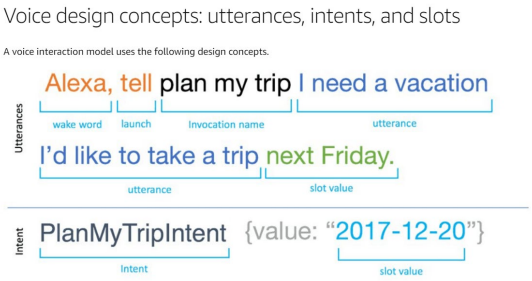
Blumer seguiu Mead e apresentou a proposição de que as pessoas interagem umas com as outras por meio de interpretação mútua das ações e definição um do outro, em vez de somente reagir às ações um do outro. Suas respostas não são dadas diretamente às ações um do outro, mas baseadas no significado que eles atribuem a tais ações. Assim, interação humana é mediada pelo uso de símbolos e significados, através de interpretação, ou determinação do significado das ações um do outro. Blumer comparou esse processo, que ele designou "interação simbólica", com as explicações behavioristas do comportamento humano que não consideram a interpretação entre estímulo e resposta, como em Schram (1949), entre outros. O que buscamos desvelar aqui é justamente a interpretação mútua que poderia ocorrer entre pessoas e dispositivos IoT com IVIU. Para tanto se faz necessário compreender como IoTs poderiam interpretar.

Desse ponto de vista, vale lembrar que a linguagem, componente desse processo, é um instinto inato e universalmente presente na espécie humana. Steven Pinker (2002) argumenta que a linguagem é uma adaptação biológica que surge naturalmente em nossa espécie e é moldada por processos evolutivos, independentemente das bases culturais, mas por essas modificada, ressignificada. Pinker entende que há uma gramática universal, que sugere que as pessoas têm uma predisposição inata para adquirir a linguagem, exemplificando como as crianças aprendem de forma rápida e eficiente, mesmo sem instrução formal. O conceito de gramática universal e a aquisição da linguagem pode ajudar a orientar o desenvolvimento de algoritmos e modelos na PLN assim como na compreensão dos processos de IVIU, para avançar o entendimento e desenvolvimento de designs que compreendam e se comuniquem de forma mais natural com as pessoas, como vemos no modelo de conversação.

Para agilizar o processo de desenvolvimento de *skills*[10] e garantir experiências de usuário consistentes, o uso de modelos de IVIU ganhou importância significativa. Os modelos de interface de voz fornecem uma base para a criação de interfaces de conversação, permitindo que os desenvolvedores se concentrem mais na funcionalidade do aplicativo, sem ter que começar do zero. Entende-se, assim, que desenvolver modelos de interação para dispositivos IoT será uma competência de profissionais de Comunicação, Design e Programação, desejada nos próximos anos. Aproveitando os modelos de IVIU, os comunicadores, desenvolvedores e designers podem economizar tempo, garantir consistência e oferecer uma experiência de voz de mais qualidade aos usuários, como percebe-se nos modelos pré-criados pela Amazon, como o Alexa Skills Kit[11] (ASK) que define um conjunto de palavras que os usuários dizem para invocar uma *skill*. Salienta-se, no entanto, a necessidade de uma avaliação crítica dos modelos existentes a partir do Interacionismo Simbólico.

Esses profissionais podem definir suas *skills* para aceitar as solicitações predefinidas. Observou-se que o ASK (figura 1) oferece vários tipos diferentes de *skills* pré-criadas para escolha, ou possibilita a construção de uma *skill* personalizada utilizando predefinições de declarações, intenções e *slots* dentro de um console de desenvolvedor Alexa. Para o desenvolvimento deste recorte, foi analisada a criação de uma *skill* para doação a uma instituição denominada Instituto Uno.

Figura 1: Conceitos do modelo ASK



Fonte: https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-the-interaction-model-for-your-skill.html

[10] Nome dado pela Amazon para aplicativos de voz que usam sistema Alexa.
[11] Disponível em: https://developer.amazon.com/en-US/docs/alexa/ask-overviews/what-is-the-alexa-skills-kit.html

Especificamente no ASK observou-se ser necessário compreender que há diferentes variáveis na construção de IVIU, como a **palavra de ativação,** que é a palavra para a Alexa começar a ouvir seus comandos, e normalmente os usuários não trocam essa palavra, mesmo sendo possível. Há a **palavra de lançamento,** que é uma palavra de ação de transição que sinaliza para o Alexa que uma invocação de *skill* provavelmente ocorrerá. Por exemplo, uma palavra de lançamento para fazer uma doação a uma instituição pode incluir "como fazer", "faça", "doe". Recomenda-se aqui uma pesquisa para descrever o máximo de palavras de lançamento possíveis para aquela intenção. Há nos modelos do ASK o **nome de invocação,** que é a expressão para começar a interagir com uma *skill*. Por exemplo, para usar a *skill* "Horóscopo Diário", um usuário pode dizer: "Alexa, abra meu horóscopo diário". Selecionar o nome de invocação, portanto, requer uma decisão fundamentada para o sucesso de uma *skill*. Um dos desafios atuais em IVIU é unir as intenções (como "quero fazer uma doação") com o nome de invocação de uma *skill* (como "abra Instituto Uno"). O modelo de IVIU na ASK também determinam que há diferentes **enunciados**, isto é**,** uma solicitação falada de um usuário, que pode invocar uma *skill*, fornecer entradas para uma *skill*, confirmar uma ação para a Alexa e assim por diante. Reforça-se aqui que é necessário aprofundar as várias maneiras pelas quais usuários podem formular suas solicitações por voz.

Naturalmente o modelo ASK possibilita a entrada de ***prompts***, isto é, uma pergunta para Alexa dirigir ao usuário para pedir informações. O desenvolvedor inclui o texto do *prompt* em sua resposta à uma solicitação de um usuário. Por exemplo, após responder a uma pergunta como "O que faz o Instituto Uno?", pode-se incluir ao final do texto descritivo o *prompt* "Para saber mais, diga Alexa, como fazer uma doação para o Instituto Uno?"

O modelo ASK a ser customizado solicita **intenções**, que são ações para atender à uma solicitação falada de um usuário. As intenções podem opcionalmente ter argumentos chamados *slots*. Os *slots* são **valores** de entrada fornecidos na solicitação falada de um usuário. Esses valores ajudam a Alexa a descobrir a intenção do usuário. Os *slots* podem ser definidos com diferentes *tipos*. O *slot* de data de viagem usa o tipo interno da Amazon para converter palavras que indicam datas (como "hoje" e "próxima sexta-feira") em um formato de data (AMAZON.DATE), enquanto de cidade e para cidade usam o *slot* (AMAZON.BR_CITY). Por exemplo, para *slots* de doação pode-se selecionar data, valor, para quem e de qual conta (figura 2).

Figura 2 – Estrutura de diálogo para doação por meio de uma skill Alexa

Proposta de diálogo para uma descoberta de doação
Alexa, gostaria de fazer uma doação.
(Alexa) Para quem você gostaria de fazer uma doação?
Alexa, me dê opções de ONGs
(Alexa) O Instituo Uno é dedicado à educação de crianças vulneráveis.
Alexa, me fale mais sobre o Instituto Uno
(Alexa) O Instituto Uno... gostaria de fazer uma doação?
Alexa, sim
Legenda por cores: palavra de ativação, palavra de lançamento, nome de invocação, enunciados, prompts, intenções, valores de slots

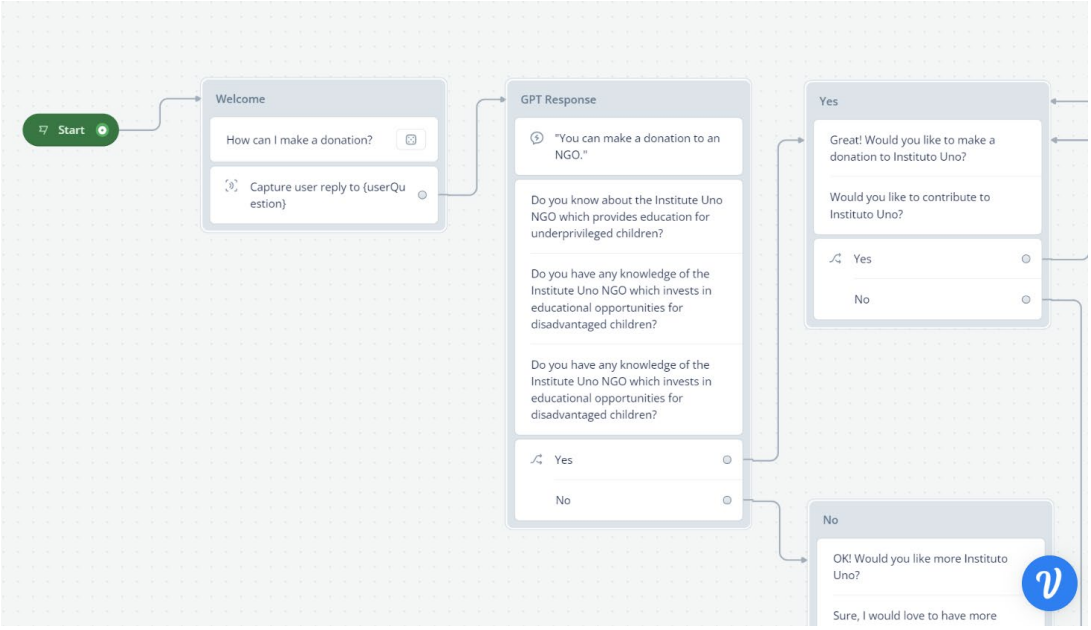
Proposta de diálogo para uma doação periódica
Alexa, gostaria de fazer uma doação.
(Alexa) Para quem você gostaria de fazer uma doação?
Alexa, para o Instituto Uno [memória]
(Alexa) Em que data você quer fazer sua doação?
Alexa, hoje
(Alexa) Qual o valor da sua doação?
Alexa, R$ 100,00
(Alexa) De qual conta você gostaria de debitar a doação?
Alexa, da minha
(Alexa) Vamos confirmar sua doação. Você confirma sua doação [hoje] para o [Instituo Uno], no valor de [R$ 100,00], da sua conta [nome da conta]?
Legenda por cores: palavra de ativação, palavra de lançamento, nome de invocação, enunciados, prompts, intenções, valores de slots

Fonte: os autores

O console do ASK salva um modelo de interação criado no formato JSON e o profissional pode editar o modelo com uma ferramenta de edição. Depois do modelo de interação JSON pronto, cria-se a função Lambda do serviço de computação em um console de gerenciamento da Amazon. Deve-se, por fim, selecionar a linguagem de programação desejada, o kit de desenvolvimento de software ASK correspondente (SDK) e a *skill* será codificada com Node.js, Python ou Java. Atualmente as *skill*s são hospedadas gratuitamente no serviço AWS Lambda para o primeiro milhão de chamadas por mês. O último passo é testar a *skill* no console do desenvolvedor Alexa que possui um simulador Alexa integrado, que é semelhante ao teste em um dispositivo habilitado para Alexa real. Entendemos que uma *skill* para fazer uma doação é complexa e tem muitas conversas de

ida e volta (conversação em vários turnos). Nesse caso, necessitamos criar um modelo de diálogo, que é uma estrutura que identifica as etapas de uma conversa de vários turnos entre sua *skill* e o usuário, como demonstrado na fig. 3 e que, neste caso, foi desenvolvida por meio do app Voiceflow:

Figura 3 – Modelo preliminar de diálogo para doação com uso de Alexa



Fonte os autores

Desse ponto de vista, parece-nos que faltam mais designs de IVIU que possam ser projetados para reconhecer e interpretar não apenas palavras isoladas, mas também os significados e intenções subjacentes às interações simbólicas entre pessoas e que o campo da Comunicação, por meio da Interação Simbólica, possa fornecer chaves de entendimento para aprofundar esses aspectos.

Como indagação, será que um design de IVIU baseado em linguagem natural que segue os princípios do Interacionismo Simbólico seria capaz de interpretar não apenas as palavras exatas de um usuário, mas também o contexto e o significado simbólico por trás delas? Do nosso ponto de vista, ele poderia considerar fatores como o contexto da conversa, as expressões emocionais e as pistas sociais para fornecer respostas mais

precisas e relevantes, o que permite uma interação mais rica e significativa entre o usuário e o sistema como um todo.

Aprendizado de Máquina e Modelos de Fundação

Sistemas como o Alexa utilizam grandes modelos de linguagem com base em aprendizado de máquina que são treinados em uma grande escala de dados, os tornando adaptáveis a uma variedade de tarefas diferentes, e podem ter como base arquiteturas de redes neurais artificiais e *transformers,* como veremos a seguir. O aprendizado de máquina depende de diferentes algoritmos para resolver problemas de dados. Os cientistas de dados apontam que não há tipos únicos de algoritmos que seja melhor para resolver um problema (MAHESH, 2020). O tipo de algoritmo empregado depende do tipo de problema que se deseja resolver, o número de variáveis, o tipo de modelo que melhor se adequa a ela e assim por diante. O aprendizado de máquina se utiliza de algoritmos de aprendizado supervisionado, que é a tarefa de aprender uma função que mapeia uma entrada para uma saída com base em exemplos, como vemos em Mahesh (*ibidem*). Ele infere uma função de treinamento de dados rotulados que consistem em um conjunto de exemplos de treinamento. O conjunto de dados de entrada pode ser dividido em treinamento e teste. O conjunto de dados do treinamento tem variável de saída que precisa ser previsto ou classificado. O algoritmo então, prevê algum tipo de padrão do conjunto de dados de treinamento e aplica-o ao conjunto de dados de teste para previsão ou classificação.

Já os algoritmos de aprendizado não supervisionados, ao contrário da aprendizagem supervisionada acima, não possui respostas corretas e não há conjuntos de dados de exemplo. São responsáveis por encontrar e apresentar uma estrutura nos dados por conta própria. Os algoritmos de aprendizado não supervisionado aprendem com poucos recursos a partir dos dados. Quando novos dados são introduzidos, ele usa os padrões previamente aprendidos para reconhecer a classe dos dados.

Entre os algoritmos aplicados ao aprendizado não-supervisionado, estão as redes neurais artificiais e estas, quando possuem múltiplas camadas internas, são chamadas de redes profundas. *Transformer,* por exemplo, é uma rede neural artificial profunda que emprega um mecanismo de auto-atenção para compreender as relações contextuais dentro de dados

sequenciais. Ao contrário das redes neurais convencionais ou versões atualizadas de redes neurais recorrentes (RNNs), como *Long Short-Term Memory* (LSTM), modelos de transformadores excelentes em lidar com longas dependências entre os elementos de sequência de entrada e habilitar o processamento paralelo (ISLAM *et. al*, 2023). Como resultado, modelos baseados em *transformers* têm atraído interesse substancial entre os pesquisadores no campo da inteligência artificial por conta do seu pontencial, não apenas em tarefas de PLN, mas também em uma ampla gama de domínios, incluindo visão computacional, áudio e processamento de fala, saúde e Internet das Coisas (IoT).

O trabalho de Bommasani *et al.* (2021), trouxe grande interesse na comunidade de pesquisa por apresentar os grandes modelos como Modelos de Fundação (MF), em suas oportunidades de aplicação e riscos. Os MFs demonstraram fortes capacidades de generalização e transferência de conhecimento como resultado do aprendizado de diversas fontes de dados, como diferentes modalidades, vários idiomas e vários domínios de aplicação. Nas aplicações em entendimento de voz, tem crescido o número de estudos de pesquisa mostrando resultados promissores e demonstrando as vantagens potenciais de tais modelos.

Na pesquisa de Narayanan *et al*. (2018), foi feita uma aplicação de um modelo de aprendizado em larga escala para reconhecimento de voz. O estudo abrange um grupo lógico de enunciados de voz que compartilham algumas características comuns. Os exemplos incluem domínios de aplicativos como pesquisa por voz, legendas de vídeo, callcenter, etc. E o treino desse modelo foi feito com um banco de dados de fala de 162.000 horas de arquivos de voz.

Dependendo se os dados de treinamento supervisionado são usados, podemos agrupar os modelos de fundação em duas categorias, modelos autossupervisionados e supervisionados (RAFFEL et al., 2020). Com o treinamento autos-supervisionado, os modelos são treinados primeiro em dados somente de áudio para aprender boas representações dos sinais de fala, depois são usados diretamente como recurso extratores para tarefas subsequentes. Nesse caso não são necessários dados rotulados, portanto essa abordagem pode ser facilmente ampliada para falas mais diversificadas, uma vez que há menos esforço humano de transcrição envolvido. O aprendizado supervisionado tem como pré-requisito ter alguns dados rotulados para tarefas nas quais os MFs são treinados (RAFFEL *et al*., 2020).

Há também a abordagem de explorar técnicas para modificar os dados de entrada para treinar um modelo de linguagem padrão em assistentes de voz fornecendo enunciados e intenções obtidas por meio de dados históricos (CHO, EUNJOON e KUMAR, 2018). Os autores obtiveram um aumento na precisão em uma tarefa de reconhecimento de fala para o Google Assistant, demonstrando que é possível obter melhorias no reconhecimento de fala para alterar a arquitetura de agentes de conversação[12], investigando a natureza assimétrica da interação com um assistente digital, onde as consultas do usuário são curtas e as respostas do agente são geralmente mais longas.

Outros estudos, visam a aplicabilidade de modelos de linguagem baseados em *transformers* disponíveis no mercado, como GPT-3 e GPT-J (WANG e KOMATSUZAKI, 2021) para aumentar os conjuntos de dados de treino para tarefas de classificação de intenção, quando o objetivo final é prever a intenção de um usuário de assistentes de voz, dado um enunciado. Aumento de dados para esse fim é particularmente desafiador porque o MF deve distinguir entre um grande número de intenções que podem ser semanticamente muito próximas umas das outras (SAHU *et al*., 2022).

É necessária uma variedade de declarações de amostra no modelo de interação porque isso ajuda a treinar o modelo de compreensão de PLN. As declarações de amostra ajudam a *skill* a atingir o objetivo do usuário. Como o desenvolvimento do modelo de PLN requer grandes quantidades de dados, e algo que tem auxiliado a treinar esses modelos é o uso da Geração de Linguagem Natural (MCDONALD, 2010), que utiliza de MFs para aumentar e gerar conjuntos de dados de forma que possam ser usados para o desenvolvimento do modelo de PLN.

No trabalho de Amin-Nejad é proposta uma metodologia que orienta a geração de conjuntos de dados com informações estruturadas usando MFs de última geração e demonstra um conjunto de dados aumentado capaz de superar a assertividade em relação a uma linha de base para um modelo de classificação (AMIN-NEJAD et al, 2020). Há espaço, portanto, para explorar a técnica de melhorar os conjuntos de dados de intenções por meio do Interacionismo Simbólico para treino dos modelos utilizados por uma Alexa, a fim de averiguar possíveis melhorias na assertividade de resposta em relação a interação com usuário.

[12] Similares a *skills*.

Os MFs podem ser usados para melhorar a capacidade de entender e responder a comandos de voz em IVIU. Os MFs, como o GPT-4, têm demonstrado habilidades avançadas de processamento de linguagem natural, permitindo que eles entendam e gerem texto de maneira mais sofisticada, embora carentes de revisões ainda. Isso pode ser aplicado para aprimorar a compreensão da fala e a resposta das interfaces de voz.

Considerações finais

A indagação que fazemos é se uma interface de voz que utiliza um MF associados ao Interacionismo Simbólico pode ser capaz de entender melhor comandos de voz complexos, responder a perguntas com mais precisão ou gerar respostas mais naturais e contextualmente relevantes. Além disso, esta associação poderia ser usada para melhorar a capacidade das interfaces de voz em realizar tarefas específicas, como fazer doações, fazer reservas ou fornecer informações detalhadas sobre determinado assunto.

Uma vez que os MFs são treinados em grandes quantidades de texto que contêm informações simbólicas e significados culturais e aprendem a compreender e gerar linguagem, incluindo símbolos e significados associados a palavras e frases é relevante observá-los por uma perspectiva teórica que se concentra na forma como os seres humanos desenvolvem o sentido e o significado por meio da interação social. Ao utilizar MFs em IVIUs, poderíamos explorar como esses modelos podem auxiliar na compreensão e na geração de linguagem simbólica. Por exemplo, em uma IVIU que segue os princípios do Interacionismo Simbólico, um MF pode ser usado para interpretar as mensagens dos usuários e gerar respostas que levem em consideração o contexto simbólico e cultural.

No trabalho de Zhong *et al* (2023), foram apresentadas avaliações das capacidades de grandes modelos de fundação com relação à cognição em nível humano. Baseando-se em exames de qualificação e concursos avançados, incluindo testes de admissão em faculdades de direito. Essas avaliações estabelecem padrões oficialmente reconhecidos para medir as capacidades humanas, tornando-os adequados para avaliar modelos de fundação no contexto de tarefas centradas no ser humano.

Ao avaliar estes modelos de fundação em tarefas centradas no ser humano e sondando suas capacidades mais profundamente, foram observados acertos significativos dos modelos. Algumas análises manuais aprofundadas no trabalho também revelarem as limitações desses grandes modelos de linguagem em termos de compreensão, utilização

do conhecimento, raciocínio e cálculo. E isso encoraja promover o desenvolvimento de modelos mais alinhados com a cognição humana, permitindo que eles lidem com uma gama mais ampla de tarefas complexas e centradas no ser humano com maior precisão e confiabilidade.

O *prompt* Chain-of-Thought (CoT) (WEI *et al*., 2022) permite que grandes modelos de linguagem quebrem uma questão complexa em uma série de passos de raciocínio decompostos. Para investigar ainda mais o raciocínio dos modelos de capacidades, os autores implementaram uma avaliação de raciocínio de "cadeia de pensamento". Este método envolve que o modelo primeiro gere uma explicação para uma pergunta dada e, posteriormente, responda à questão com base em sua explicação autogerada. Isso nos permite avaliar a capacidade do modelo de compreender tarefas complexas e identificar os elementos essenciais, como contexto e compreensão necessários para solução de problemas bem-sucedida.

Além disso, os modelos de fundação podem ser treinados em conjunto com dados que incluem interações sociais reais para capturar nuances e sutilezas da linguagem simbólica utilizada pelos seres humanos. Isso pode ajudar a tornar as interações mais naturalmente significativas e relevantes para os usuários.

No entanto, é importante ter em mente que os modelos de fundação são técnicas de aprendizado de máquina e não possuem compreensão ou intenção verdadeiras. Eles aprendem a identificar padrões nos dados de treinamento, mas até o momento, não possuem uma compreensão conceitual ou simbólica do mundo**.**

Portanto, ao utilizar esses modelos em aplicações baseadas no Interacionismo Simbólico, é necessário considerar suas limitações e garantir que a interpretação e a geração de símbolos sejam realizadas de maneira adequada e ética, levando em conta o contexto e a diversidade cultural.

Agradecimentos

Agradecemos aos organizadores do 46º Congresso Brasileiro de Ciências da Comunicação e ao Instituto Mauá de Tecnologia. Projeto "Comunicação, design e IoT: estudos para interface gráfica com design de voz". Edital de Apoio a Pesquisa 2023, Decisão 15883/17/23.

Referências

AKSU, Hydayet, *et al*. **Advertising in the IoT Era: Vision and Challenges**. Department of Electrical and Computer Engineering Florida International University, Miami, FL, USA. Recurso digital. **arXiv**, v1, 31 Jan 2018. ISSN 1802.04102. Disponível em https://arxiv.org/abs/1802.04102. Acesso em 20.01.2022

AMIN-NEJAD, A., IVE, J., e VELUPILLAI, S. Exploring transformer text generation for medical dataset augmentation. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4699-4708. 2020.

BLUMER, Herbert **Symbolic Interactionism**: Perspective and Method, Englewood Cliffs, Prentice-Hall, 1969.

BOMMASANI, R., HUDSON, D. A., "On the opportunities and risks of foundation models," **arXiv**:2108.07258, 2021

CHO, Eunjoon, e KUMAR, Shankar. "A conversational neural language model for speech recognition in digital assistants." 2018 **IEEE International Conference on Acoustics**, Speech and Signal Processing (ICASSP). IEEE, 2018.

DEIBEL, Diana; EVANHOÉ, Rebecca. **Conversations with Things**: UX Design for Chat and Voice New York: Rosenfeld Media, 2021. ISBN: 1-933820-26-8.

JURAFSKY, D. e MARTIN, J. H.. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000. 1st edition.

MAHESH, B., Machine Learninig algorithms-a review. **International Journal of Science and Research** (IJSR).[Internet], 9(1), pp.381-386. 2020

McDONALD, D.D. **Natural language generation**. Handbook of natural language processing, 2, pp.121-144, 2010.

MEAD, George Herbert. **Mind, self, and society**. Chicago: University of Chicago Press, 1934.

NARAYANAN, A.; MISRA, A.; SIM, K.C.; et al., "Toward Domain-Invariant Speech Recognition via Large Scale Training," in **Proc**. SLT, 2018.

PEREIRA, E.; MELLO, A. P. S. Internet das coisas como mídia: perspectivas e desafios para comunicação e design de anúncios. **Signos do Consumo**, [S. l.], v. 15, n. 1, p. e210974, 2023. DOI: 10.11606/issn.1984-5057.v15i1e210974. Disponível em: https://www.revistas.usp.br/signosdoconsumo/article/view/210974. Acesso em: 16 ago. 2023.

PEREIRA, E.; MELLO, A. P. S.; CORREA, J. A.; MELLITO, M. A. (2022). Comunicação, design e tecnologia: perspectivas e desafios da IoT como mídia para anúncios interativos. **Anais do 45º Congresso Brasileiro de Ciências da Comunicação**, 5 a 9 de outubro de 2022, E [recurso eletrônico]: Ciências da Comunicação contra a Desinformação / organizado por Giovandro Marcus Ferreira, Maria do Carmo Silva Barbosa e Norma Maria Meireles. ISSN 2175-4683.

PINKER, Steven. **O instinto da linguagem**. como a mente cria a linguagem. São Paulo: Martins Fontes, 2002. 640 p., ISBN-10 8533615493.

RAFFEL, C.; SHAZEER, N.; et al., "Exploring the limits of transfer learning with a unified text-to-text transformer.," **J. Mach. Learn. Res.**, vol. 21, no. 140, pp. 1–67, 2020.

SAHU, Gaurav, et al. "Data augmentation for intent classification with off-the-shelf large language models." **arXiv** preprint arXiv:2204.01959 (2022).

SCHRAMM, W. (Org.). **Mass Communications**. Urbana, IL: University of Illinois Press, 1949.

TRINDADE, E., PEREZ, C., & TEIXEIRA FILHO, C. (2019). Tendências das pesquisas em publicidade e consumos nos periódicos nacionais e internacionais de comunicação: um panorama sobre o estudo do algoritmo. In ***Anais***... Porto Alegre: Escola de Comunicações e Artes, Universidade de São Paulo. Recuperado de https://www.eca.usp.br/acervo/producao-academica/002961807.pdf

WANG, Ben e KOMATSUZAKI, Aran. **GPTJ-6B**: A 6 Billion Parameter Autoregressive Language Model. 2021 https://github.com/ kingoflolz/mesh-transformer-jax.

WEI, Jason; WANG, Xuezhi; SCHUURMANS, Dale; BOSMA, Maarten; CHI, Ed; LE, Quoc; ZHOU, Denny. Chain of thought prompting elicits reasoning in large language models. **arXiv** preprint arXiv:2201.11903, 2022.

ZHONG, W., CUI, R., GUO, Y., LIANG, Y., LU, S., WANG, Y., SAIED, A., CHEN, W. e DUAN, N. Agieval: A human-centric benchmark for evaluating foundation models. **arXiv** preprint arXiv:2304.06364. 2023