



Programa Interdisciplinar de Pós-Graduação em
Computação Aplicada
Mestrado Acadêmico

Jonathan Simon Prates

Gerenciamento de Diálogo Baseado em Modelo Cognitivo para
Sistemas de Interação Multimodal

São Leopoldo, 2015

Jonathan Simon Prates

GERENCIAMENTO DE DIÁLOGO BASEADO EM MODELO COGNITIVO PARA
SISTEMAS DE INTERAÇÃO MULTIMODAL

Dissertação apresentada como requisito parcial
para a obtenção do título de Mestre pelo
Programa de Pós-Graduação em Computação
Aplicada da Universidade do Vale do Rio dos
Sinos — UNISINOS

Orientador:
Prof. Dr. Sandro José Rigo

São Leopoldo
2015

P912g

Prates, Jonathan Simon

Gerenciamento de diálogo baseado em modelo cognitivo para sistemas de interação multimodal / Jonathan Simon Prates. – 2015.

91 f. : il. color. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2015.

"Orientador: Prof. Dr. Sandro José Rigo."

1. Processamento de linguagem natural (Computação).
2. Sistemas especialistas (Computação). 3. Interação homem-máquina. 4. Inteligência artificial. I. Título.

CDU 004.8:800.8

Jonathan Simon Prates

Gerenciamento de diálogo baseado em modelo cognitivo para sistemas de interação multimodal

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em ____/____/____

BANCA EXAMINADORA

Prof. Dr. Francisco José Serón Arbeloa - UNIZAR

Prof. Dr. Jorge Luis Victória Barbosa - UNISINOS

Prof. Dr. Sandro José Rigo (Orientador)

São Leopoldo,

Prof. Dr. Cristiano André da Costa
Coordenador PPG em Computação Aplicada

Aos nossos pais.

*If I have seen farther than others,
it is because I stood on the shoulders of giants.*
— SIR ISAAC NEWTON

AGRADECIMENTOS

Agradeço a Deus por me guiar nesta trajetória. Aos meus pais por me dar o apoio sempre quando preciso.

Agradeço ao meu orientador, professor Dr. Sandro Rigo, por ter aceitado guiar esse trabalho com calma e dedicação.

Agradeço à minha namorada, professora Simone Perotto, por me ajudar e me dar forças quando precisei e dedicar seu tempo me ajudando nesses anos.

Aos professores Atila Vasconcelos, Isabel Siqueira e Wilson Gavião por me apoiar, me incentivar e acreditar em mim. Essa jornada não teria se realizado sem eles.

Agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa durante todo o período de realização deste mestrado.

Agradeço aos professores da Unisinos por compartilhar o seu tempo e conhecimento conosco durante esses dois anos.

Os meus sinceros agradecimentos a todos vocês!

“Ninguém abre um livro sem que aprenda alguma coisa”.
(Anônimo)

RESUMO

Os Sistemas de Interação Multimodal possibilitam uma utilização mais amigável dos sistemas de computação. Eles permitem que os usuários recebam informações e indiquem suas necessidades com maior facilidade, amparados por recursos de interação cada vez mais diversos. Neste contexto, um elemento central é o diálogo que se estabelece entre os usuários e estes sistemas. Alguns dos desafios observados na área de Interação Multimodal estão ligados à integração dos diversos estímulos a serem tratados, enquanto outros estão ligados à geração de respostas adequadas a estes estímulos. O gerenciamento do diálogo nestes sistemas envolve atividades diversas associadas tanto com a representação dos assuntos tratados, como com a escolha de alternativas de resposta e com o tratamento de modelos que representam tarefas e usuários.

A partir das diversas abordagens conhecidas para estas implementações, são observadas demandas de modelos de diálogo que aproximem os resultados das interações que são geradas pelos sistemas daquelas interações que seriam esperados em situações de interação em linguagem natural. Uma linha de atuação possível para a obtenção de melhorias neste aspecto pode estar ligada à utilização de estudos da psicologia cognitiva sobre a memória de trabalho e a integração de informações.

Este trabalho apresenta os resultados obtidos com um modelo de tratamento de diálogo para sistemas de Interação Multimodal baseado em um modelo cognitivo, que visa proporcionar a geração de diálogos que se aproximem de situações de diálogo em linguagem natural. São apresentados os estudos que embasaram esta proposta e a sua justificativa para uso no modelo descrito. Também são demonstrados resultados preliminares obtidos com o uso de protótipos para a validação do modelo. As avaliações realizadas demonstram um bom potencial para o modelo proposto.

Palavras-chave: Processamento de linguagem natural (Computação). Sistemas especialistas (Computação). Interação homem-máquina. Inteligência artificial.

ABSTRACT

Multimodal interaction systems allow a friendly use of computing systems. They allow users to receive information and indicate their needs with ease, supported by new interaction resources. In this context, the central element is the dialogue, established between users and these systems. The dialogue management of these systems involves various activities associated with the representation of subjects treated, possible answers, tasks model and users model treatment.

In implementations for these approaches, some demands can be observed to approximate the results of the interactions by these systems of interaction in natural language. One possible line of action to obtain improvements in this aspect can be associated to the use of cognitive psychology studies on working memory and information integration.

This work presents results obtained with a model of memory handling for multimodal dialogue interaction based on a cognitive model, which aims to provide conditions for dialogue generation closer to situations in natural language dialogs. This research presents studies that supported this proposal and the justification for the described model's description. At the end, results using two prototypes for the model's validation are also shown.

Keywords: Natural Language Processing (Computing). Expert Systems (Computing). Human-computer interaction. Artificial Intelligence.

LISTA DE FIGURAS

Figura 1:	Protótipo de sistema de interação multimodal VOX.	27
Figura 2:	Arquitetura de um sistema multimodal e seus principais componentes. . . .	32
Figura 3:	Uma arquitetura típica de sistemas pergunta e resposta.	34
Figura 4:	Modelo da memória operacional proposto por Baddeley e Hitch (1974). . . .	37
Figura 5:	W3C Multimodal Architecture.	39
Figura 6:	EMMA - emotion representation.	40
Figura 7:	EMMA - location representation.	41
Figura 8:	AIML: Regra para saudação.	44
Figura 9:	AIML: Utilização de caracteres especiais.	45
Figura 10:	AIML: Utilização em código Python.	45
Figura 11:	Fluxograma geral da arquitetura do Florence.	52
Figura 12:	Arquitetura do modelo proposto.	56
Figura 13:	Modelo proposto: Fluxograma das etapas de controle do diálogo.	60
Figura 14:	Modelo proposto: Máquina de estado de controle de diálogo.	61
Figura 15:	Representação da entrada no formato EMMA.	63
Figura 16:	Tela de gerenciamento do RabbitMQ.	64
Figura 17:	Trecho de um diálogo exemplo gerado pelo protótipo 1.	65
Figura 18:	Protótipo 2 - Esperando entrada por voz do usuário.	66
Figura 19:	Entrada AIML em memória com base na regra e valor da ontologia.	69
Figura 20:	Diálogo gerado no cenário 1 pelo protótipo.	72
Figura 21:	Representação do RDF de Quincas Borba destacando a propriedade author. .	73
Figura 22:	Diálogo gerado no cenário 2 pelo protótipo, em continuação do cenário 1. .	74
Figura 23:	Representação do diálogo gerado no cenário 3 pelo protótipo.	75
Figura 24:	Resultados dos formulários de avaliação.	78
Figura 25:	Wikipedia - Artigos por idioma até 2008.	81

LISTA DE TABELAS

Tabela 1:	Trabalhos relacionados	53
Tabela 2:	Percepção da comparação entre as respostas geradas pelo sistema e das esperadas pelos usuários.	77

LISTA DE SIGLAS

AIML	Artificial Intelligence Markup Language
AMQP	Advanced Message Queuing Protocol
API	Application Programming Interface
ASR	Automatic Speech Recognition
EML	Emotion Markup Language
EMMA	Extensible MultiModal Annotation
FMS	Finite Machine State
FXML	Florence XML
GPS	Global Positioning System
IA	Inteligência Artificial
IHC	Interação Humano-Computador
IMM	Interação Multimodal
IP	Internet Protocol
json-LD	JavaScript Object Notation for Linking Data
LOD	Linked Open Data
MIA	Multimodal Interaction Activity
MMIF	Multimodal Interaction Framework
MOM	Message Oriented Middleware
NLTK	Natural Language Toolkit
PLN	Processamento de Linguagem Natural
QA	Question Answering
REM	Reconhecimento de Entidades Mencionadas
RDF	Resource Description Framework
TAM	Technical Architecture Modeling
TCP	Transmission Control Protocol
TREC	Text REtrieval Conference
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
WIMP	Windows, Icons, Menus, Pointer
XML	eXtensible Markup Language

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Comunicação, diálogo e memória	24
1.2	Interação humano-computador e sistemas de interação multimodal	25
1.3	Motivação	26
1.4	Questão de pesquisa	28
1.5	Objetivos	28
1.6	Metodologia	29
1.7	Organização do documento	29
2	EMBASAMENTO TEÓRICO	31
2.1	Interação multimodal	31
2.2	Sistemas pergunta e resposta	33
2.3	Processamento de linguagem natural	35
2.4	Computação ciente de contexto	35
2.5	Gerenciamento de diálogo e memória de trabalho	36
2.6	Recursos	38
2.6.1	Arquitetura do W3C e outros <i>frameworks</i>	38
2.6.2	Dados abertos ligados	41
2.6.3	Processamento distribuído e tratamento de mensagens	42
2.6.4	Linguagem de programação Python e NLTK	43
2.6.5	Processador pyAIML e Linguagem AIML	44
3	TRABALHOS RELACIONADOS	47
3.1	Mudra: A Unified Multimodal Interaction Framework	47
3.2	The SEMAINE API: towards a standards-based framework for building emotion-oriented systems	48
3.3	MATCH: An Architecture for Multimodal Dialogue Systems	49
3.4	Context Based Multimodal Fusion	49
3.5	MIND: A context-based multimodal interpretation framework in conversational systems	50
3.6	Multimodal human-robot interaction with Chatterbot system: Extending AIML towards supporting embodied interactions	51
3.7	Florence: a Dialogue Manager Framework for Spoken Dialogue Systems	51
3.8	Comparativo e considerações	52
4	MODELO PROPOSTO	55
4.1	Entrada	55
4.2	Controle	57
4.3	Saída	59
4.4	Gerenciamento do diálogo	59
5	ASPECTOS DA IMPLEMENTAÇÃO DE PROTÓTIPOS	63
5.1	Detalhes de implementação - Protótipo 1	63
5.2	Detalhes de implementação - Protótipo 2	65

6 AVALIAÇÃO	71
6.1 Avaliação por cenários	71
6.2 Avaliação com usuários	76
7 CONCLUSÃO	79
7.1 Limitações e trabalhos futuros	80
REFERÊNCIAS	83
ANEXO A FORMULÁRIO ENVIADO AOS PARTICIPANTES DO EXPERIMENTO.	89
ANEXO B DADOS RECEBIDOS DOS PARTICIPANTES DURANTE O EXPERIMENTO.	91

1 INTRODUÇÃO

O avanço no desenvolvimento de dispositivos de interação humano-computador vem trazendo um conjunto bastante amplo de possibilidades para a realização de entrada e saída de dados nos sistemas de computação. Esta diversidade é explorada nos sistemas conhecidos como sistemas de Interação Multimodal, que possibilitam uma utilização mais amigável dos sistemas computacionais. Eles permitem que os usuários recebam informações e indiquem suas necessidades com maior facilidade e adequação, amparados por recursos de interação cada vez mais diversos e pela capacidade de integração destes dispositivos de acordo com as necessidades de comunicação.

Estes sistemas são formados por componentes diversos, tais como os módulos de entrada, módulos de integração de informações, módulos de gerenciamento de diálogo e de apresentação das mensagens, além dos módulos para tratamento dos dispositivos de saída de dados (BUI, 2006). Do ponto de vista de sua utilização, o elemento central é o diálogo que se estabelece entre os usuários e estes sistemas. O diálogo é uma tarefa complexa que, embora seja realizado com naturalidade, envolve diversas funções cognitivas. Em sistemas de Interação Multimodal, o gerenciamento do diálogo envolve atividades associadas com a representação dos assuntos tratados, a escolha de alternativas de resposta, o tratamento de outras tarefas e a interação com o usuário.

Com base nas diversas abordagens descritas nos trabalhos de Dumas, Lalanne e Oviatt (2009) e Jaimes e Sebe (2007) para implementações de sistemas de interação multimodal, são observadas demandas por melhorias no âmbito do gerenciamento do diálogo. Estas demandas basicamente estão associadas aos resultados das interações geradas durante uma conversação, ou seja, durante uma interação em linguagem natural.

Neste trabalho, destaca-se como aspecto determinante o contexto do diálogo, termo que é utilizado aqui para representar o conjunto de elementos relevantes citados ao longo das interações. O contexto pode estar relacionado com diversas informações, tais como o local onde o diálogo está ocorrendo, os elementos citados nas mensagens anteriores e que podem ser considerados como o assunto principal, ou ainda as entidades ou pessoas que foram referenciadas. Uma melhoria possível de ser alcançada em sistemas de Interação Multimodal consiste em aperfeiçoar a geração do diálogo, precisamente com a utilização de informações do contexto (DEY, 2001) durante a conversação entre homem e computador. Considerando a busca de recursos para que o usuário perceba a interação de uma forma mais robusta, aproximando essa experiência como o diálogo entre duas pessoas, acredita-se que alguns pontos como a identificação do contexto e a utilização de discursos anteriores podem ajudar a avançar neste sentido.

A linha de atuação para a obtenção de melhorias neste aspecto pode estar ligada à utilização de estudos da psicologia cognitiva sobre a memória de trabalho e a integração de informações. A partir do conhecimento disponível sobre os mecanismos utilizados pelos seres humanos nestas situações de diálogo (IZQUIERDO, 2011), torna-se possível a simulação de alguns de seus

aspectos em contextos de sistemas de Interação Multimodal. Este trabalho realiza a investigação sobre o uso de um modelo de gerenciamento do diálogo para sistemas de Interação Multimodal baseado em um modelo cognitivo, que visa proporcionar condições de geração de diálogo mais próximas de situações de diálogo em linguagem natural. Para contextualizar a proposta de trabalho e para apresentar o embasamento introdutório ao assunto, serão apresentados a seguir, os aspectos gerais relacionados com a comunicação e os processos de diálogo e sua aplicação na interação humano-computador e sistemas de Interação Multimodal.

1.1 Comunicação, diálogo e memória

A multimodalidade na comunicação mostra que informações do mesmo significado, muitas vezes, podem ser expressas em diferentes formas semióticas¹. No início do século XX, livros, pinturas e apresentações artísticas mostravam uma forma única de comunicação, levantando críticas teóricas sobre o quão essas artes eram monomodais, pois cada forma continha seus próprios métodos, suas próprias hipóteses e argumentos. Ao romper esse paradigma, emerge um conceito onde múltiplos métodos semióticos juntos podem fornecer informações com mais qualidade, pois é da natureza humana utilizar diversos meios, simultaneamente, para se comunicar. A comunicação multimodal é importante para construir o significado, de forma clara, sem ambiguidades, através de informações que vão além da fala ou da escrita, expressas também através de imagens, emoções e sentimentos que podem construir o significado de tal informação (KRESS; LEEUWEN, 2001).

Os seres humanos se expressam através de uma linguagem, com uso de palavras, faladas, escritas ou gesticuladas. O homem é capaz de criar sentenças com essas palavras, dando forma ao discurso. Para a linguística, o discurso representa uma sequência coordenada de frases, que não se limita somente à fala, mas também envolve aspectos culturais e sociais (POPPEL, 1989). Vestir uma determinada roupa, por exemplo, já é um modo de comunicação gerada pelo ser humano. Durante o discurso, o emissor deve utilizar palavras de forma que a comunicação seja realizada com eficácia. As partes envolvidas devem estar em sintonia com a ideia transmitida para que a informação faça sentido. Assim, a argumentação é responsável por fortalecer a ideia transmitida entre o emissor e o receptor (ENCICLOPÉDIA, 1991).

Durante um diálogo, a memória é a responsável por manter um fluxo na conversação, preservando a lucidez para as partes envolvidas, com base em discursos e argumentações anteriores (GODOY, 2010). Um dos elementos estudados nesta área é a memória de trabalho. Para a psicologia, existem algumas definições sobre a memória de trabalho. O modelo de Baddeley e Hitch (1974), o mais aceito e estudado na psicologia cognitiva, diz que a memória de curto prazo é responsável por guardar informações temporariamente durante o seu processamento. Essa memória tem como característica principal ser limitada, guardando apenas pequenas informações

¹Semiótica é a ciência que estuda os sinais ou sistemas de sinais utilizados na comunicação e o seu significado (SEMIÓTICA, 2003)

relacionadas com o momento que está sendo vivenciado pela pessoa.

A memória como um todo é um processo complexo que apoia a manutenção dos aspectos associados com a consciência (IZQUIERDO, 2011). Ela é um importante componente cognitivo envolvido na compreensão e entendimento durante a comunicação. Essas características tão fundamentais para o diálogo entre seres humanos também são estudadas na área de interação humano-computador e utilizadas como base para diversos trabalhos.

Nesta pesquisa são analisadas as características do modelo proposto por Baddeley e Hitch (1974) e seu uso como parte do componente responsável pelo controle de diálogo em um sistema de Interação Multimodal. A abordagem foi adotada neste trabalho por sugerir a possibilidade de obtenção de melhorias na interação humano-computador, ligadas à utilização de estudos da psicologia cognitiva sobre a memória de trabalho e a integração de informações.

1.2 Interação humano-computador e sistemas de interação multimodal

A interação humano-computador (IHC) é definida como a ciência que estuda a interação entre pessoas e computadores (STONE et al., 2005). Observa-se que cada dia mais pessoas, inclusive crianças, conseguem obter resultados satisfatórios ao interagir com dispositivos tais como um *tablet*, um computador ou até mesmo interfaces de gestos. Na década de 70, quando os primeiros computadores pessoais foram lançados, somente pessoas com alto nível de conhecimento técnico conseguiam utilizar de forma proveitosa tal interação (WALKER, 1990). Segundo Neto et al. (2009), a multimodalidade na comunicação, característica tão natural entre humanos, pode ainda ser bastante ampliada nos sistemas IHC. Nas últimas décadas, vem sendo estimulado o desenvolvimento de sistemas suportando múltiplas entradas de dados simultâneas, tais como gestos, fala e expressões.

O desenvolvimento de uma interface multimodal é complexo e os processos e recursos para realizar esse trabalho também. De acordo com Cutugno et al. (2012), a interação multimodal é definida por sistemas que buscam alcançar as habilidades de comunicação humanas utilizando a fala, gestos, expressões faciais, toque e emoções. Nas últimas décadas, o uso de interfaces multimodais tornou-se parte cada vez mais desejada em sistemas de IHC. Isso se deve em parte devido à busca por atender usuários com necessidade especiais, mas também está impulsionado pela crescente evolução tecnológica de dispositivos e sensores, fomentando a possibilidade de dotar os sistemas de uma capacidade de interação mais próxima da interação realizada entre os seres humanos.

O conceito de interface multimodal teve seu início no final dos anos 70, com o notável trabalho de Bolt (1980). A interface “*Put That There*” permitia mover objetos em uma tela, combinando a fala e o apontamento de um dos braços. O processamento acontecia ao receber a palavra “*there*” junto com as coordenadas do local apontado. Recentemente, sistemas multimodais já são capazes de processar gestos em três dimensões, junto com uma análise de linguagem natural e interpretação cognitiva, porém ainda existe muito a ser pesquisado nestas áreas (OVI-

ATT, 2003). Trabalhos e experimentos recentes descrevem a utilização da captura dos cinco sentidos humanos (audição, visão, tato, olfato e paladar) até mesmo o reconhecimento de padrões de ondas cerebrais por meios de eletrodos para captar a intenção do usuário (OVIATT; COHEN, 2000).

Atualmente, existe uma variedade de implementações de sistemas de interação multimodal voltadas para diversas finalidades. Algumas interfaces possuem agentes virtuais conversacionais, com o objetivo de aproximar a percepção da interação humano-computador de uma interação entre humanos (CASSELL et al., 2000). Este efeito nem sempre é obtido, em especial devido aos formatos de renderização das interfaces e da representação dos agentes. Existem algumas implementações mais sofisticadas onde o sistema pode expressar emoções através de um rosto humano (F. J. SERÓN C. BOBED, 2014; BALDASSARRI; CERESO; SERON, 2008; CERESO et al., 2008). Dessa forma, a interface pode realizar movimentos de lábios e expressões faciais durante a fala ou quando está recebendo uma entrada de dados originada por uma ação dos usuários, melhorando a experiência durante o diálogo.

Gerenciamento de diálogo é uma área de pesquisa emergente que busca proporcionar melhorias significativas para os sistemas de interfaces humano-computador. Desenvolvida a partir dos anos 70, essa área engloba a simulação, manutenção e execução de diálogos em sistemas conversacionais. Alguns conceitos têm grande importância para a geração de discurso entre sistemas e humanos. Muitos sistemas utilizam uma abordagem linguística com processamento de linguagem natural para entendimento da entrada e geração das respostas para o usuário (HARTSON; HIX, 1989).

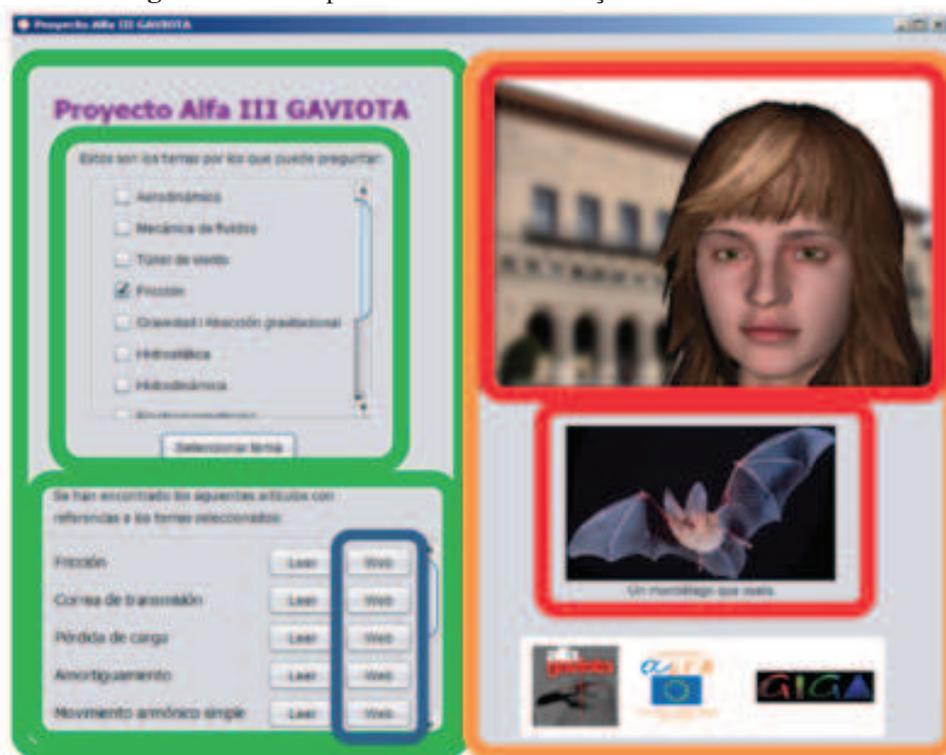
1.3 Motivação

Sistemas de Interação Multimodal oferecem uma melhor experiência para seus usuários, permitindo a estes a possibilidade de agir de forma natural, mais próxima da interação entre pessoas. Na maioria dos sistemas de conversação entre humanos e computador, para que o reconhecimento seja possível e o sistema realize as ações de forma correta, os usuários necessitam informar palavras e expressões em formatos predefinidos, que serão interpretadas por expressões regulares (LIE et al., 1998), utilizar *templates* (DWIVEDI; SINGH, 2013) ou ainda frases que atendam pré-requisitos (BUSH et al., 2001). Assim, podemos dizer que as interfaces desses sistemas de Interface Natural de Usuário² não podem ser propriamente consideradas naturais, devido ao seu distanciamento das situações cotidianas de comunicação.

Alguns sistemas de interação multimodal têm como característica principal o modelo de interação com base em perguntas e respostas, nos quais de fato não existe um diálogo (considerado como uma troca de informações entre as partes), mas sim uma pesquisa utilizando os termos informados como valores de entrada. Existem sistemas de interação multimodal em muitas

²Interface Natural do Usuário refere-se a uma interface de sistemas onde o usuário utiliza gestos, fala e outros métodos naturais para interagir com o sistema, evitando a utilização de botões, controle, teclados e outros dispositivos.

Figura 1: Protótipo de sistema de interação multimodal VOX.



Fonte: The Vox System (F. J. SERÓN C. BOBED, 2014)

áreas de atuação. Muitos deles são sistemas pergunta e resposta, seja em dispositivos móveis ou em agentes conversacionais. Um destes sistemas é o VOX, desenvolvido na Universidade de Zaragoza, na Espanha (F. J. SERÓN C. BOBED, 2014). Ele tem como objetivo localizar respostas para perguntas do usuário através do uso de recursos diversos de interação multimodal, acessando automaticamente as bases de conhecimento, tal como a DBpedia (AUER et al., 2007). A figura 1 ilustra alguns de seus componentes, onde pode ser observado uma parte da interface dedicada para a exibição de um agente comunicacional (no quadro superior à direita) que proporciona uma interação com o usuário a partir de síntese de voz das mensagens e da geração automática de animações com base nos textos gerados.

Este sistema foi utilizado como base para este estudo, a partir da identificação da possibilidade de avanços em aspectos relacionados com os diálogos gerados, que em determinados contextos poderiam não ser percebidos de forma amigável pelo usuário. O modelo proposto neste trabalho foi desenvolvido de modo a permitir sua integração em diversos sistemas de interação multimodal.

Para que exista de fato um diálogo, a memória de trabalho é um elemento necessário à interação natural entre humano e computador. Da mesma forma que o usuário mantém informações sobre o contexto da conversa, o computador deve ser capaz de fazer o mesmo e utilizar esta informação para que o diálogo possa acontecer com maior naturalidade. Outro aspecto importante do diálogo é a linguagem corporal, uma forma conhecida e estudada de comunicação não verbal entre os seres humanos. Os gestos e expressões acabam complementando a conversação. Essa

habilidade de entender o diálogo considerando não apenas palavras é um desafio para sistemas de interação multimodal e sistemas de interface natural.

Segundo Izquierdo (2011), a memória de trabalho é responsável pelo raciocínio, compreensão e aprendizagem. Esse componente funciona a partir da manutenção de um histórico para armazenamento da informação. Durante um diálogo, essa memória é fundamental para manter em foco as informações dos discursos anteriores. Acredita-se que a simulação de aspectos conhecidos sobre o funcionamento da memória de trabalho nos seres humanos pode trazer como resultado um diálogo mais natural durante a interação humano-computador, na percepção do usuário.

O presente trabalho tem como objetivo propor um modelo para o gerenciamento do diálogo, baseado em uma simulação do conceito de memória de trabalho dos seres humanos, possibilitando o desenvolvimento de sistemas de Interação Multimodal em que o usuário possa manter uma troca contínua de mensagens compondo uma conversa que seja percebida pelo usuário como mais próxima das interações naturais entre seres humanos. A motivação para esta linha de trabalho está relacionada com as possíveis vantagens do uso de um modelo de memória baseado em estudos da psicologia cognitiva. Com base na literatura (JAIMES; SEBE, 2007; DUMAS; LALANNE; OVIATT, 2009; SEBE, 2009; TURK, 2014) é possível observar que esta abordagem não é frequente em trabalhos nesta área.

1.4 Questão de pesquisa

Diante do exposto, a seguinte questão de pesquisa foi formulada:

O gerenciamento do diálogo em sistemas de interação multimodal com base em um modelo cognitivo, que leve em conta aspectos conhecidos do funcionamento da memória de trabalho, proporciona melhores condições de geração de diálogos durante a utilização desses sistemas?

1.5 Objetivos

O objetivo geral desta dissertação é avaliar a utilização do conhecimento a respeito do funcionamento de memória de trabalho em atividades de gerenciamento de diálogo em sistemas de interação multimodal. Este objetivo geral será alcançado a partir dos seguintes objetivos específicos:

1. Propor um modelo para gerenciamento do diálogo que possa ser utilizado em sistemas de interação multimodal, com base no modelo cognitivo de memória de trabalho descrito por Baddeley e Hitch (1974) e Baddeley (2000);
2. Implementar e avaliar o modelo de gerenciamento do diálogo proposto, com ênfase na investigação dos ganhos obtidos com a memória de trabalho e informações do contexto,

utilizados no sentido de manter um diálogo natural e consistente entre sistema e usuário.

1.6 Metodologia

Este trabalho consiste em uma pesquisa com objetivo exploratório, com finalidade de descrever e caracterizar aspectos do problema de pesquisa definido (KÖCHE, 2002). Uma maior familiaridade e maior quantidade de elementos associados com a descrição e natureza do problema tratado serão evidenciados com base nas pesquisas de referências bibliográficas, na experimentação proporcionada pelo protótipo a ser desenvolvido e testes a serem realizados. A pesquisa é de natureza aplicada, pois destina-se ao tratamento de um caso específico delimitado, sem derivação de características genéricas de aplicação, ficando delimitada à descrição, implementação e avaliação do modelo de gerenciamento de diálogo para um sistema de interação multimodal, conforme proposto.

Com relação ao método de trabalho (WAZLAWICK, 2009), foram definidas cinco etapas para realizar os objetivos propostos. São elas: (I) Ampliar os estudos sobre modelos cognitivos relacionados ao funcionamento da memória de trabalho e sobre o seu impacto em modelos de diálogo; (II) Descrever um modelo para implementação de um sistema de interação multimodal; (III) Descrever um modelo de gerenciamento de diálogo para contemplar a simulação de aspectos conhecidos do modelo cognitivo estudado; (IV) Desenvolver um protótipo funcional que permita a sua utilização como um sistema de interação multimodal; (V) Realizar a avaliação dos resultados obtidos, com base na utilização de aspectos quantitativos e qualitativos.

Nas etapas metodológicas descritas deve ser destacada a diferenciação entre a etapa II e III. A etapa II está destinada à descrição de um sistema completo para interação multimodal, considerando-se os aspectos mais gerais deste tipo de sistema. Nesta etapa serão seguidas diretrizes para que o sistema resultante possua flexibilidade suficiente para utilização em conjunto com diversas fontes de entrada e de saída de dados, de forma modular. Já na etapa III serão detalhados os mecanismos internos que irão delinear o funcionamento do gerenciamento do diálogo, de acordo com o modelo proposto neste estudo. Para a avaliação quantitativa foram realizadas etapas de utilização do sistema para buscar evidências de sucesso em relação a métricas relativas ao funcionamento esperado, através de cenários. Para a avaliação qualitativa foram definidos testes de utilização dos protótipos por usuários e coletados e analisados os resultados e as impressões dos usuários.

1.7 Organização do documento

O presente trabalho está organizado da seguinte maneira: no capítulo 2 é apresentado o embasamento teórico dos principais assuntos abordados. No capítulo 3 são descritos os trabalhos relacionados que possuem abordagem similar ou propostas da mesma área de pesquisa. Nos capítulos 4 é apresentado o modelo proposto para resolver o problema em questão e no capí-

tulo 5 a implementação do modelo em protótipos. Já o capítulo 6 versará sobre os resultados das avaliações baseada em cenários e teste com usuários. Para concluir, o capítulo 7 trará a conclusão, limitações do modelo proposto e trabalhos futuros.

2 EMBASAMENTO TEÓRICO

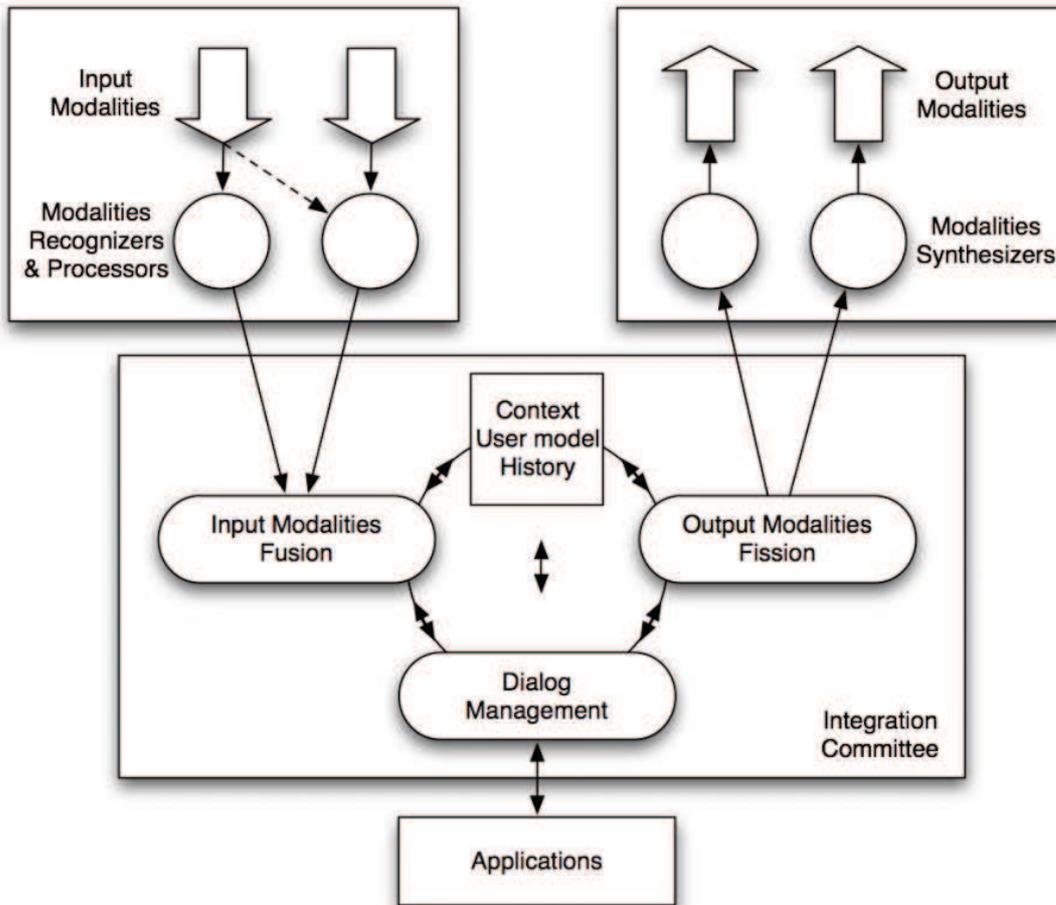
Esse capítulo apresenta o embasamento teórico dos pontos abordados nessa dissertação. Serão descritos fundamentos da IHC, analisando aspectos de sistemas de interação multimodal e sua utilização para diversos fins, como por exemplo, sistemas de interação para dispositivos móveis, sistemas baseados em agentes conversacionais, sistemas multimodais de pergunta e resposta e a utilização de dados ligados nesses sistemas. Em seguida é apresentada uma visão geral sobre sistemas pergunta e resposta, área de conhecimento responsável pela recuperação da informação durante a interação com os usuários. O processamento de linguagem natural e aspectos da computação ciente de contexto também são abordados, pois são áreas de conhecimento que proporcionam recursos para flexibilizar as possibilidades de entrada de dados pelo usuário no sistema. Para concluir, apresenta-se uma visão geral sobre sistemas que possuem gerenciamento de diálogo e sobre os recursos técnicos que são utilizados nesse trabalho.

2.1 Interação multimodal

Segundo Oviatt e Cohen (2000), existe uma mudança de paradigma na criação das interfaces de sistemas, onde se busca uma maior compreensão do comportamento de comunicação do ser humano. A multimodalidade em sistemas IHC vem se destacando principalmente devido ao uso de dispositivos móveis. É cada vez mais usual a utilização de aparelhos que permitem o controle pelo teclado, voz e toque ao mesmo tempo, junto com a utilização de dados de localização do usuário (ROQUE; PEIXOTO, 2011). Existem sistemas que possuem interfaces multimodais para muito objetivos. Destacam-se a seguir alguns cenários (JAIMES; SEBE, 2007):

- **Ambientes inteligentes:** Locais que recebem comando de voz, lidam com identificação dos usuários e utilizam sensores para compreender o contexto. Nesse caso, a entrada do usuário geralmente é feita de forma transparente ou invisível (WEISER, 1991).
- **Usuários com deficiências:** Pessoas com necessidades especiais podem ter grandes benefícios, uma vez que os sistemas não dependem mais das interfaces convencionais para funcionar. Esses usuários podem se expressar através de emoções, reconhecimento de gestos e sensores, entre outros.
- **Assistentes virtuais:** Programas que podem ajudar o usuário automatizando tarefas em quiosques e bancas de informação, por exemplo. Nesses casos, o sistema pode auxiliar através de um agente conversacional.
- **Dispositivos móveis:** As interfaces convencionais de sistemas são limitadas em dispositivos móveis. Com teclados e telas menores foi necessário o desenvolvimento de outros modos de entrada. Nessas situações, aplicativos diversos tratam outras formas de interação.

Figura 2: Arquitetura de um sistema multimodal e seus principais componentes.



Fonte: Multimodal interfaces: A survey of principles, models and frameworks (DUMAS; LALANNE; OVIATT, 2009)

Sistemas de interação multimodal combinam entradas de múltiplos dispositivos durante a interação com o usuário. Essas interfaces multimodais são diferentes das interfaces tradicionais conhecidas em *desktops* e vêm crescendo rapidamente nos últimos anos devido a melhorias de *hardware* e *software*. As pesquisas voltadas a esta área têm como objetivo desenvolver tecnologias para novos métodos de interação, trazendo um avanço na IHC e permitindo ao ser humano comunicar-se com o sistema como se estivesse conversando com outra pessoa. A interação multimodal envolve múltiplas áreas de pesquisa, incluindo engenharia da computação, ciências sociais, linguística, entre outras (TURK, 2014).

O objetivo principal dos sistemas de interação multimodal é proporcionar uma interação eficiente e natural a quem o utiliza. Segundo Oviatt e Cohen (2000), esses sistemas devem permitir modos de entradas distintos e integrados, ser eficientes, suportar entradas de dados simples e complexas, ter precisão para compreender coordenadas espaciais oriundas de um dispositivo de ponteiro ou até mesmo de direções faladas pelo usuário. Precisam ainda dar alternativas para métodos de interações ao usuário, se este preferir métodos alternativos ou possuir algum tipo de deficiência. Esses sistemas também devem evitar erros e permitir a adaptabilidade de acordo com a mudança do ambiente e do contexto.

De acordo com Blattner e Glinert (1996), as modalidades sensoriais humanas mais relevantes para a interação multimodal são divididas em quatro partes: visual, auditiva, tato e demais modalidades sensoriais. Alguns exemplos da modalidade visual seriam a identificação do rosto do usuário, gestos, expressões ou linguagem de sinais. A entrada falada e outros tipos de áudio, como música, são exemplos de entradas auditivas. Entradas através dos movimentos das mãos, dedos ou outros dispositivos de ponteiro são comparadas ao tato. Já as demais modalidades sensoriais podem incluir localização, termômetros, sensores de posição, entre outros.

Segundo Dumas, Lalanne e Oviatt (2009), os componentes em arquiteturas de sistemas de interação multimodal mais comumente encontrados são: um componente para a fusão multimodal de entrada, um gerenciador de diálogo e um gerenciador de contexto (figura 2). A entrada de dados é realizada através de vários componentes de reconhecimento. Esses componentes enviam dados processados para o componente de fusão, capaz de gerar uma interpretação única de todas as entradas. Logo após a fusão das entradas, a informação já interpretada é entregue ao gerenciador de diálogo, responsável por identificar e controlar os estados do diálogo. Já o gerenciador de contexto irá mapear o ambiente e perfil do usuário, bem como possíveis mudanças que possam ser relevantes para a próxima interação.

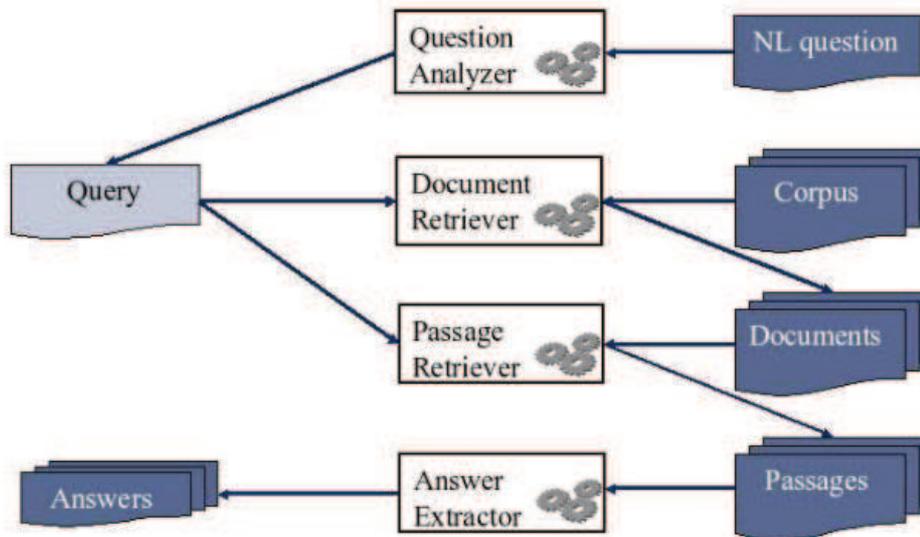
Sistemas como “*Put that there*” (BOLT, 1980) e *Quickset* (COHEN et al., 1997) são trabalhos importantes em que o conceito de interação multimodal é apresentado, através de gesto e fala e de toque, respectivamente. Mais recentemente, pesquisas com novos formatos de entradas de dados estão trazendo novos desafios. Novos experimentos consideram o estudo do comportamento do usuário, características sociais e culturais, entre outros detalhes importantes (SEBE, 2009). Como citado anteriormente, existem sistemas de interação multimodal em muitas áreas de atuação. Este trabalho é voltado a sistemas de interação multimodal destinados a pergunta e resposta (YEH; DARRELL, 2008; F. J. SERÓN C. BOBED, 2014).

2.2 Sistemas pergunta e resposta

Question Answering (QA), em português Pergunta e Resposta, é uma área de pesquisa que envolve outros campos, tais como recuperação da informação, extração da informação e processamento de linguagem natural (ALLAM; HAGGAG, 2012). Sistemas de recuperação da informação baseiam-se geralmente em buscas por palavras-chaves e exibem os resultados classificados a partir de uma ordem de relevância determinada estatisticamente. Com bases neste resultado, ainda é necessário que o usuário procure dentro destes documentos sugeridos até encontrar a resposta para sua pergunta.

Sistemas de pergunta e resposta têm como objetivo responder a uma pergunta formulada em linguagem natural, analisando diversas fontes de dados e documentos, e tentar obter a resposta mais exata para a questão (MISHRA; BANGALORE, 2010). Esses sistemas possuem características diferentes de sistemas de busca por palavras-chaves. A consulta realizada pelo usuário é feita através de uma pergunta direta, como por exemplo: “*Quem é o presidente do Brasil?*”. Em

Figura 3: Uma arquitetura típica de sistemas pergunta e resposta.



Fonte: Comparing Improved Language Models for Sentence Retrieval in Question Answering (MERKEL; KLAKEW, 2007)

sistemas de busca por palavras-chaves, serão usadas algumas palavras desta frase, por exemplo, “*presidente*” e “*Brasil*”, para identificar documentos onde se constata a sua ocorrência.

O resultado gerado por sistemas pergunta e resposta não deve trazer documentos que levem à resposta, e sim, a resposta em si. Devido à complexidade de identificar respostas exatas, técnicas de processamento de linguagem natural são empregadas para atingir esse objetivo. Muitos sistemas executam essa tarefa em duas etapas. A primeira consiste em identificar os parágrafos de textos dos documentos que são candidatos a respostas. A segunda etapa tenta encontrar as respostas exatas com base nos parágrafos candidatos (LI et al., 2002).

Os sistemas pergunta e resposta podem ser classificados como *closed-domain* ou *open-domain* (LIN, 2002). Sistemas *closed-domain* têm o objetivo de encontrar respostas para questões sobre um determinado domínio. Estes sistemas podem fazer uso de ontologias sobre aquele assunto a fim de obter as respostas de forma precisa. Já sistemas pergunta e resposta *open-domain* tentam responder questões sobre vários domínios. Esses sistemas podem usar diferentes abordagens para encontrar uma resposta satisfatória. A arquitetura mais comum (figura 3) encontrada entre sistemas pergunta e resposta é constituída por três etapas fundamentais executadas por componentes distintos: a classificação da questão (*question processing*), a recuperação de informação (*information retrieval*) e o processamento e extração da resposta (*answer processing*).

O componente de classificação das questões é fundamental para o sistema pergunta e resposta. Ele é responsável por encontrar o tópico da pergunta. Já o componente responsável pela recuperação de informação obtém os documentos ou registros das fontes de dados do sistema que se encaixam com o tópico reconhecido. Por fim, a fase de extração de respostas (*answer extractor*) classifica as possíveis respostas e as validam (GUPTA; GUPTA, 2012). De acordo

com Voorhees e Harman (2000), a *Text Retrieval Conference - TREC* fez crescer o número de pesquisas nas últimas décadas e um progresso significativo em sistemas pergunta e resposta.

2.3 Processamento de linguagem natural

Um dos desafios de praticamente todos os sistemas de interação multimodal é o reconhecimento, interpretação e inferência da fala ou da entrada de texto do usuário. O significado da entrada falada, escrita ou por linguagem de sinais (gestos), é um comando explícito, ou seja, o usuário quer passar a informação ao sistema. Nesse caso, o processamento de linguagem natural é uma peça fundamental para o gerenciamento do diálogo.

O Processamento de Linguagem Natural (PLN) é uma tarefa de Inteligência Artificial (IA) crucial para sistemas que necessitam inferir a entrada de dados do usuário. Utilizando-se de técnicas e padrões linguísticos, o PLN procura automatizar a compreensão da linguagem humana. O PLN pode analisar dados em nível fonético (relacionando palavras através de seus sons), em nível morfológico (analisando a formação da palavra e seu radical), em nível sintático (considerando o papel de cada palavra na frase), em nível semântico (buscando o significado da palavra através de combinações com outras) e em nível pragmático (analisando o contexto para inferir ou mudar o significado) (GONZALES; LIMA, 2003).

Sistemas de interação multimodal com características de sistemas pergunta e resposta podem utilizar o PLN junto com uma abordagem estatística e/ou baseada em regras ou *templates* para localizar a melhor resposta ao usuário. É importante citar que alguns desses sistemas possuem sensores informando dados extras, obtidas do próprio ambiente e que podem mudar o contexto da frase já tratada e analisada por técnicas de PLN.

O PLN é fundamental para identificação do significado de uma entrada escrita ou falada. Nesses sistemas, o analisador léxico e morfológico é responsável por identificar a função gramatical de cada palavra na frase. O analisador sintático é responsável pela análise e construção de uma estrutura sintática da sentença de entrada. Já o analisador semântico interpreta os componentes da sentença, buscando localizar sujeito, substantivo, verbos e a quem eles se referem. É de responsabilidade do analisador pragmático interpretar a intenção da sentença de acordo com o contexto da frase (LIDDY, 2001). O PLN é uma tarefa complexa, uma vez que muitos idiomas possuem tratamentos específicos de acordo com a cultura social (GROSSMAN, 2004). A maioria dos algoritmos de PLN possui um alto custo computacional, logo, os sistemas devem personalizar seu uso de acordo com cada necessidade.

2.4 Computação ciente de contexto

Uma característica importante dos sistemas de interação multimodal é a entrada de comandos e dados de forma implícita. Devido a evolução de possibilidades de geração de informações em contextos diferenciados (DEY, 2001), os sistemas atualmente podem utilizar dados vindos

de sensores e dispositivos a fim de obter informações sobre o ambiente, bem como dados do usuário com quem está interagindo. As entradas podem ter fontes variadas (gestos, fala, piscar de olhos etc.) captadas por diferentes dispositivos (JAIMES; SEBE, 2007). Portanto, a computação ciente de contexto tem um papel importante em sistemas de interface multimodal. Os seres humanos podem conversar sobre temperatura do ambiente ou localização geográfica sem citar o contexto do assunto, pois são capazes de saber se está frio ou calor e em que cidade estão.

A computação ubíqua permite a sistemas multimodais adaptarem o comportamento e interações futuras, de acordo com informações obtidas do ambiente em que o usuário está. Ela permite que os sistemas capturem, de forma imperceptível e automatizada, ações do usuário e do contexto. As mais simples informações de ambiente podem ser obtidas através de GPS, termômetro ou relógio, entre outros. Essas informações são consideradas entradas implícitas (NOCK; IYENGAR; NETI, 2004).

A computação ubíqua tem como objetivo tornar a interação do usuário invisível (WEISER, 1991), ou seja, as pessoas não se dão conta de que estão utilizando tecnologias integradas no seu dia-a-dia. Os comandos dados pelos usuários são implícitos e os computadores, portados com sistemas específicos e inteligentes, estariam em comunicação constantemente. A computação ciente de contexto utiliza diversos recursos a fim de melhorar a interação com o usuário. Destacam-se, nesse aspecto, técnicas para definir a localização do usuário, como triangularização de antenas, proximidade relativa e análise de cenas, bem como a utilização de dispositivos de GPS, rede de computadores local, radar e dispositivos de rádio frequência (HIGHTOWER; BRUMITT; BORRIELLO, 2002).

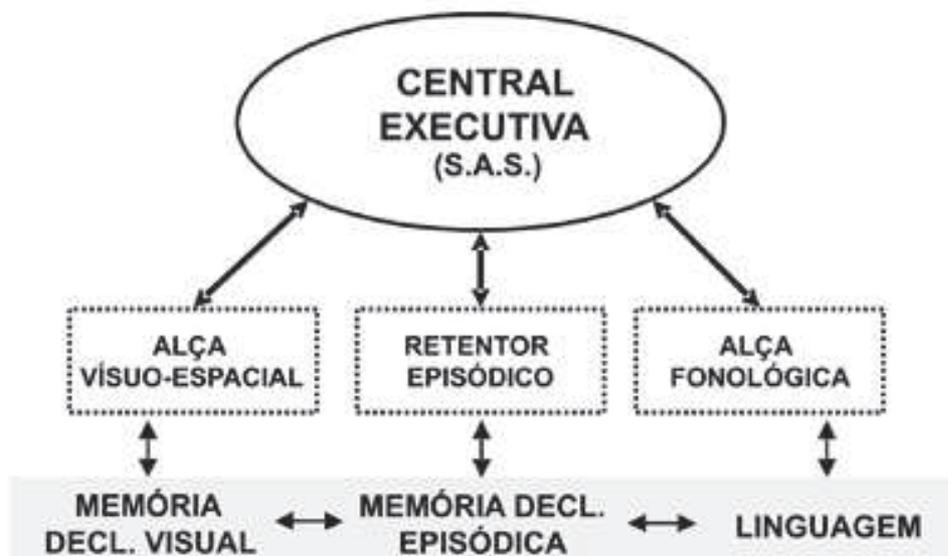
Outro ponto importante está relacionado à dificuldade de personalização e privacidade dos usuários em locais com mais pessoas. Nesse aspecto a computação vestível (*wearable computing*) possui características que complementam a computação ubíqua e permitem uma melhor identificação de cada usuário (RHODES; MINAR; WEAVER, 1999). Em sistemas de interação multimodal é possível observar requisitos similares, como a importância da identificação, personalização e informações do contexto.

2.5 Gerenciamento de diálogo e memória de trabalho

Neste item serão resumidos aspectos conhecidos sobre o funcionamento da memória de trabalho nos seres humanos e possíveis impactos deste funcionamento nas atividades de gerenciamento de diálogo. Estes subsídios pretendem ser utilizados no presente trabalho como forma de qualificar o módulo de gerenciamento de diálogo a ser desenvolvido.

A memória dos seres humanos possui uma tarefa muito complexa e é um dos principais elementos responsáveis por garantir a consciência de cada indivíduo. A memória é fundamental para localizar informações passadas e, a partir disso, tomar decisões. Ela também é responsável por garantir o senso de continuidade e lucidez, tão necessário para os humanos (GODOY, 2010).

Figura 4: Modelo da memória operacional proposto por Baddeley e Hitch (1974).



Fonte: A construção da atenção a partir da memória (HELENE; XAVIER, 2003)

De acordo com Baddeley e Hitch (1974), a memória de trabalho, um componente cognitivo relacionado com a memória, é responsável pelo raciocínio, compreensão e aprendizagem. Esse componente funciona como a manutenção de um pequeno histórico de informação. Durante um diálogo, esse componente é fundamental para manter as informações dos discursos anteriores. O modelo da memória operacional proposto pelos autores afirma que pelo menos três subsistemas de apoio são responsáveis pelo processamento da memória de curto prazo.

Segundo Baddeley e Hitch (1974) e Helene e Xavier (2003), o modelo da memória operacional possui uma central executiva responsável por coordenar os subsistemas de apoio. Os três subsistemas são: a alça visuo-espacial, capaz de guardar informações adquiridas através de imagens, como cores, tamanhos e localização de determinado objeto; a alça fonológica, que tem como função processar informações linguísticas; e um retentor episódico, um subsistema de capacidade limitada que é capaz de buscar a informação da memória de longa duração. É tarefa da central executiva direcionar a atenção e descartar informações não mais relevantes (figura 4). A memória é capaz de armazenar palavras e discursos anteriores para manter uma conversação.

Para Jurafsky e Martin (2000), a formação da linguagem possui divisões mais específicas do que palavras e frases. Um discurso não é formado por informações isoladas em frases soltas, mas sim por um grupo de sentenças diretamente relacionadas. Discurso é um termo genérico utilizado para definir um grupo de sentenças em uma linguagem, podendo ser unidirecional, conhecido como monólogo, ou multidirecional, conhecido como diálogo. A maneira em que as sentenças são combinadas define a estrutura do discurso (GROSZ; SIDNER, 1986). Muitos trabalhos sobre sistemas conversacionais descrevem uso de um gerenciador de diálogo implementado conforme a teoria destes autores (ROTARU, 2008).

O papel do gerenciador de diálogo no modelo de Grosz e Sidner (1986) é representar o

contexto linguístico, as informações interpretadas do usuário e o plano para respostas do sistema. As interações anteriores são armazenadas em um modelo chamado trilha de diálogo. Esse componente atua como um histórico das ações. Através da trilha de diálogo é possível gravar informações para restaurá-las (GROSZ; WEINSTEIN; JOSHI, 1995). O trabalho de Chai, Pan e Zhou (2005) utiliza uma memória que armazena o contexto, domínio e interações anteriores para aprimorar a experiência na conversação. O contexto, a memória, o discurso e demais características citadas nessa sessão são fundamentais para o trabalho desenvolvido nessa dissertação.

2.6 Recursos

Neste seção serão detalhados os recursos técnicos utilizados nesse trabalho. Os recursos aqui descritos são utilizados para desenvolvimento e validação do modelo proposto. A utilização efetiva de cada uma destas tecnologias para o desenvolvimento do protótipo está descrita de forma detalhada nos capítulos 4 e 5.

2.6.1 Arquitetura do W3C e outros *frameworks*

Um sistema de interação multimodal deve possuir diversos componentes para tratar cada um dos itens citados em suas definições de arquitetura (capítulo 4). Ainda não existe um padrão definido e amplamente adotado para a criação e desenvolvimento de componentes para uma arquitetura multimodal. Entretanto, desde 2002, o W3C (*World Wide Web Consortium*)¹ trabalha em uma proposta a fim de padronizar os elementos e componentes para sistemas de interação multimodal na web, através da *Multimodal Interaction Activity*. Além da iniciativa do W3C, a utilização de *APIs*² e *frameworks* pode ser visto em diversos trabalhos (VO; WOOD, 1996; JOHNSTON; BANGALORE; VASIREDDY, 2001; REITHINGER; SONNTAG, 2005; SCHRÖDER, 2010) devido a facilidade de reuso e modularização.

Mantida pelo W3C, a *Multimodal Interaction Activity*³ é uma iniciativa que procura estender a *web* para que os usuários possam utilizar os sistemas da maneira que seja mais apropriado a eles, incluindo usuários com deficiências ou limitações. Essa padronização deve permitir aos desenvolvedores a criação de interfaces mais eficazes e para um público maior. O W3C *Multimodal Interaction (MMI) Working Group* vem há mais de uma década trabalhando em uma proposta para a padronização de desenvolvimento de aplicações multimodais. Em 2005, o grupo publicou o *Multimodal Architecture and Interfaces Framework (MMIF)*⁴, aceito como recomendação em 2012. A proposta define os componentes básicos e suas principais caracte-

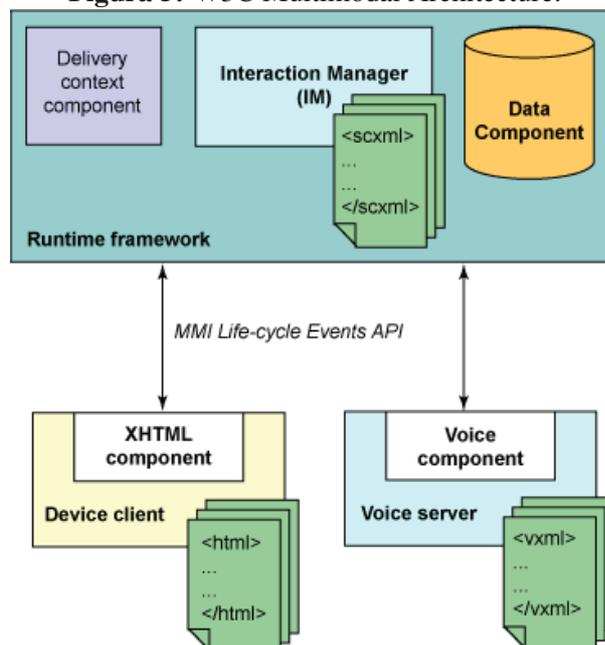
¹O W3C é um consórcio que visa promover padrões para a evolução e interoperabilidade na *web*. É formado por membros de empresas e instituições de desenvolvimento de software e demais tecnologias relacionadas à *web*.

²*Application Programming Interface (API)* é um conjunto de padrões, funções e rotinas de um software expostas para a utilização em outros serviços.

³<http://www.w3.org/2002/mmi/>

⁴<http://www.w3.org/TR/mmi-arch/>

Figura 5: W3C Multimodal Architecture.



Fonte: The W3C Multimodal Architecture, Part 1: Overview and challenges (MCCOBB, 2007)

rísticas (BARNETT et al., 2012).

A grande motivação do trabalho desenvolvido pelo W3C se dá pelo rápido crescimento do número de computadores pessoais e dispositivos capazes de entender outras formas de interação, não apenas baseadas no teclado e *mouse*. O trabalho especifica sua arquitetura de forma distribuída para que tarefas complexas, como reconhecimento de fala e reconhecimento de expressões, possam ser processadas em um servidor ou computador remoto ao invés de serem processadas por um dispositivo pequeno, com *hardware* limitado. Em uma visão geral e resumida, o *framework* está dividido em:

- ***Runtime framework***: controla a comunicação entre os componentes sistema.
- ***Delivery context***: armaneza as condições do contexto e preferências do usuário.
- ***Interaction manager***: responsável por manter o fluxo de diálogo e estado atual da interação.
- ***Data component***: contém o modelo de dados.
- ***Modality components***: componentes responsáveis por tarefas como reconhecimento de fala, captação de imagens, entre outras. São fracamente acoplados e devem interagir com os demais componentes. A tecnologia utilizada nesses componentes deve ser a mais apropriada para cada caso.

O *framework* descrito na figura 5 propõe que a comunicação entre os componentes deve ser realizada utilizando o formato XML e EMMA. O formato *eXtensible Markup Language*

Figura 6: EMMA - emotion representation.

```

<emma:emma
  version="2.0"
  xmlns:emma="http://www.w3.org/2003/04/emma"
  xmlns="http://www.example.com/example"
  xmlns:emo="http://www.w3.org/2009/10/emotionml">
  <emma:group id="result1"
    emma:start="1241035886246"
    emma:end="1241035888246"
    emma:medium="visual"
    emma:mode="video"
    emma:verbal="false">
    <emma:interpretation id="emo1">
      <emo:emotion dimension-set="http://www.w3.org/TR/emotion-voc/xml#pad-dimensions"
        category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">
        <emo:dimension name="arousal" value="0.3"/>
        <emo:dimension name="pleasure" value="0.9"/>
        <emo:dimension name="dominance" value="0.8"/>
        <emo:category name="bored" confidence="0.1"/>
      </emo:emotion>
    </emma:interpretation>
    <emma:group-info>emotion</emma:group-info>
  </emma:group>
</emma:emma>

```

Fonte: Elaborada pelo autor

(XML), recomendação do W3C para o desenvolvimento e criação de informações organizadas de forma hierárquica, é uma linguagem de marcação extensível que permite definir os elementos dessa marcação (BRAY et al., 2004). Utilizando o XML, a *Extended Multimodal Annotation Language (EMMA)* é um conjunto de especificações para sistemas de interação multimodal. Essa especificação pode ser usada para representar dados falados, gestos, expressões e demais entradas possíveis em sistemas de interação multimodal. A utilização do formato EMMA permite separar o processo de reconhecimento da entrada, da fusão e do processamento dos dados (JOHNSTON et al., 2009).

O EMMA foi criado com o objetivo de padronizar a troca de informações, possibilitando a combinação dos dados oriundos de múltiplas fontes. A especificação descreve as marcações que devem ser utilizadas antes, durante e após a interpretação das entradas. Essas marcações contêm as informações geradas pelo componente de captura de entrada juntamente com a taxa de confiança e a marcação temporal de quando o evento ocorreu (JUNIOR; REIS, 2007).

A figura 6 mostra um exemplo da representação no formato EMMA, extraída pelo componente responsável pela análise da emoção. Nesse XML, a marcação `<emma:emma>` representa a abertura da marcação principal, ou *root tag*, com seu devido *namespace* identificado pela URI `http://www.w3.org/2003/04/emma`. A marcação seguinte, `<emma:group>`, é um elemento que serve como agrupador de uma série de interpretações de entradas do usuário. Já a marcação `<emma:interpretation>` é a informação gerada pelo componente responsável pela leitura da entrada. No exemplo da figura 6, pode ser visto que o XML dentro da interpretação está no formato *Emotion Markup Language (EML)*⁵.

⁵<http://www.w3.org/TR/2014/REC-emotionml-20140522/>

Figura 7: EMMA - location representation.

```

<emma:emma version="1.1"
  xmlns:emma="http://www.w3.org/2003/04/emma"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2003/04/emma
  http://www.w3.org/TR/2009/REC-emma-20090210/emma.xsd"
  xmlns="http://www.example.com/example">
  <emma:location
    latitude="42.361860"
    longitude="-71.091840"
    altitude="6.706"
    accuracy="20.5"
    altitudeAccuracy="1.6"
    heading=""
    speed=""
    description="W3C MIT office"
    address="32 Vassar Street, Cambridge, MA 02139 USA"/>
  </emma:location>
  <emma:interpretation id="nlul"
    emma:medium="acoustic"
    emma:mode="voice"
    emma:tokens="flights from boston to denver">
    <origin>Boston</origin>
    <destination>Denver</destination>
  </emma:interpretation>
</emma:emma>

```

Fonte: Emma: Extensible multimodal annotation markup language (JOHNSTON et al., 2009)

A figura 7, mostra uma integração baseada na localização e reconhecimento de fala do usuário. Nesse caso, a marcação `<emma:location>` possui dados como latitude, longitude, altitude, endereço e outras informações relevantes. A entrada do usuário, representada em `<emma:interpretation>` foi falada por ele. O usuário estava no endereço informado no atributo `address` durante a ação.

Como o W3C é composto por importantes atores da área de tecnologia atuando nesta temática e suas propostas de padrões tendem a ser estáveis, completas e de ampla adoção, considerou-se relevante utilizar os recursos propostos para manter maior interoperabilidade com o modelo aqui desenvolvido. O padrão EMMA foi adotado para a troca de mensagens entre componentes nesse trabalho.

2.6.2 Dados abertos ligados

O W3C também é responsável pela definição de padrões e iniciativas para divulgação de dados ligados na *web*. Esse conceito, consiste na implementação de um conjunto de práticas para a publicação de informação na *web* de maneira estruturada. Sugerido por Berners-Lee, Hendler e Lassila (2001) e mantido pelo W3C, o projeto *Linked Open Data (LOD)* pretende estruturar as informações encontradas na *web*, fazendo ligações e referências entre elas. Essas informações são armazenadas no formato *Resource Description Framework (RDF)*, uma recomendação do W3C para representação de dados ligados (BIZER; HEATH; BERNERS-LEE, 2009). Essas tecnologias fazem parte do projeto *Semantic Web*, uma iniciativa que pretende tornar os dados

na *web* legíveis para homens e máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001).

Atualmente, muitos trabalhos com sistemas de interação multimodal, pergunta e resposta e agentes conversacionais têm dados abertos ligados como base de conhecimento (BOBED; ESTEBAN; MENA, 2012; DAMOVA et al., 2013; SHEKARPOUR; NGONGA NGOMO; AUER, 2013; BERANT et al., 2013; HAKIMOV et al., 2013; DAMLJANOVIĆ et al., 2013). A DBPe-dia é um projeto aberto que busca extrair informações estruturadas da Wikipedia⁶ e disponibilizar esta informação na *web*, possibilitando assim a realização de consultas em dados ligados de forma que possam ser entendida por humanos e outros sistemas (AUER et al., 2007).

Recentemente, diversos órgãos públicos vêm declarando a adoção destas recomendações e divulgando bases de dados com a filosofia de dados abertos. Alguns dos mais volumosos e conhecidos podem ser citados, como a prefeitura de Nova Iorque (EUA)⁷, a prefeitura de São Francisco (EUA)⁸ e a prefeitura de Londres (GB)⁹. Além destes exemplos, podem ser citados os contextos das prefeituras nacionais de Novo Hamburgo (RS)¹⁰ e Porto Alegre (RS)¹¹, bem como o portal de dados abertos do governo federal (BR)¹². Entre outras possibilidades, estes conjuntos de bases de dados abertos e ligados são importantes por proporcionarem que sistemas acessem estas informações de modo mais efetivo, pelo fato de estarem estruturadas. Deste modo, podem representar um elemento importante para sistemas de pergunta e resposta ou sistemas de interação multimodal que utilizem estas bases para consulta.

2.6.3 Processamento distribuído e tratamento de mensagens

Os sistemas de interação multimodal devem ser considerados como sistemas que atuam em um contexto de processamento distribuído, no caso de se desejar manter flexibilidade e escalabilidade. Os principais fatores que corroboram esse argumento estão ligados aos diferentes dispositivos de entrada e saída a serem considerados, bem como à sua diversidade e o surgimento de novas possibilidades. Além disso, algumas atividades podem demandar processamento com recursos específicos, tais como algumas rotinas de aplicações de processamento de linguagem natural, reconhecimento de padrões em imagens, síntese de voz, entre outras. Esses processamentos podem ser distribuídos entre diversos dispositivos para otimizar o resultado.

A escalabilidade é definida como a capacidade do sistema de expandir de acordo com a demanda. Neste trabalho foi estudada a utilização de um *message oriented middleware (MOM)* para permitir essa escalabilidade. O MOM permite a troca de mensagens, de forma assíncrona, entre os componentes do modelo, através do protocolo *Advanced Message Queuing Protocol (AMQP)*. Esse protocolo define um padrão de comunicação aberto permitindo que diversas

⁶<http://www.wikipedia.org/>

⁷<https://nycopendata.socrata.com>

⁸<https://data.sfgov.org>

⁹<http://data.london.gov.uk>

¹⁰<https://dados.novohamburgo.rs.gov.br>

¹¹<http://www.datapoa.com.br>

¹²<http://dados.gov.br>

tecnologias e plataformas distintas sejam integradas (VINOSKI, 2006).

O funcionamento deste *middleware* pode ser resumido da seguinte forma: cada mensagem é armazenada em filas e permanecem lá até que a aplicação destino esteja pronta para recebê-las. Para evitar que as ligações entre as partes se restrinjam entre componentes específicos, são utilizados componentes conhecidos como *Message Brokers*. Estes são serviços que permitem o armazenamento, roteamento, tratamento e persistência das mensagens. Garantir a entrega da mensagem é uma característica muito importante nos MOMs. A mensagem só é removida da fila após receber uma confirmação do cliente. No meio do processo, se o sistema cliente sofrer algum problema, a mensagem permanece na fila. Esse sistema cliente também pode recusar uma mensagem. A flexibilidade dos sistemas de troca de mensagens permite que uma aplicação seja destinatário e remente ao mesmo tempo.

O AMQP é um protocolo de padrão aberto originado em meados de 2003 que foi crescendo com a ajuda e participações de empresas que trabalhavam com tecnologias proprietárias e outros padrões. Atualmente, o *AMQP Working Group* é responsável pelas especificações do protocolo. Sendo um protocolo assíncrono, binário e com múltiplos canais de comunicação, ele permite transações complexas, dando a opção de interoperabilidade entre as aplicações. O RabbitMQ, serviço utilizado nesse trabalho, é um *message broker* de código aberto, escrito na linguagem de programação Erlang, que implementa o protocolo AMQP. O serviço possui suporte à alta disponibilidade e tem bibliotecas clientes para as principais linguagens de programação (CARMO, 2012).

2.6.4 Linguagem de programação Python e NLTK

Criada por Guido van Rossum em 1991, Python é uma linguagem de programação interpretada de alto nível. Possui como característica a sua velocidade de desenvolvimento e legibilidade. Atualmente, a linguagem possui seu código aberto e é mantida pela *Python Software Foundation*, uma instituição sem fins lucrativos. É uma linguagem multiparadigma, misturando principalmente aspectos de orientação a objetos com programação funcional.

O *Natural Language Toolkit (NLTK)* é um conjunto de programas, bibliotecas e utilitários, escrita em Python, para o processamento de linguagem natural. O NLTK disponibiliza funcionalidades essenciais para o PLN, fornecendo *sentence delimiters*, *tokenizers*, *stemmers*, *taggers*, *stopwords* e *corpus* de muitos idiomas. O objetivo do NLTK é prover a facilidade de uso em aplicações que necessitem PLN.

No NLTK, *sentence delimiters* procuram definir o escopo de cada sentença através de padrões utilizados em cada idioma, como pontuação de final de frase, letras maiúsculas e demais características possíveis. Os *tokenizers* são responsáveis por dividir o texto em *tokens*, onde cada *token* representa uma palavra. Basicamente, eles são responsáveis por localizar uma sequência de caracteres, separadas por um delimitador, geralmente um espaço ou um ponto. Os *stemmers* fazem a análise morfológica das palavras, associando-as com suas variantes, de

Figura 8: AIML: Regra para saudação.

```

<aiml>
  <category>
    <pattern>OI</pattern>
    <template>OI HUMANO!</template>
  </category>
</aiml>

```

Fonte: Elaborada pelo autor

acordo com seu radical. Já os *taggers* baseiam-se nos *tokenizers* e *sentence delimiters* a fim de achar a função de cada palavra na frase. Nesse ponto cada um dos *tokens* da sentença são classificados como substantivos, adjetivos, verbos etc. *Stopwords* são listas de palavras que consideradas irrelevantes em cada idioma. Essas palavras geralmente aparecem repetidas vezes e prejudicam o processamento da sentença. *Corpora* é uma grande coleção de documentos (*corpus*) previamente processados, utilizado como base para análise estatística das sentenças. O NLTK possui *corpus* para diversos idiomas. Para o português, possui o MacMorpho (*corpus* de artigos publicados no jornal da Folha de São Paulo em 1994, com mais de 1 milhão de palavras classificadas), o *Floresta Portuguese Treebank* (*corpus* baseados em notícias com mais de 100 mil palavras classificadas) e o *Machado Text Corpus* (*corpus* das obras de Machado de Assis)(LOPER; BIRD, 2002).

2.6.5 Processador pyAIML e Linguagem AIML

Artificial Intelligence Markup Language (AIML) é uma linguagem para a criação de *softwares* para interação em linguagem natural (BUSH et al., 2001). É baseada em XML com um conjunto específico de marcações. As principais marcações da linguagem AIML são:

- `<aiml>`: Define início e fim do documento AIML.
- `<category>`: Define uma unidade de conhecimento que representa uma possível interação com o usuário.
- `<pattern>`: Contém um padrão que deve ser confrontado com a entrada do usuário. Cada entrada do usuário será avaliada conforme os nodos `<pattern>`.
- `<template>`: Contém a resposta para a entrada do usuário.

O funcionamento do AIML consiste em localizar o padrão apropriado de maneira simplificada. A figura 8 mostra a estrutura básica de um arquivo AIML. Nessa imagem o padrão é “OI”, ou seja, assim que o usuário digitar “OI”, o sistema irá responder “OI HUMANO!”. Para deixar

Figura 9: AIML: Utilização de caracteres especiais.

```

<aiml>
  <category>
    <pattern>OI, EU SOU *</pattern>
    <template>OI <star index="1" /></template>
  </category>
</aiml>

```

Fonte: Elaborada pelo autor

o sistema mais rico, o AIML possui caracteres curingas, como “*” e “_”. Eles permitem que os termos capturados nesses padrões sejam reutilizados na resposta, como mostrado na figura 9. Nesse trabalho foi utilizado o AIML para a geração do diálogo. As regras são criadas de forma dinâmica após consultar a base de dados ligados. O pyAIML¹³ é um interpretador para o AIML desenvolvido em Python, com o objetivo de prover 100% de compatibilidade com a especificação AIML 1.01. A figura 10 mostra a utilização básica da biblioteca em um *script* python. Nessa imagem, o arquivo *std-startup.xml* contém as regras de conhecimento da linguagem.

Figura 10: AIML: Utilização em código Python.

```

import aiml
k = aiml.Kernel()
k.learn("std-startup.xml")
while True:
    print k.respond(raw_input("> "))

```

Fonte: Elaborada pelo autor. Código de exemplo da documentação do pyAIML.

¹³<http://pyaiml.sourceforge.net/>

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados trabalhos relacionados com o tema proposto nessa dissertação. Foram usadas duas abordagens para a definição dos trabalhos estudados. Em um primeiro momento, o estudo de trabalhos relacionados buscou identificar, com pesquisas de referências em revistas científicas qualificadas e em eventos da área, exemplos de implementações de sistemas de interação multimodal com o objetivo de auxiliar no estudo e na descrição dos seus aspectos gerais de arquitetura e dos detalhes de incorporação de recursos diversos por estes sistemas. No segundo momento, o aspecto estudado está relacionado com os exemplos de trabalhos versando mais especificamente sobre o tratamento do diálogo em sistemas de interação multimodal. A seguir são apresentados trabalhos relacionados que mostram a importância da padronização de componentes entre sistemas de interação multimodal, além de aspectos de gerenciamento de diálogo de sistemas de interação multimodal.

3.1 Mudra: A Unified Multimodal Interaction Framework

Segundo Hoste, Dumas e Signer (2011) as interfaces baseadas no conceito WIMP (do inglês, *window, icon, menu, pointing device*) estão progressivamente cedendo espaços para sistemas de interação multimodal. A combinação de múltiplas entradas de dados, conhecida como fusão multimodal de entradas, torna-se um desafio para o desenvolvimento de cada sistema. Primeiramente, o processamento das múltiplas entradas deve acontecer em tempo real e o reconhecimento e fusão deve ser realizado de maneira paralela. Em segundo lugar, as entradas são oriundas de dispositivos diferentes, como reconhecimento de fala e de toque. De acordo com o autor, somente algumas ferramentas possuem a habilidade de tratar dados de fontes distintas.

Os *frameworks* de sistemas de interação multimodal podem realizar a fusão de entrada de dados em diferentes níveis de abstração: em baixo nível (fluxo de dados diretamente do dispositivo que deu a entrada) ou em alto nível (soluções onde a intenção ou significado da entrada já foi identificado). De acordo com Hoste, Dumas e Signer (2011), existe uma lacuna nestas ferramentas, pois não oferecem suporte a uma fusão de dados nativa, não permitindo ao sistema de interação multimodal ter acesso a diferentes níveis de fluxo de informação.

O trabalho apresenta o MUDRA, um *framework* de interação multimodal que suporta um processo integrado desde da fusão de dados de baixo nível até uma fusão de dados de mais alto nível, onde a informação já foi tratada através de uma linguagem de regras declarativas. De acordo com o autor, a arquitetura inovadora do trabalho incentiva o uso de princípios de engenharia de software, como a modularização e a escalabilidade para suportar um conjunto crescente de tipos de entrada.

3.2 The SEMAINE API: towards a standards-based framework for building emotion-oriented systems

O número de sistemas interativos que lidam com emoções é crescente. Esses sistemas buscam identificar padrões e inferir emoções de diversas maneiras. De acordo com o trabalho de (SCHRÖDER, 2010), a análise da emoção, estudada na área de computação afetiva, utiliza informações obtidas do rosto do usuário, comportamento e voz, entre outros. Estes sistemas acabam apresentando características em comum, como por exemplo, o fato de que todos necessitam representar os estados emocionais para processá-los posteriormente. Os componentes do sistema precisam trocar informação em cada etapa e, dessa maneira, a reutilização de componente e a padronização da comunicação é cada vez mais necessária.

O autor ressalta que a utilização de padrões ajuda na reusabilidade e interoperabilidade, fazendo um comparativo da utilização de padrões em páginas *web* e documentos eletrônicos. Já formatos fechados comerciais garantem uma vantagem competitiva para a empresa que os criou. Devido às limitações de plataformas comerciais, cada vez mais comuns na área de computação afetiva, o trabalho apresenta a SEMAINE API, um *framework* modular de código aberto para o desenvolvimento de sistemas orientados à emoção, utilizando padrões de abertos de comunicação.

O objetivo do trabalho é a criação de uma API que permita a integração dos componentes do sistema de forma robusta e em tempo real, além da reutilização de componentes e da possibilidade de estender o sistema. O SEMAINE API utiliza um *middleware* de troca de mensagens, baseado no protocolo aberto AMQP, para trocar informações entre os componentes do sistema, bem como padrões de representação de dados recomendados pelo W3C para a criação e desenvolvimentos de plataformas multimodais.

Para demonstrar a facilidade de uso na construção de sistemas orientados a emoções e validar a utilização da SEMAINE API, o autor criou três pequenos projetos. O primeiro projeto, chamado “*Hello World*”, consiste em um sistema baseado em entradas de texto que utiliza a API para inferir o estado afetivo do usuário após o texto digitado. Segundo o autor, apesar de simples, o projeto apresenta características de um sistema orientado a emoções. Após a análise da entrada, o sistema foi capaz de extrair dados de valência e alerta, retornando a emoção mais apropriada identificada.

O segundo projeto, chamado de “*Emotion Mirror*”, é uma variação do primeiro projeto. Nesse caso, as emoções não são deduzidas a partir de palavras de entrada, e sim através do reconhecimento do tom de voz, baseado no projeto *openSMILE* (EYBEN; WÖLLMER; SCHULLER, 2010). Para finalizar, o autor apresenta o terceiro projeto, um jogo que utiliza discurso emocional¹, onde o personagem é um nadador. O objetivo é evitar que o nadador afunde através de palavras positivas. Como resultados, a pesquisa mostra que a SEMAINE API é um *fra-*

¹O discurso emocional tem como característica transmitir sua mensagem através da emoção mais do que o significado presente no texto.

mework para a criação de sistemas simples e complexos, apresentando escalabilidade e reuso de componentes.

3.3 MATCH: An Architecture for Multimodal Dialogue Systems

Os autores apresentam a aplicação *Multimodal Access To City Help* (MATCH) e uma arquitetura geral para interação multimodal destinada a aplicações móveis (JOHNSTON; BANGALORE; VASIREDDY, 2001). Interfaces móveis devem permitir que o sistema se adapte às preferências do usuário e seu modo de interagir. Celulares, PDAs, *tablets* e *smartphones* possuem limitações quando falamos sobre entrada de dados, devido a pequenos tamanhos de telas e não possuírem um teclado físico ou mouse. Através de uma interação multimodal é possível resolver esse problema, combinando interações por voz com entradas de toques. A combinação de entradas deve acontecer de maneira mais apropriada a fim de ajudar a interação do usuário. O MATCH é um guia da cidade e um sistema de navegação que permite aos usuários acessar desde de restaurantes a informações de metrô na cidade de Nova Iorque. A interação se dá através de toque, marcando pontos do mapa, através de fala ou combinando ambos.

A arquitetura do MATCH utiliza um componente chamado MCUBE para enviar mensagens para um grupo de agentes via TCP/IP. Esses agentes são responsáveis por tarefas específicas como reconhecimento de fala, servidor de ASR, perfil do usuário, banco de dados, entre outros.

O processo de desenvolvimento foi realizado junto com pequenos testes, sendo possível verificar problemas de maneira mais eficaz. De acordo com os autores, o MATCH permite o desenvolvimento rápido de aplicações multimodais. Adicionalmente, a mesma arquitetura foi utilizada em duas outras aplicações: em uma interface multimodal de informações de diretórios corporativos e em uma aplicação médica para auxiliar médicos em salas de emergência. Esse aplicativos demonstraram a utilidade da arquitetura para rápida prototipação (JOHNSTON; BANGALORE; VASIREDDY, 2001).

3.4 Context Based Multimodal Fusion

Pfleger (2004) apresenta uma abordagem chamada *context based multimodal integration*, onde cada interação entre usuário e computador é enriquecida com informações do contexto em que ele está. Segundo o autor, as informações de contexto são dados de todas as entradas de dispositivos disponíveis no momento da interação. O modelo possui um componente capaz de processar as múltiplas entradas e fazer a fusão das informações. O resultado desse processamento procura tentar informar ao sistema a intenção do usuário, expressada através de regras de sequência de ações.

Para o tratamento do diálogo, o modelo utiliza um sistema de regras de produção desenvolvido para atuar na integração com sistemas multimodais, apelidado de *PATE (a Production rule system based on Activation and Typed feature structure Elements)*. Assim como os demais sis-

temas de regra de produção, a abordagem proposta possui uma memória de trabalho que serve como uma central de informações. Cada regra de produção consiste em três componentes: peso, condição e ação. Para a validação, o modelo foi aplicado ao sistema chamado COMIC, onde o autor mostrou que a abordagem baseada em regras de produção atende às demandas do componente responsável pela integração entre as múltiplas entradas em um sistema multimodal (PFLEGER, 2004).

3.5 MIND: A context-based multimodal interpretation framework in conversational systems

Sistemas de interação conversacionais devem tratar entradas de dados muitas vezes com falhas, imprecisas ou sem a informação completa. No trabalho de (CHAI; PAN; ZHOU, 2005) o autor argumenta que a fusão multimodal de dados não é suficiente para garantir a interpretação correta da informação. O *MIND (Multimodal Interpreter for Natural Dialog)* é um *framework* para interpretação multimodal baseada em contexto. O trabalho utiliza um modelo semântico que dá o significado à entrada do usuário. O *framework* também utiliza informações de contexto, histórico e domínio para melhorar o entendimento das informações.

No tratamento da entrada do usuário, o MIND possui dois objetivos. O primeiro é entender o significado da entrada do usuário enviado pelo componente responsável pela conversação. O segundo é capturar diferentes tipos de entradas de dados, como por exemplo, conversação ou gestos, para responder o usuário de forma mais apropriada. Para isso, as entradas do usuário são classificadas entre intenção (objetivo da ação do usuário), atenção (conteúdo da entrada do usuário), preferências de apresentação (a melhor resposta para o usuário) e interpretação da situação, com uma avaliação geral sobre o entendimento das entradas do usuário.

O MIND possui também um modelo capaz de prover uma visão geral do estado da conversação e outro modelo capaz de obter informações do contexto do usuário. O trabalho utiliza o *IBM ViaVoice* para realizar o reconhecimento de discurso e um componente para processamento de linguagem natural baseado em estatísticas. Para o reconhecimento de gestos, o grupo de pesquisa desse trabalho desenvolveu seu próprio componente. Já a análise do contexto é dividida em três partes: contexto conversacional, contexto do domínio e contexto visual.

O contexto conversacional provê ao sistema um histórico da conversação, permitindo que o usuário não precise repetir dados mencionados na conversa anterior. O contexto de domínio fornece ao sistema informações extras sobre o domínio do assunto, ajudando a resolver ambiguidades. Por último, o contexto visual permite que o usuário se refira a objetos representados na tela, através de informações espaciais e atributos do objeto a que ele se refere. Como resultado, o autor conclui que o MIND é capaz de processar uma grande variedade de entradas do usuário, incluindo entradas incompletas, ambíguas ou complexas (CHAI; PAN; ZHOU, 2005).

3.6 Multimodal human-robot interaction with Chatterbot system: Extending AIML towards supporting embodied interactions

Uma das principais metas na interação homem-robô é o desenvolvimento da habilidade de entender expressões humanas, de forma verbal ou não verbal. Para atingir esse objetivo é necessária a integração de algumas modalidades de comunicação e entendimento. O trabalho apresentado por Tan, Duan e Inamura (2012) propõe adaptar a interação baseada em textos, utilizada em *chatterbots*², para o desenvolvimento de uma interação multimodal homem-robô.

O trabalho utiliza o AIML para a geração do diálogo, porém os autores argumentam que a linguagem é limitada a interações verbais. De acordo com Tan, Duan e Inamura (2012), existe uma lacuna para atender as demandas da interação homem-robô, uma vez que as interações não verbais são tão importantes quanto as interações verbais nessa área. O objetivo deste trabalho é realizar a interação multimodal homem-robô em com base em um sistema de diálogo simples. O trabalho apresenta uma maneira de utilizar o AIML para representar entradas de gestos, utilizando o elemento *<set>*. As expressões e as emoções do usuário são representadas utilizando o elemento *<topic>*. O aprendizado do robô é realizado utilizando o elemento *<learn>*. De acordo com os autores, os estudos de comparação com sistemas convencionais de diálogos complexos serão realizados no futuro para verificar a usabilidade e a eficácia do desenvolvimento. Em testes preliminares, a proposta se mostrou funcional.

3.7 Florence: a Dialogue Manager Framework for Spoken Dialogue Systems

Os avanços nas tecnologias de sistemas de diálogos falados vêm crescendo. Esses sistemas permitem que os usuários realizem interações de voz com máquinas em vez de serem obrigados a navegar nos menus de opções dos sistemas. O *VoiceXML*³ surgiu como um avanço para a comunicação entre sistemas de baseados em voz, pois permite uma forma padronizada de comunicação entre as pontas, com seu padrão aberto.

Atualmente, os sistemas de diálogo baseados em voz, que utilizam linguagem natural, realizam interações mais complexas com seus usuários, utilizando modelos computacionais mais complexos. Nesse ponto, o *VoiceXML* é limitado, não permitindo uma abordagem mais geral. O trabalho apresenta o Florence, um gerenciador de diálogo para abordagens mais gerais de forma flexível, permitindo a interoperabilidade entre estratégias de diálogo. A linguagem XML descrita no Florence (*FXML*) facilita o desenvolvimento de aplicações em linguagem natural e permite que o autor do diálogo possa reutilizar recursos entre diversas aplicações. O Florence também possui características para atender processamento em larga escala, de maneira distribuída.

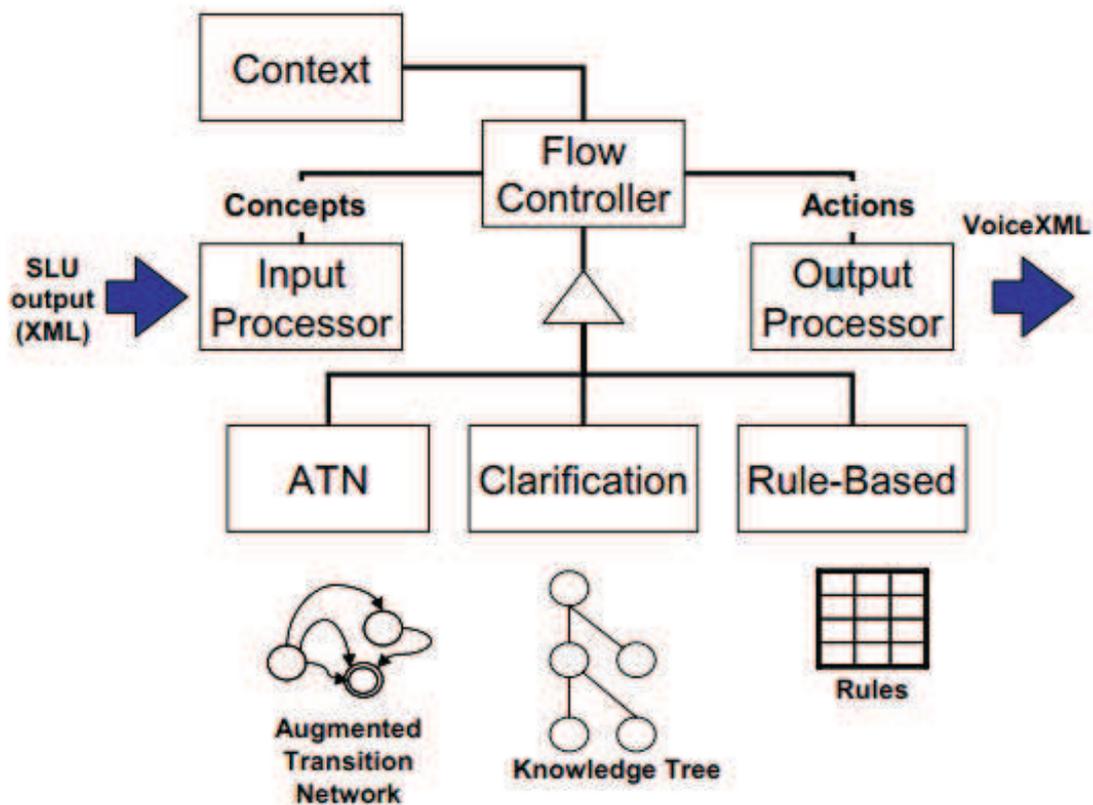
O conceito chave da arquitetura do Florence é a *dialogue stack*. Esta é uma estrutura de

²Chatterbot é um programa de computador que tenta realizar um diálogo com as pessoas.

³<http://www.w3.org/TR/2004/REC-voicexml20-20040316/>

dados que controla a ordem em que os diálogos foram realizados e permite um controle sobre esses diálogos. O modelo possui os *flow controllers*, que realizam o rastreamento sobre um diálogo na pilha. Os contextos local e global são usados para manter informações variáveis que podem ser enviados de um controlador para o outro durante um diálogo. O sistema pode gerar ações, indicando para se comunicar com o usuário, através do *VoiceXML*. A figura 11 mostra o fluxograma realizado pelo Florence para controle do diálogo.

Figura 11: Fluxograma geral da arquitetura do Florence.



Fonte: Florence: a Dialogue Manager Framework for Spoken Dialogue Systems (DI FABBRIZIO; LEWIS, 2004)

Segundo Di Fabbrizio e Lewis (2004), o Florence se mostrou um *framework* para o gerenciamento de diálogo flexível que utiliza boas práticas descritas em trabalhos anteriores e permite a interoperabilidade com sistemas utilizados pela indústria. Devido ao suporte a novos algoritmos, o Florence torna-se uma opção para o desenvolvimento de sistemas de linguagem natural faladas em larga escala.

3.8 Comparativo e considerações

Nessa seção são analisados os detalhes dos trabalhos apresentados. A tabela 1 traz uma comparação entre as principais características dos trabalhos relacionados.

No trabalho de Hoste, Dumas e Signer (2011), os autores apresentam uma abordagem nova para a fusão de entradas multimodais, permitindo a utilização dessas informações em vários

níveis, além de uma modelagem que possibilita a modularização do sistema. Essa característica é fundamental para um modelo expansível, assim como proposto na presente pesquisa.

A proposta de Schröder (2010) apresenta um trabalho modular que permite a interação através de mensagens no padrão XML. Uma das vantagens desse modelo é a utilização de padrões recomendados pelo W3C, permitindo assim interoperabilidade com outros sistemas e componentes. O SEMAINE API utiliza um *message-oriented middleware* para enviar mensagens entre seus componentes permitindo desacoplamento e fácil reuso.

Johnston, Bangalore e Vasireddy (2001) apresentam uma abordagem que possibilita o desenvolvimento modular e rápido de aplicações multimodais com suporte às informações do contexto. Apesar de suportar perfis de usuários, o MATCH não mantém um histórico de diálogo para interações futuras.

No trabalho de Pflieger (2004), a memória de trabalho é uma representação controlada dos objetos que descrevem a situação do contexto através de regras. A aplicação de cada regra é feita através da validação contra as regras de produção. Para cada regra é verificado se as condições satisfazem a configuração atual da memória de trabalho.

Tabela 1: Trabalhos relacionados

	Mudra	SEMAINE API	MATCH	PATE	MIND
Domínio	Modelo genérico	Sistemas orientados a emoções	Dispositivos móveis	Modelo genérico	Modelo genérico
Uso de componentes	Sim	Sim	Sim	Não	Sim
Usa alguma info de contexto (localização, sensores etc.)	Não	Sim	Não	Não	Sim
Uso de memória de trabalho	Não	Não	Não	Não	Não
Arquitetura escalável	Sim	Sim	Sim	Não	Não

Fonte: Elaborado pelo autor.

Já no trabalho de Chai, Pan e Zhou (2005), o sistema explora várias abordagens para entender a entrada do usuário, entre elas, um histórico de conversação. O MIND possui como característica utilizar contextos variados (diálogo e domínio, por exemplo) para melhorar a fusão multimodal de entrada. Essas funcionalidades permitem uma melhor interação com o usuário. Os componentes do modelo sugerido em MIND são bastante acoplados, o que prejudica a expansão por terceiros.

A tabela 1 traz uma comparação entre características estudadas dos trabalhos relacionados. Estas características são: a) o domínio possível para utilização dos sistemas; b) a utilização de componentes de software e possibilidades de modularização; c) características dos modelos

quanto ao uso de informação de contexto para a sua operação; d) utilização de memória de trabalho; e) possibilidades do sistema quanto à sua escalabilidade. Portanto nesta tabela são apresentadas de forma resumida as propriedades que são relevantes nessa dissertação e que estimularam a definição do modelo proposto.

Os trabalhos de Di Fabbrizio e Lewis (2004) e Tan, Duan e Inamura (2012) não tratam diretamente de sistemas de interação multimodal (e por isso não foram adicionados a tabela 1), porém foram relacionados aqui devido a algumas características importantes. O trabalho de Tan, Duan e Inamura (2012) destaca-se pela maneira que o AIML foi utilizado para expressar emoções, gestos e outras ações não verbais. Essa abordagem foi utilizada no protótipo dessa dissertação. Já o trabalho de Di Fabbrizio e Lewis (2004) mostra a utiliza um gerenciador de diálogo com aspectos importantes para tratar um diálogo em linguagem natural.

Os trabalhos apresentados corroboram com a ideia de Oviatt, Coulston e Lunsford (2004). É notável o avanço da percepção multissensorial em interfaces multimodais e, ao mesmo tempo, existe uma lacuna na relação da utilização das informações da entrada do usuário com uma memória de trabalho. Esta dissertação apresenta um modelo para a memória de trabalho, capaz de utilizar algumas características do modelo de Baddeley e Hitch (1974), acrescido de informações do contexto do usuário, obtidas por sensores, a fim de melhorar a experiência do usuário em sistemas de interação multimodal. A utilização de protocolos abertos e o desacoplamento de componentes possibilitam que o modelo seja aplicado em sistemas multimodais já existentes e permitem a expansão para novos tipos de entradas de dados.

A principal contribuição deste trabalho, portanto, consiste na definição de um modelo de gerenciamento de diálogo com base em um modelo cognitivo (BADDELEY; HITCH, 1974), o que não é contemplado nos trabalhos estudados. De forma complementar, deve ser destacado que o modelo proposto foi definido de modo a permitir a inclusão de novos componentes para possibilitar a ampliação gradual dos fatores tratados pelo sistema de interação multimodal no futuro, de acordo com a expectativa de surgimento e disponibilização de novos recursos de interação.

4 MODELO PROPOSTO

Neste capítulo são apresentadas as escolhas que compõem o modelo proposto no presente trabalho. Com a finalidade de possibilitar a implementação desse modelo em um protótipo que seja funcional e que atenda às necessidades de validação do modelo proposto, existem dois níveis de descrição relevantes. O primeiro está descrito a partir da arquitetura geral para o sistema de interação multimodal, envolvendo todos os elementos necessários para a sua construção (itens 4.1, 4.2 e 4.3). Já o segundo nível de descrição está relacionado com o gerenciamento de diálogo, sendo descrito em maiores detalhes no item 4.4.

O modelo proposto é dividido em 3 macro etapas. A figura 12 apresenta uma visão geral da arquitetura do modelo e os elementos que compõem a entrada, o controle e a saída. Para a descrição deste modelo foi adotada a notação TAM (*Technical Architecture Modeling*)¹. O modelo proposto visa um diálogo mais robusto humano-computador em sistemas de interação multimodal, envolvendo também a ampliação dos recursos de entrada de dados a partir da possibilidade de futura expansão dos tipos de informação de entrada, através de módulos que possam interpretar dados descritos através do formato EMMA (Johnston et al. (2009)).

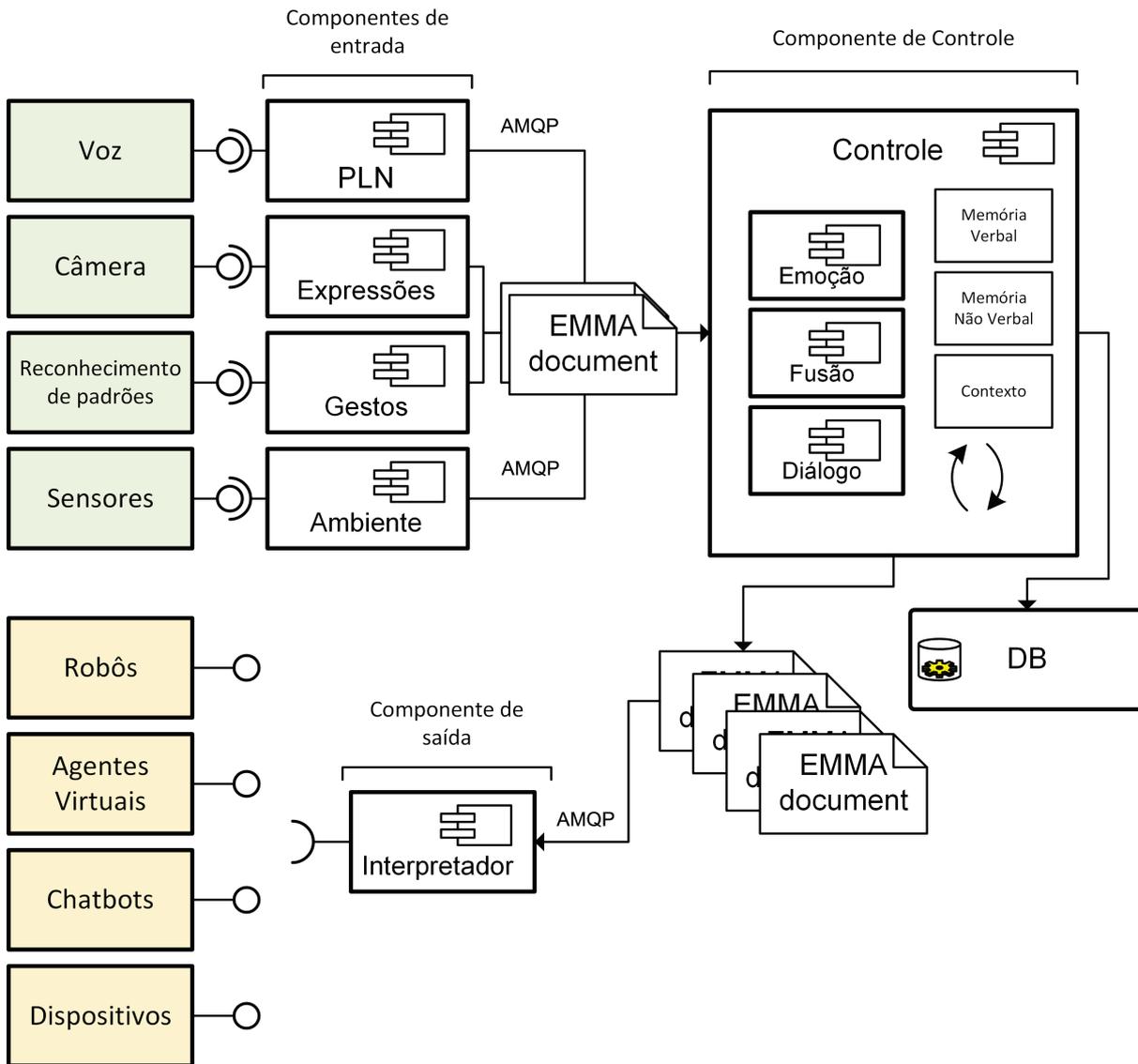
4.1 Entrada

Sistemas de interação multimodal têm a capacidade de combinar múltiplas entradas de dados, provenientes de fontes e tipos distintos. As diversas modalidades de entradas de dados se complementam, evitando ambiguidade e aumentando a acurácia a fim de buscar um significado da intenção do usuário com maior precisão.

A figura 12 mostra os elementos envolvidos na entrada do modelo. Assim como descrito na especificação do modelo proposto pelo W3C (BARNETT et al., 2012), as entradas de dados podem ser originadas em fontes tais como teclados, sistemas de reconhecimento de fala, câmeras, dispositivos de reconhecimento de padrões, sensores, entre outros. O *hardware*, o *software*, os métodos utilizados e as arquiteturas envolvidas em cada uma dessas entradas acabam sendo muito específicos para cada situação devido à complexidade de cada tarefa. Cada um desses métodos de entradas precisa ser analisado, tratado e interpretado para que sua informação seja útil ao sistema de interação multimodal.

Para atender esta demanda, cada um dos tipos de entrada possui um interpretador, elemento necessário para tratar dispositivos diversos. Podemos dizer que um interpretador de imagens originadas de uma câmera é capaz de inferir se o usuário está triste ou feliz, ou até mesmo saber se o usuário saiu e outro entrou no sistema. O mesmo acontece com a fala: é necessário um interpretador para que a entrada de voz do usuário transforme-se em uma entrada de texto. Esse componente ainda pode variar devido às diferenças de idiomas, linguagens, culturas etc. Assim que estes dados são interpretados começa a etapa de processamento. Nesse ponto, é criada uma

¹<http://www.fmc-modeling.org/fmc-and-tam>

Figura 12: Arquitetura do modelo proposto.

Fonte: Elaborada pelo autor.

mensagem para enviar ao modelo, de acordo com cada caso. Essa mensagem é enviada através do protocolo AMQP (*Advanced Message Queuing Protocol*) de forma assíncrona (ou síncrona, dependendo exclusivamente do suporte do *software* que está utilizando o modelo).

O EMMA (*Extensible MultiModal Annotation markup language*) define atributos de marca-temporal para dizer quando uma ação começou e terminou. Isso permite saber se as entradas originadas de diversos dispositivos foram efetuadas no mesmo momento. As tecnologias EMMA e AMQP são descritas no capítulo 2. Conforme Johnston et al. (2009), cada um dos dispositivos deve enviar o XML de acordo com seu contexto, por exemplo, o componente responsável por ler informações do GPS irá enviar o atributo identificando-se como “sensor”. As principais formas e interpretações de entrada são:

- **Informação de contexto/sensores:** Durante um diálogo entre seres humanos, não é pre-

ciso informar um ao outro qual a temperatura atual para depois sentir frio ou calor. Informações do nosso dia-a-dia, como horário, localização e temperatura estão implícitas durante uma conversação e em um sistema multimodal esses dados devem ser captados através de sensores. A informação de localização, por exemplo, permite ao usuário manter uma interação mais rica sobre sua região geográfica. Já informações de temperatura permitem ao usuário ser avisado sobre possíveis mudanças de tempo, caso este aspecto seja relevante ou relacionado com o diálogo em andamento.

- **Texto:** Nesse modelo, a principal entrada para o sistema é realizada através de texto. Deve-se deixar claro que esse tipo de entrada pode vir de fontes como um teclado de computador, um dispositivo de reconhecimento de som ou fala, ou qualquer outro meio que possa transformar as informações em uma cadeia de caracteres.
- **Expressão:** Detectar e identificar rostos de seres humanos em determinada imagem não é uma tarefa complexa, diferentemente da detecção de certos padrões, que requerem uma análise mais detalhada. Por outro lado, o reconhecimento de expressões é uma tarefa computacionalmente difícil. A identificação de uma expressão é um ponto relevante durante uma conversa, pois ao mesmo tempo em que um ser humano diz "sim", ele pode demonstrar uma expressão convicta ou duvidosa. Esse interpretador tem como objetivo recolher esses dados, identificar padrões de emoção e entregá-los ao modelo através do formato especificado no padrão EMMA. Assim é possível inferir se o usuário necessita ou não de mais informações, se está feliz ou triste durante a conversa etc.
- **Gestos e Sinais:** Uma gama de técnicas de análise e processamento de imagem, com o objetivo de fazer com que o computador entenda o gesto é estudada na área de reconhecimento de gestos e padrões. Na maioria dos casos estas imagens são capturadas por uma câmera ou dispositivo específico. Para um diálogo, o simples gesto de balançar a cabeça de cima para baixo, pode ser deduzido como se o usuário estivesse entendendo o que se trata ou concordando. Já o balançar da cabeça de um lado para outro pode dar uma informação ao modelo de que não está entendendo ou discordando. É importante citar que o reconhecimento de sinais pode substituir a entrada falada ou escrita utilizando linguagem de sinais.

Cada uma dessas informações é tratada em etapas de aquisição, interpretação, análise e processamento. Depois destas etapas, as informações são enviadas ao componente de controle descrito no modelo geral. Essas entradas são enviadas no formato EMMA.

4.2 Controle

De modo semelhante ao exibido na arquitetura geral de Dumas, Lalanne e Oviatt (2009), que mostra que sistemas de interação multimodal possuem um componente responsável pela

integração das informações, no modelo proposto o componente responsável pela integração - o componente de controle - é dividido em um módulo de fusão de entradas multimodais, um componente de acesso a dados ligados, o gerenciador de diálogo, a memória verbal, a memória não verbal e as informações de ambiente e contexto.

Como diferencial em relação ao trabalho de Dumas, Lalanne e Oviatt (2009), destaca-se: o gerenciador de diálogo, responsável pela interpretação das entradas e o controle do tópico em atenção; os elementos de memória verbal e memória não verbal, que armazenam um histórico sobre as interações anteriores; os componentes de fusão e acesso a dados ligados. Estes aspectos são representados no componente de controle da figura 12. Indo ao encontro do modelo cognitivo proposto por Baddeley e Hitch (1974), é possível simular a atenção e acessar dados armazenados em uma memória operacional, buscando a geração de um diálogo mais natural.

Para que as múltiplas entradas sejam processadas e efetivamente sejam úteis, é necessária a integração da informação, chamada de fusão de entradas multimodais. A literatura traz duas principais estratégias para essa fusão (CORRADINI et al., 2005). A primeira, chamada de *feature level fusion*, é utilizada geralmente com entradas sincronizadas onde as informações dependam uma das outras. A segunda estratégia é utilizada nesse trabalho, conhecida como *semantic level fusion*, é a mais aplicada quando as entradas do sistema multimodal possuem características diferentes, principalmente em tempo e tamanho das entradas. Cada entrada é marcada com uma marca-temporal, assim é possível definir quais informações enviadas pelos componentes de entrada pertencem ao mesmo contexto temporal. Mas também é possível utilizar outros critérios para a identificação de elementos de um contexto. A abordagem *semantic level fusion* oferece algumas vantagens em relação ao *feature level fusion*, como escalabilidade e modularidade (CORRADINI et al., 2005). A fusão é realizada quando as marcas-temporais de cada entrada respeitam um limite de tempo de interação. Por exemplo, se duas mensagens chegam em um intervalo de dois segundos, uma deve considerar a outra para a interpretação e fusão.

O gerenciador de diálogo é responsável pela interpretação após a fusão multimodal. No momento em que o usuário trocar de assunto, é realizada uma consulta a uma base de conhecimento. Essa base serve para o modelo como uma memória de longo prazo, onde o sistema irá resgatar dados sobre determinado assunto. Neste trabalho foi utilizada a DBPedia, já que esta possui um grande acervo de informações. Outras bases de dados ligados mais específicas poderiam ser utilizadas, diminuindo o escopo e aumentando a acurácia.

O entendimento do diálogo acontece durante a recuperação da informação. Através de uso de recursos de Processamento de Linguagem Natural, o sistema busca identificar se o usuário está ainda falando do mesmo assunto da interação anterior. Se o usuário trocar de assunto (por exemplo, se o usuário mencionar outro tópico), será realizada uma consulta na DBpedia. A memória verbal ajuda na identificação do tópico, enquanto a memória não verbal informa ao sistema o estado emocional do usuário. As informações do ambiente e contexto ajudam a responder perguntas mais específicas, como por exemplo, “Será que chove hoje?”. Após a

consulta, algumas regras AIML são criadas e o sistema irá aprender sobre elas, com base no atributo *rdf:type*.

As mensagens enviadas ao componente de controle podem alterar os estados da conversação, informando ao sistema alterações de localização, emoções ou até mesmo identificando que um novo usuário está presente. Quando isso acontece, o modelo não faz o processamento de linguagem natural, mas utiliza as informações para compor novos padrões de interação, como previsão de tempo, mensagens sobre o humor do usuário, conversação sobre o contexto, entre outros cenários possíveis. O resultado das interpretações, após o seu processamento, é enviado via protocolo AMQP em formato EMMA para os componentes de saída do modelo.

4.3 Saída

A saída acontece após a interpretação necessária realizada no componente de controle. Em sistemas de interação multimodal, a aplicação cliente dessa interpretação pode ter diversos fins. Conforme a figura 12, o sistema poderia interpretar os dados e enviar para um agente virtual ou um braço robótico. O componente de controle envia para um interpretador de saída a mensagem no padrão EMMA. Esse componente é responsável por analisar o conteúdo e enviar à aplicação cliente uma resposta adequada. Isso permite uma expansão para diversos cenários, uma vez que o AMQP e o EMMA são padrões abertos.

4.4 Gerenciamento do diálogo

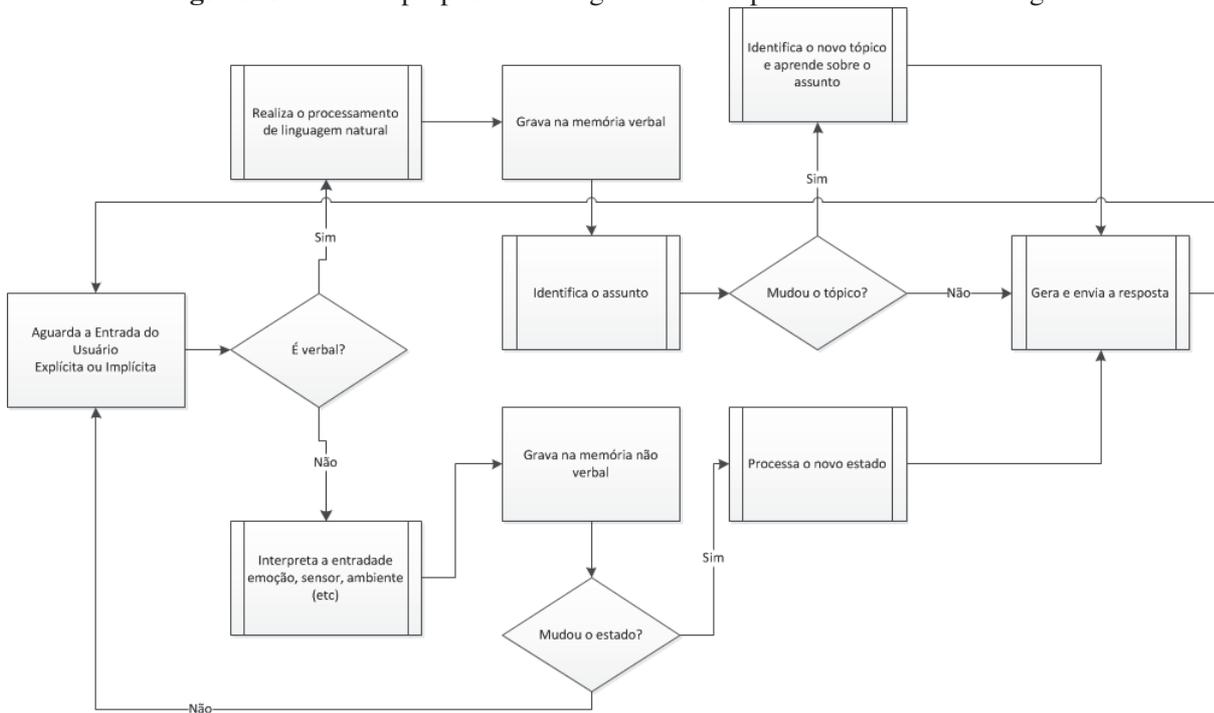
O fluxograma apresentado na figura 13 mostra a sequência do funcionamento que foi adotada no modelo para alcançar o objetivo de promover um melhor gerenciamento do diálogo. A seguir, são mostrados mais detalhes sobre cada item.

Dada uma determinada entrada através de um dos componentes do sistema, ela será interpretada. Cada componente envia mensagens de acordo com a sua função. Para exemplificar, um componente responsável pelo *speech-to-text*² irá enviar essas informações como texto. Já um componente responsável pela emoção pode obter imagens de uma câmera, inferir o estado emocional do usuário e enviar essa informação ao sistema. Os componentes se comunicam através de mensagens assíncronas, ou seja, mensagens que podem ser processadas sem bloquear demais funcionalidades do sistema. Ao receber a mensagem, o modelo faz sua primeira análise com base em sua origem, classificada como entrada explícita ou implícita.

As interações explícitas são verbais ou gestuais (expressando um sentido em linguagem de sinais, por exemplo). Já interações implícitas são informações lidas de sensores, câmeras etc. Se esta interação é verbal, o sistema realiza o processamento de linguagem natural (no fluxograma como “identificação do assunto”) com o objetivo de identificar a que assunto o usuário está

²*Speech-to-text* é uma funcionalidade dos softwares de reconhecimento de voz que permite ao usuário utilizar a própria voz ao invés de digitar mensagens de texto. Geralmente esse reconhecimento é utilizado para enviar comandos a um aplicativo ou falar o texto para um computador.

Figura 13: Modelo proposto: Fluxograma das etapas de controle do diálogo.



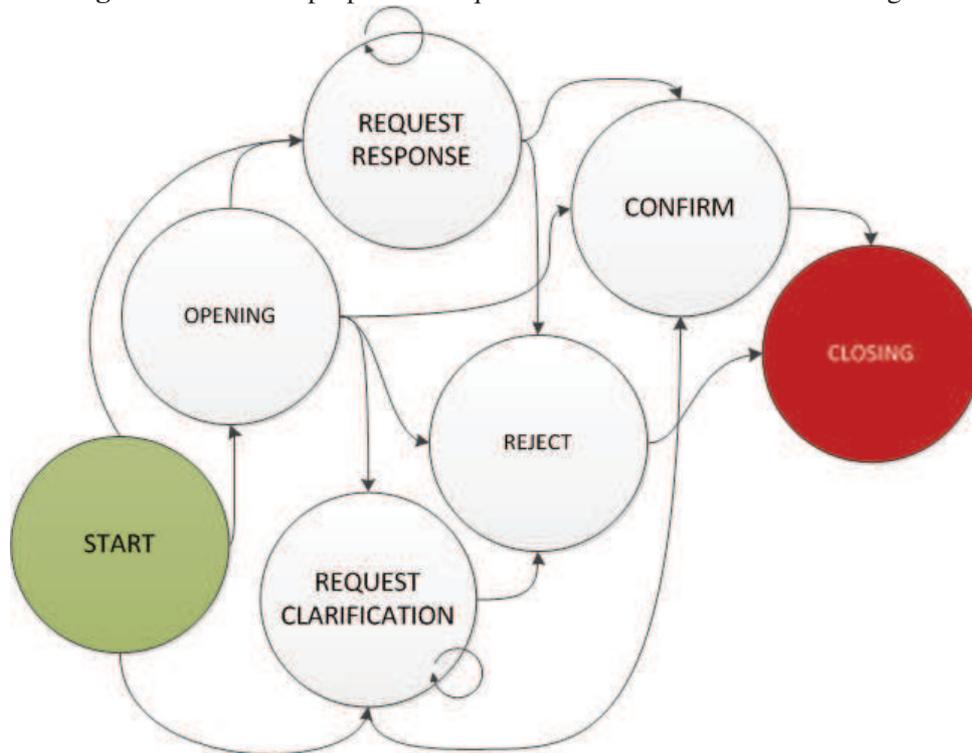
Fonte: Elaborada pelo autor.

se referindo e de que forma. Essa informação é gravada na memória verbal. Nesse ponto, o modelo identifica o tópico da questão através do processamento de linguagem natural. O tópico é importante para que o algoritmo defina o ponto de atenção sobre a conversação. Sempre que o ponto de atenção da conversa é alterado depois de uma nova entrada do usuário, o sistema deve consultar uma base de dados ligados para recuperar informações sobre esse assunto. Essa informação é gravada na memória e gera respostas para possíveis perguntas do usuário.

No caso de interações implícitas, o modelo utiliza a memória não verbal para armazená-las. Esses dados serão mantidos para identificar localização, estado emocional atual e demais atributos que podem influenciar no diálogo. Sempre que essas informações mudarem, o sistema pode ou não gerar uma nova resposta. O estado da conversação, bem como o tópico em atenção, pode ser alterado se a localização, emoção ou identificação do usuário mudar. O controle dos estados do diálogo é representado através de uma máquina de estado finito (*finite machine state - FMS*). A figura 14 mostra os possíveis estados da máquina durante um diálogo. O objetivo da máquina de estados é evitar que o sistema processe as entradas do usuário de maneira incorreta, quando elas não possuírem relação ao tópico em questão. O gerenciamento de diálogo também permite que perguntas como “*Qual a previsão do tempo para hoje?*”, possam ser respondidas utilizando dados da memória não verbal, como a localização obtida pelo GPS, caso seja esse o objetivo do sistema. Abaixo são descritos os possíveis estados da máquina e seus respectivos significados:

- **START:** estado inicial da máquina.

Figura 14: Modelo proposto: Máquina de estado de controle de diálogo.



Fonte: Elaborada pelo autor.

- **OPENING:** estado de entrada para o diálogo (ex: *Olá*).
- **REQUEST_RESPONSE:** o usuário faz uma pergunta direta (ex: *Quem escreveu Quincas Borba?*).
- **REQUEST_CLARIFICATION:** o usuário faz uma pergunta relacionada ao contexto das perguntas anteriores (ex: *E qual sua data de nascimento?*).
- **CONFIRM:** o usuário faz uma afirmação ou aceita a resposta (ex: *Sim, entendo*).
- **REJECT:** o usuário faz uma negação ou não aceita a resposta (ex: *Acho que não*).
- **CLOSING:** o usuário encerra o diálogo (ex: *Tchau*).

É importante destacar que a máquina de estados somente muda de estado após a análise da entrada do usuário, pois é fundamental o reconhecimento de alguns padrões de conversação na entrada do usuário para que o sistema tome as decisões corretas.

5 ASPECTOS DA IMPLEMENTAÇÃO DE PROTÓTIPOS

Para demonstrar e avaliar a viabilidade de integração dos itens citados no modelo proposto, dois protótipos foram criados. O primeiro trata-se de um programa de perguntas e respostas, baseado em interação através de mensagens em formato textual. Seu foco principal é voltado para a validação dos elementos envolvidos na troca de mensagens, permitindo o baixo acoplamento entre seus componentes. Esse protótipo também permite a validação do uso do padrão EMMA na representação dos dados, a utilização do processador AIML e das bibliotecas e serviços auxiliares de processamento de linguagem natural. O segundo protótipo trata-se de uma aplicação para *smartphones* de perguntas e respostas, no qual é possível observar aspectos relacionados com a interação multimodal através de emoções, sensores e reconhecimento de fala. Ambos os protótipos têm como objetivo destacar os itens envolvidos no gerenciamento do diálogo, na utilização de informações do ambiente e na utilização de base de dados ligados para responder as perguntas dos usuários.

5.1 Detalhes de implementação - Protótipo 1

Para permitir a integração de dados em diferentes formatos originados em dispositivos distintos, foi utilizado o padrão EMMA, descrito anteriormente, pois ele garante uma perspectiva de compatibilidade e interoperabilidade (JOHNSTON et al., 2009). A figura 15 exemplifica o uso deste padrão para representar dados obtidos a partir de uma entrada textual ou de um dispositivo de reconhecimento de fala. Nesta figura podem ser observados os trechos do arquivo em formato XML que descreve uma palavra “Olá” identificada a partir de um componente de reconhecimento de voz.

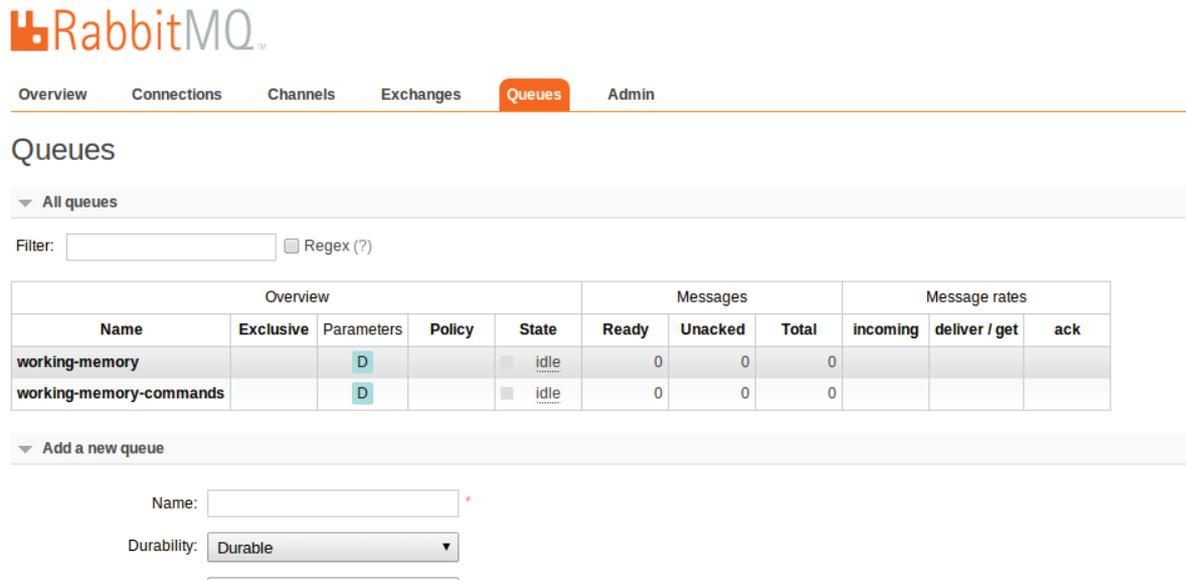
Figura 15: Representação da entrada no formato EMMA.

```
<emma:emma version="1.0"
  xmlns:emma="http://www.w3.org/2003/04/emma"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.w3.org/2003/04/emma
  http://www.w3.org/TR/2007/CR-emma-20071211/emma.xsd"
  xmlns="http://www.example.com/example">
  <emma:interpretation id="raw"
    emma:medium="acoustic" emma:mode="voice">
    <literal>Olá</literal>
  </emma:interpretation>
</emma:emma>
```

Fonte: Elaborada pelo autor.

A integração entre os dispositivos externos e os módulos do protótipo é realizada através de um *Message Oriented Middleware* (MOM), permitindo que exista uma completa indepen-

Figura 16: Tela de gerenciamento do RabbitMQ.



Overview

Connections Channels Exchanges **Queues** Admin

Queues

▼ All queues

Filter: Regex (?)

Overview					Messages			Message rates		
Name	Exclusive	Parameters	Policy	State	Ready	Unacked	Total	incoming	deliver / get	ack
working-memory		D		idle	0	0	0			
working-memory-commands		D		idle	0	0	0			

▼ Add a new queue

Name: *

Durability:

Fonte: Elaborada pelo autor.

dência entre os diferentes dispositivos e componentes do sistema. Nesse trabalho foi utilizado o *message broker* RabbitMQ. A figura 16 mostra a tela de administração deste serviço, destacando as filas utilizadas pelo modelo. A fila *working-memory* é responsável por armazenar as mensagens recebidas de componentes do modelo enquanto a fila *working-memory-commands* armazena as mensagens após processamento e interpretação. Essas mensagens são destinadas ao componente de saída e geração do diálogo. As configurações, os nomes das filas e serviços podem ser substituídos desde que utilizem protocolos similares.

Para a geração do diálogo utilizou-se um processador para AIML baseado na linguagem Python, a biblioteca PyAIML¹. Durante a interação com o usuário, o sistema realiza consultas à uma base de dados ligados. A DBPedia (AUER et al., 2007) foi utilizada nesse trabalho. Os seus resultados das consultas, retornados no formato JSON-LD², são manipulados e utilizados na geração dos contextos AIML. O processamento das entradas textuais é realizado através da biblioteca NLTK³, para a geração de informações morfossintáticas, e do serviço AlchemyAPI⁴, para realizar o processamento de linguagem natural e identificação do assunto e análise de sentimentos.

O ponto principal de interesse para a avaliação desse protótipo foi a busca de evidências do resultado positivo deste modelo para o tratamento do contexto de diálogo (seu ponto de atenção e tópico), para que ele seja realizado de forma mais natural, em situações especificamente destacadas como problemáticas em outros sistemas de interação multimodal. Assim, para a avaliação do protótipo desenvolvido, foram tratadas situações em que se verificada a necessi-

¹<http://pyaiml.sourceforge.net/>

²<http://json-ld.org/>

³<http://www.nltk.org/>

⁴<http://alchemyapi.com/>

dade de melhorias em resultados observados em outros sistemas de interação multimodal, tal como o sistema VOX (F. J. SERÓN C. BOBED, 2014). Neste sistema, considera-se que o gerenciamento do diálogo pode ser ampliado para uma forma que apresente maior qualidade de conversação. Foram escolhidas situações para validar a capacidade do modelo proposto para: a) identificar um contexto prévio referido, de modo que a percepção dos usuários seja adequada; b) identificar uma troca de contexto, evitando uma continuidade de diálogo de forma inadequada.

Figura 17: Trecho de um diálogo exemplo gerado pelo protótipo 1.

```

Usuário: Olá!
Sistema: Olá!
Usuário: O que você gosta de comer?
Sistema: Churrasco
Usuário: E o que é isso?
Sistema: Churrasco é o nome do prato feito a base de carne in natura ou processada, assada sobre o fogo ou brasas, com a utilização de estacas de madeira ou metal – chamados de espetos – ou de grelhas
Usuário: O que vai no churrasco?
Sistema: Comida típica brasileira: fraldinha, picanha, galinha.
Usuário: E quem foi o diretor do filme Titanic?
Sistema: James Cameron
Usuário: E qual a sinopse do filme?
Sistema: Titanic é um filme estadunidense de 1997 escrito e dirigido por James Cameron. É uma história ficcionalizada do naufrágio do RMS Titanic, estrelado por Leonardo DiCaprio e Kate Winslet.

```

Fonte: Elaborada pelo autor.

A situação descrita no item “a” será considerada satisfeita quando o sistema permitir que um assunto previamente citado seja utilizado ao longo das interações futuras, sem necessidade de menções explícitas. O item “b” será considerado satisfatório quando o sistema responder adequadamente a uma mudança no assunto principal de conversação. Um trecho de diálogos gerados pelo sistema e que exemplificam estas duas situações é apresentado na figura 17. Ela ilustra um diálogo gerados pelo sistema e demonstra as possibilidades do modelo desenvolvido para o tratamento dos itens de avaliação elencados. Deve ser destacado que as perguntas como “*E o que é isso?*” estão associadas corretamente ao contexto da pergunta anterior, bem como as perguntas sobre novos assuntos são corretamente tratadas. Durante a execução deste diálogo, os conteúdos foram recuperados na DBPedia após a análise sintática e do reconhecimento do tópico principal das mensagens enviadas pelo usuário.

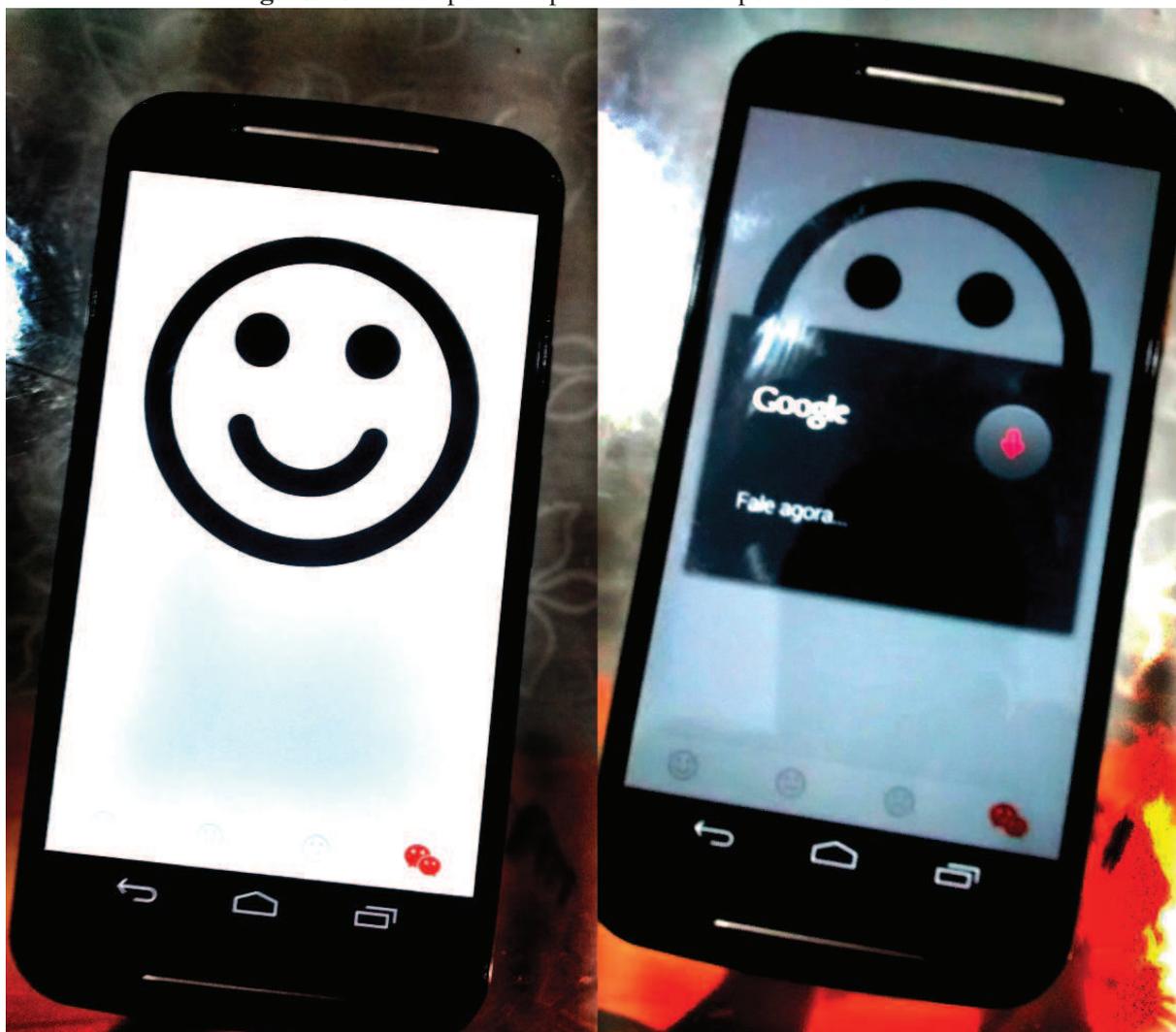
5.2 Detalhes de implementação - Protótipo 2

O protótipo 2, ilustrado na figura 18, possui uma implementação com mais funcionalidades, em relação ao primeiro protótipo, para demonstrar o funcionamento do modelo proposto nesse trabalho. Esse protótipo trata-se de uma aplicação de perguntas e respostas para *smartphones*, com integração de dados oriundos da câmera e do sensor de GPS.

Além disso, o protótipo realiza o reconhecimento de voz para tratar a entrada de dados e sín-

tese de voz para a saída de dados, realizada pela interface gráfica. Essa aplicação foi dividida em duas partes para seu funcionamento devido a limitações de processamento e desenvolvimento na plataforma *mobile*. A primeira parte é composta pela aplicação cliente, que é executada no dispositivo móvel. Esta aplicação foi desenvolvida utilizando Phonegap⁵. O Phonegap é um *framework* de código aberto que permite a criação de aplicações na linguagem HTML5⁶ para as principais plataformas móveis. Apesar da possibilidade da geração de pacotes para outras plataformas, esse protótipo teve como alvo a plataforma Android⁷.

Figura 18: Protótipo 2 - Esperando entrada por voz do usuário.



Fonte: Elaborada pelo autor.

O aplicativo possui uma interface simples, já que a tela é dividida em apenas duas partes. Na parte superior, um *smiley* representa o humor atualmente adotado pelo sistema. Este estado de humor representa uma parte importante da comunicação gerada e está previsto no controle do diálogo e, dessa maneira, é possível interagir com o usuário de formas mais dinâmicas. Na

⁵<http://phonegap.com/>

⁶<http://www.w3.org/TR/html5/>

⁷<https://www.android.com/>

parte inferior da tela, uma barra com pequenos botões representa os possíveis estados de humor do usuário naquele momento. Nesse protótipo, esses botões simulam a entrada da câmera após o reconhecimento da emoção. Essa entrada foi simulada uma vez que técnicas de processamento de imagem e computação afetiva não fazem parte do foco principal do trabalho, que é o gerenciamento do diálogo.

A segunda parte dessa aplicação é executada em um servidor. Isso permite maior flexibilidade nas escolhas de desenvolvimento e técnicas de processamento das informações. Com o objetivo de facilitar a comunicação entre aplicação e servidor, o protótipo utiliza o protocolo HTTP (síncrono) ao invés de utilizar o AMQP, como no protótipo 1. Isso limita em alguns aspectos a troca de informações nos sistemas de interação multimodal, porém foi adotado provisoriamente por facilitar o desenvolvimento e os testes do escopo desse trabalho.

O funcionamento do protótipo começa assim que o usuário executa a aplicação em seu *smartphone*. Após a inicialização, o usuário pode utilizar os botões na barra inferior independentemente de sua ordem. Os primeiros botões enviam ao servidor uma mensagem simulando entradas da câmera. Para reduzir a complexidade dessas entradas, o sistema envia valores representando humor positivo, neutro ou negativo. No sistema, as emoções apresentadas pelo *smiley* superior são um resultado das entradas do usuário, através de comandos simulados da câmera somados a análise de sentimentos obtidos pelas entradas por voz. A análise de sentimentos é realizada pela solução da AlchemyAPI⁸. Para exemplificar esse caso, a frase “*you are ugly*” é identificada como um sentimento negativo para o sistema. Esse valor é considerado juntamente com as últimas interações e assim um *smiley* de resposta (positivo, neutro ou negativo) é apresentado. Do mesmo modo, a expressão “*you are beautiful*” é identificada como sentimento positivo.

As emoções podem influenciar na interação com o usuário dependendo do objetivo de cada sistema. Em um tutor inteligente, por exemplo, uma expressão negativa pode sugerir que o aluno precisa de mais informações. É importante lembrar que em um sistema multimodal, as entradas de câmera não dependem da entrada do usuário em um botão. Elas são feitas de maneira implícita. O modelo prevê uma adoção flexível para o componente responsável pela interpretação destas entradas de dados.

Além das entradas simuladas de emoção, o usuário pode também conversar com o sistema, através de entradas por voz. No capítulo 6 são apresentados possíveis cenários de uso que exemplificam essa funcionalidade. No cenário 1, o usuário inicia o diálogo enviando uma mensagem com o termo “*Hello*” para o sistema. Quando isso acontece, a máquina de estados, descrita em 4.4, passa do seu estado inicial (*START*) para o estado de início de diálogo (*OPENING*). Esse controle é necessário para evitar que o sistema responda de maneira incorreta. Após a mudança de estado, o sistema pode gerar frases de acordo com o seu objetivo (educação, saúde, etc.) ou até mesmo gerar uma resposta de boas-vindas ou outro tipo de cumprimento.

Ao receber uma entrada do usuário, o sistema necessita saber o que essa entrada representa

⁸<http://www.alchemyapi.com/products/alchemy/language/sentiment-analysis/>

antes de mudar o estado atual da máquina de estados e, para isso, o modelo faz uma análise em duas etapas. A primeira etapa consiste em avaliar se a entrada é compatível com um cumprimento na língua portuguesa, através de expressões regulares. Isso permite uma classificação de maneira simplificada. No final desse trabalho, é apresentada uma breve discussão sobre como melhorar essa identificação através de técnicas de classificação do diálogo para a língua portuguesa. A entrada “*Olá*” combina com uma regra de entrada e assim a máquina de estados muda do estado *START* para o estado *OPENING*. Para a adoção do estado de encerramento (*CLOSING*) também são utilizadas regras e expressão regulares.

Caso a entrada não combine com as regras que identificam *OPENING* e *CLOSING*, como “*Quem escreveu Quincas Borba?*”, por exemplo, o sistema irá realizar o processamento de linguagem natural e o reconhecimento de entidades mencionadas. Isso auxilia o sistema a identificar o tópico, ou seja, o assunto principal da entrada do usuário. Nesse caso, o tópico identificado é “*Quincas Borba*”. O sistema deve decidir agora qual o melhor estado da máquina de estados para manter o diálogo de forma apropriada.

Nesse trabalho, o tópico está relacionado a atenção no modelo de Baddeley e Hitch (1974), portanto é necessário saber se o usuário já estava conversando sobre o mesmo assunto na interação anterior. Nesse exemplo, é a primeira vez que o usuário dá uma entrada desse tipo ao sistema, então, o tópico atual, que é nulo, passa a ser “*Quincas Borba*”. O estado da máquina move-se do estado *OPENING* para o estado *REQUEST_RESPONSE*.

Quando a máquina de estados está no estado *REQUEST_RESPONSE*, o sistema entende que, nesse momento, é necessário aprender sobre o tópico em questão. Todo aprendizado sobre o tópico atual será enviado para a pilha da memória verbal, comparada a alça fonológica no modelo de Baddeley e Hitch (1974).

A aprendizagem acontece através do acesso à DBPedia, base de dados ligados utilizada nesse trabalho. Uma vez identificado o tópico, “*Quincas Borba*”, o sistema irá realizar uma busca das possíveis entradas da DBPedia que correspondem ao termo. Para esse exemplo, se obtém apenas um resultado. Outros tipos de pergunta podem gerar resultados não tão exatos. Na pergunta “*Quem dirigiu o filme Titanic?*”, o tópico identificado foi *Titanic*. Ao tentar identificar entradas na DBPedia que possuem relação com o termo *Titanic*, obtemos o seguinte resultado: *Titanic (1997)*, *Titanic 3D*, *Titanic (1953)*, *Titanic (1996)*, *Titanic II*, *Titanic (trilha sonora)*, *Titanic (2012)* e *Titanic (1943)*. Nesse caso, o sistema analisará todas as entradas e buscará obter informações sobre cada uma delas. Nesse mesmo exemplo, a pergunta do usuário pode ser respondida através da entrada *Titanic (1997)*, pela entrada *Titanic (1953)* ou pela entrada *Titanic (1943)*. Esse prototipo foi programado para que o resultado mais recente seja utilizado, *Titanic (1997)*, podendo ser alterado para que o sistema pergunte sobre qual item o usuário está falando.

Para tratar o tópico “*Quincas Borba*”, o sistema carregará o RDF obtido através da consulta no formato JSON-LD. Esse formato foi escolhido apenas para facilitar a implementação, po-

dendo ser utilizado outros formatos baseados em RDF como N-Triples, RDF+XML ou N3⁹. Cada propriedade do RDF possui possíveis valores para responder perguntas do usuário. Uma vez que o sistema encontra uma referência a outra ontologia, essa ontologia também é carregada e seu valor é definido como tópico secundário, que será utilizado a seguir. O modelo utiliza AIML para gerar respostas para determinada pergunta, portanto, é necessário saber qual padrão AIML (*category/pattern*) de pergunta está relacionada a cada propriedade na ontologia.

As propriedades da ontologia são mapeadas com estes padrões através de um arquivo de regras AIML. Esse arquivo serve como base para saber que perguntas podem ser feitas para cada propriedade da ontologia, em português. No final desse trabalho, é apresentada uma breve discussão sobre como substituir esse arquivo de regras por um *corpus* para identificação dessas relações. Durante a análise e carregamento do JSON-LD, o sistema encontra a propriedade *author*. Essa propriedade possui como valor uma URI, que aponta para outra ontologia. O sistema então sabe que deve carregar mais informações dessa URI. No arquivo de regras AIML, existe uma entrada que relaciona o padrão “*quem escreveu **” com a propriedade `<http://dbpedia.org/ontology/author>`. Nesse caso, uma entrada AIML em memória será gerada, conforme a figura 19. É importante observar que o sistema normaliza a *string* para evitar erros na resposta.

Figura 19: Entrada AIML em memória com base na regra e valor da ontologia.

```
<category>
  <pattern>quem escreveu *</pattern>
  <template>Joaquim Maria Machado de Assis</template>
</category>
```

Fonte: Elaborada pelo autor.

Ao encontrar uma relação entre as ontologias, o valor “*Joaquim Maria Machado de Assis*” passa a ser um tópico secundário do assunto. Isso ajuda a proporcionar um diálogo contínuo entre o sistema e o usuário, assim como na conversa entre duas pessoas. Essa mesma abordagem é realizada para todas as outras propriedades recursivamente. Assim, o sistema também carregará a propriedade `<http://dbpedia.org/property/birthDate>`, pois ela corresponde ao padrão “** qual sua data de nascimento*” no arquivo de regras. Após analisar todas as propriedades do JSON-LD e suas ontologias ligadas, o sistema então busca as entradas de AIML em memória e grava em um arquivo AIML temporário. Isso é necessário para executar o método “*aiml.learn()*”, passando como parâmetro o nome desse arquivo.

Após o processo de aprendizagem, o sistema pode responder as perguntas do usuário pois já aprendeu sobre o tópico “*Quincas Borba*”. A última interação do usuário, “*Quem escreveu Quincas Borba?*”, que ainda não foi respondida. O sistema deve então executar o método `aiml.respond(“Quem escreveu Quincas Borba?”)` para obter a resposta “*Joaquim Maria Machado de Assis*”. A resposta é enviada ao usuário e lida através do serviço de *text-to-speech* do

⁹<http://www.w3.org/RDF/>

dispositivo. Também é realizada a análise de sentimento para essa entrada, conforme explicado anteriormente. A pergunta “*Quem escreveu Quincas Borba?*” é classificada como neutra.

A próxima entrada realizada foi “*E qual sua data de nascimento?*”. A máquina de estados está em *REQUEST_RESPONSE*. Novamente o sistema precisa saber como tratar essa pergunta. Primeiramente, é realizado o processamento de linguagem natural e reconhecimento de entidades mencionadas. Essa frase não possui nenhuma entidade reconhecida, então o sistema não considera uma mudança de tópico. Se o tópico não mudar, a máquina de estados passa para *REQUEST_CLARIFICATION*. Essa pergunta será enviada ao método *aiml.respond()* (“*E qual sua data de nascimento?*”), onde a resposta está pré-carregada, pois faz parte de um contexto relacionado e aprendido anteriormente. Dessa forma, o sistema é capaz de responder tanto perguntas relacionadas a “*Joaquim Maria Machado de Assis*” e “*Quincas Borba*” (e demais assuntos possíveis relacionados), realizando uma integração entre regras AIML e os valores retornados pela DBPedia.

Quando o usuário muda o tópico, exemplificado no cenário 2, com a pergunta “*Qual é a extensão da área da bacia do rio Amazonas?*”, o sistema entenderá que o tópico é “*Rio Amazonas*”. Esse será o ponto de atenção, assim como no modelo de Baddeley e Hitch (1974). Não é possível manter o tópico anterior e seus tópicos relacionados, pois o sistema deve respeitar o contexto correto. Após a mudança de assunto, a máquina de estados volta para *REQUEST_RESPONSE* e o tópico anterior é substituído por “*Rio Amazonas*”. Da mesma forma, esse tópico e seus tópicos relacionados serão carregados da DBPedia. O usuário pode também entrar com afirmações ou negações, movendo a máquina de estados para *CONFIRM* e *REJECT* respectivamente. Esses estados não foram implementados no protótipo, mas dão flexibilidade para o sistema tomar ações variadas, como sugerir mais informações ou dar mais opções ao usuário.

6 AVALIAÇÃO

A avaliação do modelo proposto nesse trabalho acontece em duas etapas. A primeira consiste na adoção da avaliação por cenários. O objetivo é elencar e destacar as principais características do modelo em funcionamento a partir de experimentos realizados nos protótipos descritos. O foco principal dessa etapa é apresentar o gerenciamento de diálogo e sua memória de trabalho em situações diferentes. A segunda etapa tem como objetivo principal validar o gerenciamento do diálogo proposto com uso de interações geradas por usuários e obter e comparar com seu resultado esperado.

6.1 Avaliação por cenários

Nesse etapa foi utilizada a avaliação baseada em estudos de cenários, método que pode ser utilizado na validação de projetos que envolvem sensibilidade ao contexto (DEY, 2001). A seguir, são apresentados cenários que destacam as principais características do modelo proposto.

Cenário 1 - Estudante de ensino médio com dúvidas em literatura:

Este cenário pretende demonstrar a utilização do modelo proposto no âmbito educacional, auxiliando um estudante com dúvidas em literatura brasileira. Neste cenário, o sistema mantém um diálogo de fluxo contínuo com o usuário.

Descrição do cenário:

João é estudante da escola estadual de ensino médio da região em que mora. Para o próximo final de semana, a professora pedirá a João e seus colegas que comecem a ler o livro *Quincas Borba*. Marcelo, colega de classe de João, ouviu falar neste livro apenas uma vez. João e Marcelo então decidem ir à biblioteca em busca do livro. A biblioteca da escola é organizada por autores. João e Marcelo não sabem quem é o autor do livro, então se dirigem ao quiosque inteligente da biblioteca a fim de buscar mais informações. João pergunta ao quiosque sobre o autor desse livro. Após ouvir a resposta, Marcelo comenta que este livro deve ser um clássico, pois é de Machado de Assis. João então pergunta algo relacionado à Machado de Assis, sem citar seu nome. O sistema entende que essa pergunta não tem relação com o livro *Quincas Borba*, e sim, relacionada ao seu autor. Marcelo conclui realizando uma pergunta sobre o livro novamente. A figura 20 ilustra o diálogo no protótipo 2.

Detalhes do cenário:

Para a execução do cenário 1, o protótipo inicia o reconhecimento do usuário através de uma câmera. O reconhecimento de um novo usuário permite ao sistema ativamente enviar um “Olá”. A imagem e o resultado do reconhecimento de sentimento são gravados na memória não

Figura 20: Diálogo gerado no cenário 1 pelo protótipo.

```

Quiosque: Olá
João: Bom dia
João: Quem escreveu Quincas Borba?
Quiosque: Joaquim Maria Machado de Assis
João: E qual sua data de nascimento?
Quiosque: 21 de junho de 1839.
Marcelo: E qual o enredo?
Quiosque: Quincas Borba é um romance escrito por Machado de Assis,
desenvolvido em princípio como folhetim na revista A Estação,
entre os anos de 1886 e (...)

```

Fonte: Elaborada pelo autor.

verbal, ou seja, a representação da alça visuo-espacial no modelo de Baddeley e Hitch (1974). Nesse momento, quando João responde “*Bom dia*”, a máquina de estados modifica o seu estado inicial. A figura 20 exibe o resultado obtido pelo protótipo com este cenário.

Através de expressões regulares, o sistema é capaz de identificar possíveis valores de entrada usados frequentemente, como cumprimentos e despedidas. Na próxima interação, onde João pergunta “*Quem escreveu Quincas Borba?*”, o sistema irá utilizar processamento de linguagem natural para identificar a entrada do usuário. Assim, é possível identificar a classe gramatical das palavras envolvidas e então realizar o reconhecimento de entidade mencionada (REM).

Para essa entrada, o sistema reconhecerá “*Quincas Borba*” como entidade mencionada nessa questão. Esse valor será enviado à memória verbal, ou seja, a representação da alça fonológica no modelo de Baddeley e Hitch (1974). Durante essa operação, a máquina de estados tem seu estado alterado. Para que o termo “*Quincas Borba*” tenha algum sentido para o diálogo, o sistema irá buscar informações sobre esse assunto através de consultas realizadas na DBPedia, utilizando o termo “*Quincas Borba*”. Para cada atributo do RDF retornado, uma nova consulta é realizada para obter novas informações relacionadas. O atributo *author*, representado na figura 21, ilustra essa ligação e indica como o processo de coleta de informações relacionadas pode recuperar outras informações complementares, tais como o autor da obra, neste exemplo.

Nesse exemplo, o sistema também seguirá esse atributo para obter informações de “*Joaquim Maria Machado de Assis*”. Esse termo também será adicionado à memória de trabalho do sistema, uma vez que existe relação direta entre “*Quincas Borba*” e “*Joaquim Maria Machado de Assis*”. Assim que a consulta é concluída, o sistema carrega regras que respondem aos atributos do RDF. Com base nessas regras, são geradas as respostas em tempo de execução, através de arquivos AIML. Agora, o sistema pode responder tanto perguntas sobre “*Quincas Borba*” e “*Joaquim Maria Machado de Assis*”. Nesse cenário é possível mostrar o funcionamento do diálogo e seu tópico de atenção, bem como o tratamento da mudança de assunto.

Figura 21: Representação do RDF de Quincas Borba destacando a propriedade author.

```

<dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:Novels_by_Joaquim_Maria_Mi
<dcterms:subject rdf:resource="http://dbpedia.org/resource/Category:1891_novels"/>
<dbpprop:hasPhotoCollection rdf:resource="http://wifo5-03.informatik.uni-mannheim.de/flickrwrapp/
<dbpedia-owl:wikiPageExternalLink rdf:resource="http://metalibri.wikidot.com/title:contos-fluminenses
<dbpedia-owl:wikiPageExternalLink rdf:resource="http://metalibri.wikidot.com/authors:joaquim-maria-
<dbpedia-owl:wikiPageExternalLink rdf:resource="http://metalibri.wikidot.com/title:dom-casmurro"/>
<dbpedia-owl:wikiPageExternalLink rdf:resource="http://metalibri.wikidot.com/title:memorias-postuma
<dbpedia-owl:wikiPageExternalLink rdf:resource="http://metalibri.wikidot.com/title:quincas-borba"/>
<dbpprop:author rdf:resource="http://dbpedia.org/resource/Joaquim_Maria_Machado_de_Assis"/>
<dbpprop:country xml:lang="en">Brazil</dbpprop:country>
<dbpprop:language xml:lang="en">Portuguese</dbpprop:language>
<dbpprop:name xml:lang="en">Quincas Borba</dbpprop:name>
<dbpprop:releaseDate rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">1891</dbpprop:
<dbpedia-owl:abstract xml:lang="pt">Quincas Borba é um romance escrito por Machado de Assis, dese
ser publicado definitivamente pela Livraria Garnier. No processo de adaptação de folhetim para l

```

Fonte: Elaborada pelo autor.

Cenário 2 – Mudança de assunto durante o esclarecimento de dúvidas:

Este cenário pretende demonstrar a utilização do modelo proposto no âmbito educacional, sendo uma continuação do cenário anterior. Neste cenário, o sistema mantém um diálogo de fluxo contínuo, porém identifica a mudança de tópico e deve continuar o diálogo com o usuário.

Descrição do cenário:

Após a última pergunta ao sistema, descrita no cenário 1, João e Marcelo já sabem em que prateleira devem ir buscar o livro Quincas Borba. Antes disso, Marcelo lembra-se que a professora de geografia pediu-lhes para que estudassem sobre o Rio Amazonas e seus afluentes. Marcelo não realizou essa tarefa e tem poucos minutos antes de começar a aula. Ainda na frente do quiosque, João resolve perguntar sobre o Rio Amazonas, para ajudar Marcelo. O quiosque entende que o tópico atual não é mais “*Quincas Borba*” e deve então “aprender” sobre “*Rio Amazonas*” antes de responder. Nesse ponto, o tópico em questão, ou seja, a atenção, gerenciada pela central executiva no modelo de Baddeley e Hitch (1974), é alterado de “*Quincas Borba*” para “*Rio Amazonas*”. Marcelo então faz a próxima pergunta sobre o novo assunto. A figura 22 ilustra o diálogo no protótipo.

Detalhes do cenário:

O sistema identifica o tópico como “*Rio Amazonas*”, porém vê que não é o mesmo da interação anterior. Assim como no cenário 1, as consultas são realizadas na DBPedia para buscar informações com base nos tópicos em questão. Nesse cenário, o ponto fundamental é o reconhecimento de entidades mencionadas e o processamento de linguagem natural. É possível mostrar através deste cenário o controle sobre a troca de contexto da conversação.

Cenário 3 – Decisão de diálogo conforme emoção:

Este cenário pretende demonstrar a utilização do modelo proposto combinando entradas textuais e emoções.

Figura 22: Diálogo gerado no cenário 2 pelo protótipo, em continuação do cenário 1.

Quiosque: Olá
 João: Bom dia
 João: Quem escreveu Quincas Borba?
 Quiosque: Joaquim Maria Machado de Assis
 João: E qual sua data de nascimento?
 Quiosque: 21 de junho de 1839.
 Marcelo: E qual o enredo?
 Quiosque: Quincas Borba é um romance escrito por Machado de Assis, desenvolvido em princípio como folhetim na revista A Estação, entre os anos de 1886 e (...)
 João: Qual é a extensão da área da bacia do rio Amazonas?
 Quiosque: 7.050.000 quilômetros quadrados
 Marcelo: Quais os afluentes do Rio Amazonas?
 Quiosque: Rio Napo, Rio Javari, Rio Jandiatuba, Rio Içá, Rio Jutai, Rio Juruá, Rio Japurá, Rio Tefé, (...)

Fonte: Elaborada pelo autor.

Descrição do cenário:

Maria utiliza uma aplicação em seu *smartphone* para responder perguntas rápidas. Hoje, Maria não está com bom humor devido ao seu dia difícil. Maria executa a aplicação e realiza uma pergunta. Como Maria está na rua, barulhos de carros impediram o reconhecimento correto da frase de Maria. Ela então responde de forma inadequada para o sistema. Com base nas últimas interações, através da câmera e das palavras utilizadas por Maria, o estado emocional do sistema passa a ser considerado como negativo. Esse sistema foi programado para ajudar o usuário em situações ruins. Para tentar melhorar o humor de Maria, o sistema fala uma frase engraçada. Após alguns minutos, o humor detectado através da câmera de Maria é positivo. Maria realiza novamente a pergunta ao sistema, dessa vez realizado com sucesso. O sistema responde e termina com uma frase positiva. A figura 23 ilustra o diálogo gerado pelo protótipo.

Detalhes do cenário:

Nesse cenário, o sistema realiza a fusão multimodal com informações vindas da câmera e da entrada de voz a fim de tratar o humor atual na conversação. A decisão de falar frases positivas ou negativas após analisar interações e estados de humor das entradas anteriores mostra a influência do humor do sistema no diálogo.

*Cenário 4 – Aplicação de *smartphone* para respostas rápidas (futuro).*

Este cenário pretende demonstrar a utilização do modelo proposto nas atividades do cotidiano através de perguntas e respostas rápidas a fim de ajudar o usuário no seu dia-a-dia. É

Figura 23: Representação do diálogo gerado no cenário 3 pelo protótipo.

Maria: Olá
 Sistema: Olá
 Maria: Onde eu es... (ruído).
 Sistema: Não entendi. Tente falar novamente.
 Maria: Sistema estúpido!
 Sistema: Entendo.
 Sistema: Cara feia pra mim é falta de maquiagem.
 Maria: Onde eu estou?
 Sistema: Você está na Rua Anita Garibaldi,
 bairro Boa Vista, Porto Alegre.
 Sistema: Quem bom que está feliz!

Fonte: Elaborada pelo autor.

importante deixar claro que este cenário foi citado apenas como possível cenário futuro e não foi validado pelo protótipo atual. Acredita-se que as informações disponíveis sejam suficientes para garantir o seu funcionamento. Este cenário não foi validado por conter elementos na sua implementação que fogem do escopo dessa dissertação.

Descrição do possível cenário:

Paula trabalha em um escritório no centro da cidade e cursa Administração na Universidade Federal no turno da noite. Paula não lê jornais antes de sair de casa, entretanto, costuma utilizar uma aplicação em seu *smartphone* para acompanhar as notícias. A aplicação é seu assistente para perguntas rápidas, a fim de ajudar Paula no seu dia-a-dia.

- Paula: Olá
- Aplicação: Bom dia
- Paula: Qual a previsão do tempo?
- Aplicação: Sol com poucas nuvens durante o dia. Períodos nublados, porém sem a possibilidade de chuva

Para a execução do cenário 4, o sistema realiza a fusão multimodal a nível semântico, obtendo dados do GPS e interpretando-os para responder sobre a previsão do tempo de uma determinada cidade ou região. Após receber a informação, Paula, que não gosta de dias chuvosos, dá um sorriso. Através da câmera, o sistema interpreta a entrada como uma emoção positiva, exibindo um sorriso na tela, falando de forma correspondente:

- *Aplicação: Que bom! Você parece estar feliz! Tenha um ótimo dia!*

Nesse cenário, o ponto fundamental é a utilização das entradas de contexto e a utilização das mesmas junto com a entrada de voz do usuário. A fusão de entradas multimodais é importante para deixar o sistema mais robusto. Durante o diálogo, a informação sobre a localização é armazenada na memória não verbal, ou seja, a alça visuo-espacial no modelo de Baddeley e Hitch (1974).

6.2 Avaliação com usuários

Nesta etapa foi solicitado a um grupo de dez pessoas que escrevessem algumas questões diretas para o sistema responder, bem como a resposta que esperavam para aquelas perguntas. As perguntas criadas por estes usuários foram utilizadas em sessões de interação com o protótipo 2 e então devolvidas com as respostas geradas pelo sistema. Estas respostas foram comparadas às respostas esperadas por cada usuário. Esse experimento foi importante para saber como as pessoas fazem suas questões, de maneira geral, criando assim um cenário de testes mais eficaz, evitando atrasos devido a disponibilidade de cada participante e foco na validação do modelo ao invés do protótipo.

As pessoas participantes deste experimento foram selecionadas por conveniência e possuem um perfil adequado para o experimento. A faixa etária varia entre 22 e 57 anos. As profissões dos participantes são predominantemente das áreas de tecnologia da informação, administração, saúde e comunicação. Os convidados para o experimento receberam instruções para preencherem um formulário com os elementos necessários para a realização do teste. Este formulário está apresentado no anexo A desse trabalho, enquanto as respostas recebidas dos participantes deste experimento estão descritas no anexo B.

Foi solicitado a cada usuário a indicação de um conjunto de dados que permite validar se o protótipo realiza a sua operação de acordo com o esperado, ou seja, detectar o contexto atual e responder a uma pergunta relacionada. Deste modo, os participantes indicaram uma pergunta inicial e sua respectiva resposta, além de uma segunda pergunta sobre o contexto geral da primeira e da resposta esperada. Um exemplo deste conjunto de dados que ilustra esta atividade pode ser exemplificada com a pergunta “*Qual a capital da Noruega?*”, sendo que a resposta esperada pelo participante foi “*Oslo*”. Já a segunda pergunta foi “*Quantos habitantes ela possui?*”, para a qual espera-se a resposta “*593 mil habitantes*”.

Na avaliação utilizou-se a escala Likert (1932) para obter os níveis de precisão após medir a qualidade das funcionalidades descritas nesse trabalho. A tabela 2 mostra os resultados obtidos após a comparação entre as repostas geradas pelo sistema e as respostas esperadas pelos usuários. Nessa tabela, estão listados os diferentes testes realizados com o material enviado pelos participantes para uso na avaliação do protótipo.

No primeiro teste aplicado (*Identificou o tópico*), o usuário deveria escolher assuntos de determinadas categorias e realizar uma pergunta. O sistema deve identificar o assunto e responder

qual o tópico principal a que o usuário está se referindo. Para entender melhor o motivo das categorias pré-selecionadas, serão abordadas as limitações de processamento de linguagem natural da língua portuguesa no capítulo 7. É importante lembrar que, embora as frases geradas pelos usuários foram aplicadas através de um protótipo, o objetivo é avaliar o modelo proposto no escopo desse trabalho.

Tabela 2: Percepção da comparação entre as respostas geradas pelo sistema e das esperadas pelos usuários.

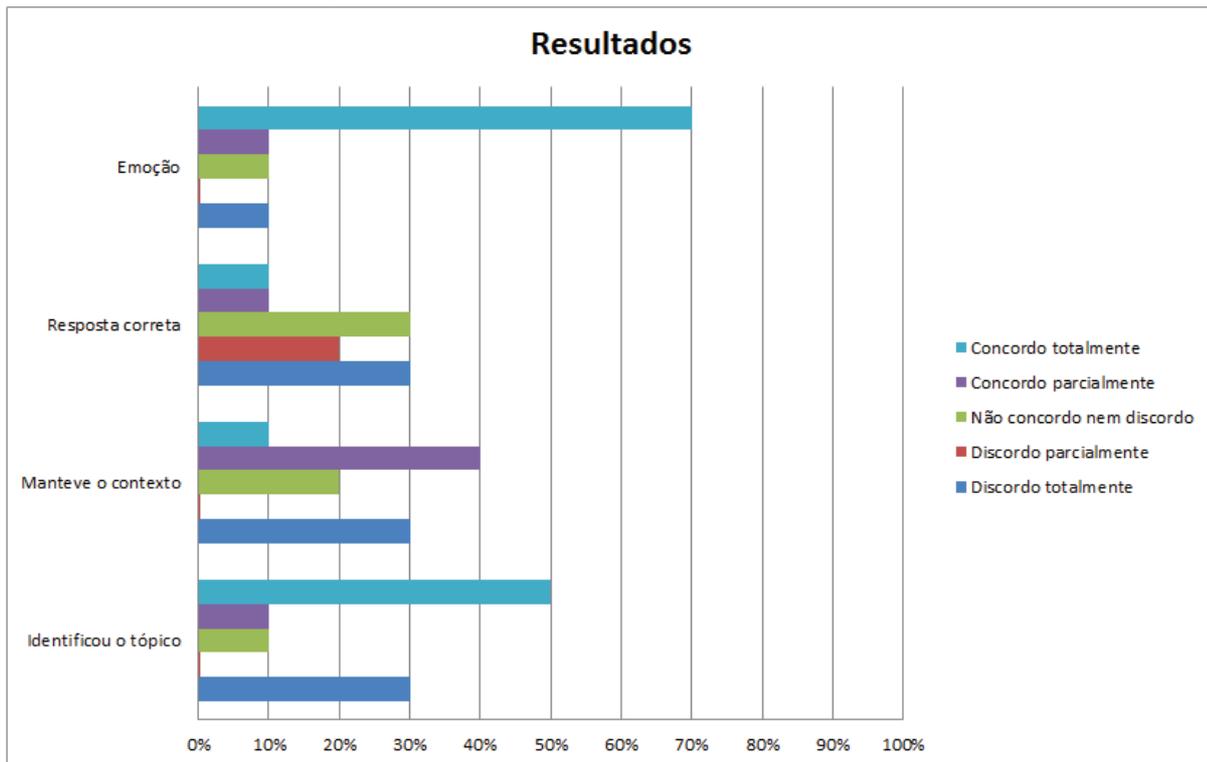
Questão	Discordo totalmente	Discordo parcialmente	Não concordo nem discordo	Concordo parcialmente	Concordo totalmente
1. Identificou o tópico	30%	0%	10%	10%	50%
2. Manteve o contexto	30%	0%	20%	40%	10%
3. Resposta correta	30%	20%	30%	10%	10%
4. Emoção	10%	0%	10%	10%	70%

Fonte: Elaborada pelo autor.

No segundo teste (*Manteve o contexto*) foi solicitado a cada usuário que realizasse uma pergunta correspondente ao resultado da primeira pergunta, sem que mencionasse o assunto principal. O terceiro teste (*Resposta correta*) compara as respostas geradas pelo sistema com as respostas esperadas pelos usuários. No último teste (*Emoção*) foi solicitado que o usuário enviasse duas frases simples, uma que considerasse negativa e outra que considerasse positiva.

É possível observar que na maioria das situações o tópico da pergunta foi identificado corretamente. É necessário trazer algumas observações nesse ponto, pois alguns dos resultados são influenciados pelos componentes auxiliares, como o processamento de linguagem natural e a disponibilidade de informações existentes na DBPedia. Por exemplo, um dos usuários perguntou “*Quantos meses têm o ano?*” e “*Qual o mês do natal?*”. Não foi possível encontrar “*Ano*” como entidade mencionada e também não foi possível encontrar uma relação entre “*12 meses*” e “*Natal*” descrita na DBPedia. Deste modo, o resultado avaliado foi considerado negativo, entretanto, é possível identificar que as necessidades para superar esta situação estão relacionadas com os componentes de processamento de linguagem natural e com representação de conhecimento.

Em outro exemplo analisado, o participante fez as perguntas “*Quem escreveu O Capital?*” e “*Qual sua data de falecimento?*”. Nesse caso também não foi possível reconhecer “*O Capital*” como uma entidade mencionada em português. Porém, ao citar *Karl Marx*, o sistema conseguiu reconhecer a sua data de falecimento corretamente. O problema ocorreu também com as perguntas “*Onde estão Scooby-Doo e Salsicha?*” e “*O que eles procuram?*”. O sistema não conseguiu reconhecer “*Scooby-Doo*” como entidade mencionada. A segunda pergunta foi subjetiva e relacionada com o conhecimento do enredo que envolve estes dois personagens, logo o protótipo não foi planejado para tal questão. Já as perguntas “*Quem matou Osama Bin Laden?*”

Figura 24: Resultados dos formulários de avaliação.

Fonte: Elaborada pelo autor.

e “*Onde ele foi morto?*” foram tratadas corretamente, porém a DBPedia não possui a informação esperada pelo usuário, “*Rob O’Neill*”. Nesse caso, entretanto, a segunda pergunta deste grupo foi respondida corretamente. Para as perguntas “*Quem é Stephen William Hawking?*” e “*Qual a sua doença?*”, a DBPedia não possui uma ligação direta entre “*Stephen William Hawking*” e a sua doença, o que não permitiu tratar adequadamente a segunda questões. A figura 24 mostra um comparativo entre os resultados obtidos.

É possível perceber que cada usuário tem uma maneira de realizar as perguntas, mesmo seguindo as instruções básicas informadas. Nota-se que o reconhecimento de entidades mencionadas em português, com proposta adotada, pode ser considerado insuficiente em alguns casos. Quando o tópico é identificado corretamente, o modelo proposto nesse trabalho trata de responder às questões de forma mais natural, mantendo o contexto da conversação.

7 CONCLUSÃO

Neste trabalho foram explorados subsídios para a melhoria do controle do diálogo em sistemas de interação multimodal, através da pesquisa de características conhecidas dos modelos de memória de trabalho e comunicação humana. Buscou-se a identificação de aspectos que possam atuar na geração de melhores formatos de diálogo, com a manutenção e tratamento de informações que descrevem a memória deste diálogo e que incluam os novos eventos no contexto geral da conversação, de forma integrada às informações anteriores. Além disso, no modelo proposto nessa dissertação, é simulada a utilização de memória de longo prazo, a partir de consultas realizadas a bases de dados ligados ou bases de conhecimento, complementando as informações disponíveis e permitindo que o sistema apresente aspectos mais flexíveis.

Nesse trabalho foram estudados conceitos de diversas áreas que se relacionam diretamente com a interação multimodal. Algumas áreas da ciência da computação podem representar aportes importantes para o desenvolvimento e ampliação do trabalho. São elas: o processamento de linguagem natural, a recuperação de informação, a computação móvel, a computação ubíqua, a computação afetiva e a web semântica. Também foram estudados conceitos relacionados a áreas como linguística e psicologia, que apoiaram a concepção do modelo proposto. Com base na literatura, foi possível conhecer e entender melhor muitos conceitos e arquiteturas propostas em trabalhos anteriores.

A questão de pesquisa abordada neste trabalho tratou de identificar possibilidades de melhorias nos resultados de diálogos gerados por sistemas de interação multimodal, no sentido de permitir que sejam percebidos como mais próximos das interações realizadas entre duas pessoas. A investigação principal esteve associada com o controle do diálogo e teve sua concepção influenciada por um modelo cognitivo que considerou alguns dos aspectos conhecidos sobre o funcionamento da memória de trabalho, tal como definidos no modelo de Baddeley e Hitch (1974). No modelo proposto foram representados os elementos principais que compõem a estrutura sugerida para a memória de trabalho, com a expectativa de que, desta forma, os protótipos desenvolvidos proporcionariam condições de geração de diálogos percebidos como naturais pelos usuários.

A partir da análise dos trabalhos relacionados, pode-se considerar que o modelo proposto apresenta como diferencial a utilização de aspectos da memória de trabalho e do contexto do diálogo de forma integrada. Com isso, busca-se a geração de diálogos mais robustos e com percepção mais adequada por parte do usuário. As validações realizadas com base na avaliação de cenários e testes de usabilidade permitem identificar a possibilidade da criação de contextos mais complexos para diálogos a partir de informações multimodais. O modelo também apresenta aspectos relevantes de flexibilidade, podendo ser utilizado em diferentes configurações de sistemas de interação multimodal. Com a implementação de dois protótipos, descritos em 5.1 e 5.2, foi possível validar os protocolos e formatos propostos para o modelo, como a utilização do padrão EMMA e do protocolo AMQP. A utilização da memória, da atenção, do contexto e

das decisões tomadas pelo sistema em cada situação ficou foi descrita nas seções 4.4 e 5.2.

Por fim, as avaliações realizadas a partir de questões propostas por usuários permitiram identificar uma percepção positiva com relação ao contexto das mensagens. Estas avaliações também possibilitaram constatar a dependência dos protótipos quanto aos seus componentes, em especial os componentes de tratamento da base de conhecimento utilizada e de processamento de linguagem natural. Nos casos em que estes componentes não possuem a informação desejada ou em que não conseguem processar adequadamente as mensagens, o resultado não é percebido de forma satisfatória. É possível observar estes aspectos em outros sistemas em que a dependência de conhecimento específico de uma área pode ser determinante para o resultado do sistema de interação multimodal.

Considera-se que as principais contribuições do trabalho estão relacionadas com o esforço de integração do estudo nas diversas áreas de conhecimento e técnicas exploradas para a composição do modelo aqui proposto. Outro aspecto a destacar é a característica de arquitetura genérica e flexível, o que permitiu a sua utilização em dois contextos diferenciados, para dois protótipos de validação implementados. Também considera-se relevante a experimentação com a integração de base de dados ligadas como parte do modelo, através uso das informações disponíveis na DBPedia, que possibilitaram a simulação dos diálogos para o contexto de perguntas e respostas utilizado como elemento de validação.

O desenvolvimento de melhores recursos para o tratamento de interação multimodal é um tema relevante atualmente, tendo em vista o aumento de opções de dispositivos que permitem a captura de dados representativos de diversos aspectos da comunicação humana. As aplicações de interação multimodal possibilitam melhorar a utilização de sistemas de computação em situações específicas, como em educação ou saúde, nas quais os usuários podem ser amplamente beneficiados com a maior flexibilidade gerada. Além disso, é cada vez maior a quantidade de bases de dados abertos e ligados disponíveis para consultas e utilização de forma adicional aos sistemas de interação multimodal, com a vantagem de apresentarem descrições semânticas relevantes.

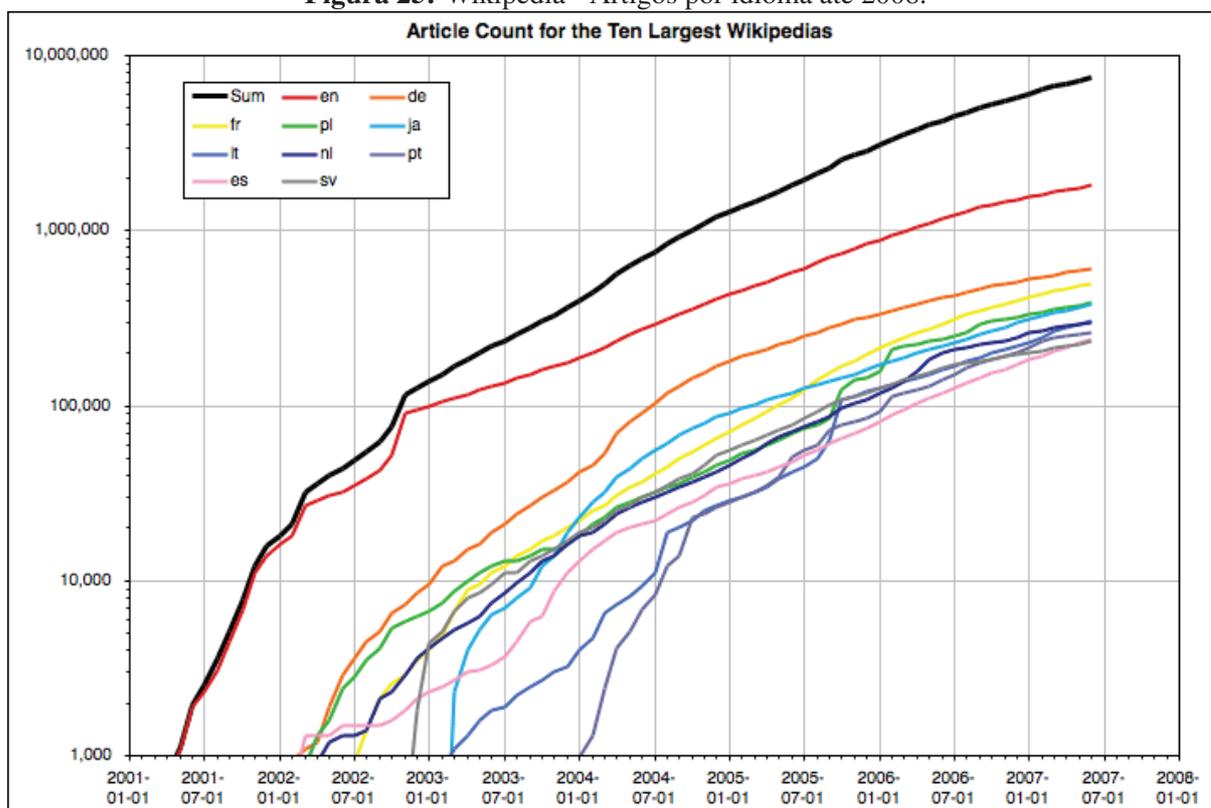
7.1 Limitações e trabalhos futuros

Nesse trabalho, procurou-se tratar as entradas considerando a língua portuguesa. A DBPedia é um projeto que tem como base a Wikipedia¹. A figura 25 mostra a diferença entre artigos em inglês e outros idiomas até 2008. A primeira limitação desse trabalho foi encontrar informações sobre alguns itens na DBPedia em português, pois muitas vezes a informação somente estava disponível na DBPedia em inglês. A estratégia implementada no protótipo foi a de trazer a informação em inglês quando não disponível em português.

A segunda limitação desse modelo está relacionada a área de processamento de linguagem natural. Nesse caso, estudos da língua portuguesa estão muito atrás da língua inglesa, assim, o

¹<http://pt.wikipedia.org/>

Figura 25: Wikipedia - Artigos por idioma até 2008.



Fonte: (WIKIMEDIA, 2014)

reconhecimento de entidades mencionadas fica prejudicado em muitos aspectos. Nos protótipos desenvolvidos, quando não é possível reconhecer uma entidade mencionada, o sistema traduz a frase para língua inglesa, a fim de obter um resultado. Essa abordagem resolveu alguns problemas, porém gera muitos falsos-positivos. Como trabalho futuro, pode-se estudar melhores maneiras de tratar a língua portuguesa nessas situações.

Ainda relacionado a língua portuguesa, muitos nomes de livros, por exemplo, possuem palavras simples. O título da série *O Tempo e o Vento*, de Érico Veríssimo, não foi identificada em nenhuma ocasião, assim como demais exemplos citados na avaliação. Outro ponto está relacionado à identificação de cumprimentos e despedidas através de expressão regular. É necessário encontrar uma abordagem robusta para a língua portuguesa, como *dialogue act tagging*, para tornar essa identificação mais precisa. Falsos-positivos podem acontecer também utilizando a busca por *keywords*. A implementação de algoritmos de busca por relevância como TF-IDF e Okapi BM25 ou soluções mais abrangentes pode evitar erros ao localizar tópicos na DBPedia, como sugerido no artigo de Bobed, Esteban e Mena (2012).

A área de interação multimodal é composta por muitas áreas de conhecimento relacionadas. Esse trabalho procurou trazer cenários e implementações voltadas a sistemas de interação multimodal que pudessem responder perguntas de maneira mais natural, através do controle de diálogo. Uma futura aplicação do método proposto será a integração do modelo aqui proposto com o sistema VOX (F. J. SERÓN C. BOBED, 2014), envolvendo o aspecto de interação com

o ambiente já delimitado de agentes comunicacionais.

O AIML possui uma abordagem simples e funcional, utilizada para responder ao usuário nesse trabalho. Cada propriedade de uma ontologia disponível a partir da DBPedia é mapeada através de um arquivo de regras. Uma possível melhoria seria encontrar uma abordagem para saber quais perguntas têm como resposta o valor de uma propriedade da ontologia. Por exemplo: Sendo uma entidade do tipo *Book* e a propriedade *dbpprop:author*, deve-se encontrar com base em um *corpus* perguntas como “*Quem escreveu?*” ou “*Quem é o autor?*”, através de técnicas apropriadas, de maneira automática.

Em relação ao controle do diálogo, é possível observar na avaliação de cenários que o sistema não mantém um fluxo de diálogo para dois usuários diferentes interagindo ao mesmo tempo com o sistema. Uma das possíveis linhas de atuação futura seria acrescentar essa distinção ao modelo a fim de obter uma interação mais natural com múltiplos usuários.

Pretende-se realizar testes com maior quantidade de usuários, em ambiente educacional, voltado para temas que sejam disponíveis na base de conhecimento utilizada, com vistas a tratar dos resultados possíveis para uma grande quantidade de usuários e para observar especificamente aspectos de melhoria possíveis com o uso de melhores recursos de PLN.

Por fim, identificou-se como um trabalho futuro interessante a adoção de mecanismos de clarificação de perguntas, que seriam adotados como complementares ao diálogo e serviriam para tratar aquelas situações em que as mensagens dos usuários não são adequadamente identificadas ou nas situações em que o conhecimento disponível nas bases de conhecimento consultadas não atende exatamente às demandas recebidas dos usuários do sistema. O código utilizado nesse trabalho está disponível em <https://github.com/jonathansp/>.

REFERÊNCIAS

- ALLAM, A. M. N.; HAGGAG, M. H. The Question Answering Systems: a survey. **International Journal of Research and Reviews in Information Sciences (IJRRIS)**, [S.l.], 2012.
- AUER, S.; BIZER, C.; KOBILAROV, G.; LEHMANN, J.; CYGANIAK, R.; IVES, Z. **Dbpedia: a nucleus for a web of open data**. [S.l.]: Springer, 2007.
- BADDELEY, A. The episodic buffer: a new component of working memory? **Trends in cognitive sciences**, [S.l.], v. 4, n. 11, p. 417–423, 2000.
- BADDELEY, A. D.; HITCH, G. Working memory. **Psychology of learning and motivation**, [S.l.], v. 8, p. 47–89, 1974.
- BALDASSARRI, S.; CERESO, E.; SERON, F. J. Maxine: a platform for embodied animated agents. **Computers & Graphics**, [S.l.], v. 32, n. 4, p. 430–437, 2008.
- BARNETT, J.; BODELL, M.; RAGGETT, D.; WAHBE, A. Multimodal architecture and interfaces. **W3C Working Draft**, [S.l.], v. 11, 2012.
- BERANT, J.; CHOU, A.; FROSTIG, R.; LIANG, P. Semantic Parsing on Freebase from Question-Answer Pairs. In: EMNLP, 2013. **Anais...** [S.l.: s.n.], 2013. p. 1533–1544.
- BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, [S.l.], v. 284, n. 5, p. 1–5, 2001.
- BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data-the story so far. **International journal on semantic web and information systems**, [S.l.], v. 5, n. 3, p. 1–22, 2009.
- BLATTNER, M. M.; GLINERT, E. P. Multimodal Integration. **IEEE Multimedia**, Los Alamitos, CA, USA, v. 3, n. 4, p. 14–24, 1996.
- BOBED, C.; ESTEBAN, G.; MENA, E. Ontology-driven Keyword-based Search on Linked Data. In: KES, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1899–1908.
- BOLT, R. A. ‘Put-that-there’: voice and gesture at the graphics interface. In: ANNUAL CONFERENCE ON COMPUTER GRAPHICS AND INTERACTIVE TECHNIQUES, 7., 1980, New York, NY, USA. **Proceedings...** ACM, 1980. p. 262–270. (SIGGRAPH '80).
- BRAY, T.; PAOLI, J.; SPERBERG-MCQUEEN, C. M.; MALER, E.; YERGEAU, F. **Extensible markup language (xml) 1.0**. 2004.
- BUI, T. H. **Multimodal Dialogue Management - State of the Art**. Enschede: Centre for Telematics and Information Technology, University of Twente, 2006. (TR-CTIT-06-01).
- BUSH, N.; WALLACE, R.; RINGATE, T.; TAYLOR, A.; BAER, J. Artificial Intelligence Markup Language (AIML) Version 1.0. 1. **Prieiga per internetą: <http://www.alicebot.org/TR/2001/WD-aiml> [žiūrėta 2005 m. kovo 15 d.]**, [S.l.], 2001.

CARMO, T. d. R. **Uso do padrão AMQP para transporte de mensagens entre atores remotos**. 2012. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, 2012.

CASSELL, J.; SULLIVAN, J.; PREVOST, S.; CHURCHILL, E. **Embodied Conversational Agents**. 2000. [S.l.]: MIT Press, 2000.

CEREZO, E.; BALDASSARRI, S.; HUPONT, I.; SERON, F. J. Affective embodied conversational agents for natural interaction. **ISBN**, [S.l.], p. 978–3, 2008.

CHAI, J.; PAN, S.; ZHOU, M. Mind: a context-based multimodal interpretation framework in conversational systems. In: KUPPEVELT, J. van; DYBKJÆR, L.; BERNSEN, N. (Ed.). **Advances in Natural Multimodal Dialogue Systems**. [S.l.]: Springer Netherlands, 2005. p. 265–285. (Text, Speech and Language Technology, v. 30).

COHEN, P. R.; JOHNSTON, M.; MCGEE, D.; OVIATT, S.; PITTMAN, J.; SMITH, I.; CHEN, L.; CLOW, J. Quickset: multimodal interaction for distributed applications. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 1997. **Proceedings...** [S.l.: s.n.], 1997. p. 31–40.

CORRADINI, A.; MEHTA, M.; BERNSEN, N. O.; MARTIN, J.; ABRILIAN, S. Multimodal input fusion in human-computer interaction. **NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES**, [S.l.], v. 198, p. 223, 2005.

CUTUGNO, F.; LEANO, V. A.; RINALDI, R.; MIGNINI, G. Multimodal Framework for Mobile Interaction. In: INTERNATIONAL WORKING CONFERENCE ON ADVANCED VISUAL INTERFACES, 2012, New York, NY, USA. **Proceedings...** ACM, 2012. p. 197–203. (AVI '12).

DAMLJANOVIĆ, D.; AGATONOVIĆ, M.; CUNNINGHAM, H.; BONTCHEVA, K. Improving habitability of natural language interfaces for querying ontologies with feedback and clarification dialogues. **Web Semantics: Science, Services and Agents on the World Wide Web**, [S.l.], v. 19, p. 1–21, 2013.

DAMOVA, M.; DANNELLS, D.; ENACHE, R.; MATEVA, M.; RANTA, A. Natural language interaction with semantic web knowledge bases and lod. **Towards the Multilingual Semantic Web**. Springer, Heidelberg, Germany, [S.l.], 2013.

DEY, A. K. Understanding and Using Context. **Personal Ubiquitous Comput.**, London, UK, UK, v. 5, n. 1, p. 4–7, Jan. 2001.

DI FABBRIZIO, G.; LEWIS, C. Florence: a dialogue manager framework for spoken dialogue systems. In: INTERSPEECH, 2004. **Anais...** [S.l.: s.n.], 2004.

DUMAS, B.; LALANNE, D.; OVIATT, S. Multimodal interfaces: a survey of principles, models and frameworks. In: **Human Machine Interaction**. [S.l.]: Springer, 2009. p. 3–26.

DWIVEDI, S. K.; SINGH, V. Research and Reviews in Question Answering System. **Procedia Technology**, [S.l.], v. 10, p. 417–424, 2013.

ENCICLOPÉDIA, E. **Lisboa**: imprensa. [S.l.: s.n.], 1991.

EYBEN, F.; WÖLLMER, M.; SCHULLER, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In: INTERNATIONAL CONFERENCE ON MULTIMEDIA, 2010, New York, NY, USA. **Proceedings...** ACM, 2010. p. 1459–1462. (MM '10).

FERNÁNDEZ, M. I. (Ed.). The Vox System. In: _____. **Augmented Virtual Realities for Social Developments. Experiences between Europe and Latin America.** [S.l.]: Universidad de Belgrano, ISBN 978-950-757-046-9, 2014. p. 137–166.

GODOY, J. P. M. C. **Integração de informações visuais e verbais na memória de trabalho.** 2010. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, 2010.

GONZALES, M.; LIMA, V. S. Recuperação de Informação e Processamento da Linguagem Natural. In: ACM, 2003. **Anais...** [S.l.: s.n.], 2003.

GROSSMAN, D. A. **Information retrieval: algorithms and heuristics.** [S.l.]: Springer, 2004. v. 15.

GROSZ, B. J.; SIDNER, C. L. Attention, intentions, and the structure of discourse. **Computational linguistics**, [S.l.], v. 12, n. 3, p. 175–204, 1986.

GROSZ, B. J.; WEINSTEIN, S.; JOSHI, A. K. Centering: a framework for modeling the local coherence of discourse. **Computational linguistics**, [S.l.], v. 21, n. 2, p. 203–225, 1995.

GUPTA, P.; GUPTA, V. Article: a survey of text question answering techniques. **International Journal of Computer Applications**, [S.l.], v. 53, n. 4, p. 1–8, September 2012. Published by Foundation of Computer Science, New York, USA.

HAKIMOV, S.; TUNC, H.; AKIMALIEV, M.; DOGDU, E. Semantic question answering system over linked data using relational patterns. In: JOINT EDBT/ICDT 2013 WORKSHOPS, 2013. **Proceedings...** [S.l.: s.n.], 2013. p. 83–88.

HARTSON, H. R.; HIX, D. Human-computer interface development: concepts and systems for its management. **ACM Computing Surveys (CSUR)**, [S.l.], v. 21, n. 1, p. 5–92, 1989.

HELENE, A. F.; XAVIER, G. F. A construção da atenção a partir da memória. **Revista Brasileira de Psiquiatria**, [S.l.], v. 25, p. 12 – 20, 12 2003.

HIGHTOWER, J.; BRUMITT, B.; BORRIELLO, G. The location stack: a layered model for location in ubiquitous computing. In: MOBILE COMPUTING SYSTEMS AND APPLICATIONS, 2002. PROCEEDINGS FOURTH IEEE WORKSHOP ON, 2002. **Anais...** [S.l.: s.n.], 2002. p. 22–28.

HOSTE, L.; DUMAS, B.; SIGNER, B. Mudra: a unified multimodal interaction framework. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES, 13., 2011, New York, NY, USA. **Proceedings...** ACM, 2011. p. 97–104. (ICMI '11).

IZQUIERDO, I. **Memória.** 2. ed. [S.l.]: Editora Artmed, 2011.

JAIMES, A.; SEBE, N. Multimodal human–computer interaction: a survey. **Computer vision and image understanding**, [S.l.], v. 108, n. 1, p. 116–134, 2007.

JOHNSTON, M.; BAGGIA, P.; BURNETT, D.; CARTER, J.; DAHL, D.; MCCOBB, G.; RAGGETT, D. Emma: extensible multimodal annotation markup language. **World Wide Web Consortium Recommendation REC-emma-2009021. Technical report, W3C**, [S.l.], 2009.

JOHNSTON, M.; BANGALORE, S.; VASIREDDY, G. MATCH: multimodal access to city help. In: AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING, 2001. ASRU '01. IEEE WORKSHOP ON, 2001. **Anais...** [S.l.: s.n.], 2001. p. 256–259.

JUNIOR, I.; REIS, V. dos. **Um framework para desenvolvimento de interfaces multimodais em aplicações de computação ubíqua**. 2007. Dissertação (Mestrado em Ciência da Computação) — Universidade de São Paulo, 2007.

JURAFSKY, D.; MARTIN, J. H. **Speech & language processing**. [S.l.]: Pearson Education India, 2000.

KÖCHE, J. C. Fundamentos de metodologia científica: teoria da ciência e iniciação à pesquisa. atual. **Petrópolis: Vozes**, [S.l.], 2002.

KRESS, G.; LEEUWEN, T. van. **Multimodal Discourse: the modes and media of contemporary communication**. 1. ed. New York: Oxford University Press, 2001.

LI, W.; SRIHARI, R. K.; LI, X.; SRIKANTH, M.; ZHANG, X.; NIU, C. Extracting Exact Answers to Questions Based on Structural Links. In: CONFERENCE ON MULTILINGUAL SUMMARIZATION AND QUESTION ANSWERING - VOLUME 19, 2002., 2002, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2002. p. 1–9. (MultiSumQA '02).

LIDDY, E. D. Natural language processing. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2001. **Anais...** [S.l.: s.n.], 2001.

LIE, D.; HULSTIJN, J.; AKKER, R. op den; NIJHOLT, A. A transformational approach to natural language understanding in dialogue systems. **Proceedings Natural Language Processing and Industrial Applications (NLP+ IA'98)**, [S.l.], v. 2, p. 163–168, 1998.

LIKERT, R. A technique for the measurement of attitudes. **Archives of psychology**, [S.l.], 1932.

LOPER, E.; BIRD, S. NLTK: the natural language toolkit. In: ACL-02 WORKSHOP ON EFFECTIVE TOOLS AND METHODOLOGIES FOR TEACHING NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL LINGUISTICS - VOLUME 1, 2002, Stroudsburg, PA, USA. **Proceedings...** Association for Computational Linguistics, 2002. p. 63–70. (ETMTNLP '02).

MCCOBB, G. **The W3C Multimodal Architecture, Part 1: overview and challenges**. [S.l.]: IBM, 2007. Disponível em: <<http://www.ibm.com/developerworks/web/library/wa-multimodarch1/>>. Acesso em: 20 mai. 2014.

MERKEL, A.; KLAKEW, D. Comparing Improved Language Models for Sentence Retrieval in Question Answering. In: IN PROCEEDINGS OF CLIN, 2007. **Anais...** [S.l.: s.n.], 2007.

MISHRA, T.; BANGALORE, S. Qme!: a speech-based question-answering system on mobile devices. In: HUMAN LANGUAGE TECHNOLOGIES: THE 2010 ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 2010. **Anais...** [S.l.: s.n.], 2010. p. 55–63.

NETO, A. T.; BITTAR, T. J.; FORTES, R. P. M.; FELIZARDO, K. Developing and Evaluating Web Multimodal Interfaces - a Case Study with Usability Principles. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2009., 2009, New York, NY, USA. **Proceedings...** ACM, 2009. p. 116–120. (SAC '09).

NOCK, H. J.; IYENGAR, G.; NETI, C. Multimodal Processing by Finding Common Cause. **Commun. ACM**, New York, NY, USA, v. 47, n. 1, p. 51–56, Jan. 2004.

OVIATT, S. Multimodal interfaces. **The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications**, [S.l.], p. 286–304, 2003.

OVIATT, S.; COHEN, P. Perceptual user interfaces: multimodal interfaces that process what comes naturally. **Communications of the ACM**, [S.l.], v. 43, n. 3, p. 45–53, 2000.

OVIATT, S.; COULSTON, R.; LUNSFORD, R. When Do We Interact Multimodally?: cognitive load and multimodal communication patterns. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES, 6., 2004, New York, NY, USA. **Proceedings...** ACM, 2004. p. 129–136. (ICMI '04).

PFLEGER, N. Context based multimodal fusion. In: MULTIMODAL INTERFACES, 6., 2004. **Proceedings...** ACM, 2004. p. 265–272.

POPPEL, E. **Fronteiras da Consciência - A Realidade e a Experiência do Mundo**. [S.l.]: Edições 70, 1989.

REITHINGER, N.; SONNTAG, D. An integration framework for a mobile multimodal dialogue system accessing the semantic web. In: INTERSPEECH, 2005. **Anais...** [S.l.: s.n.], 2005. p. 841–844.

RHODES, B. J.; MINAR, N.; WEAVER, J. Wearable computing meets ubiquitous computing: reaping the best of both worlds. In: WEARABLE COMPUTERS, 1999. DIGEST OF PAPERS. THE THIRD INTERNATIONAL SYMPOSIUM ON, 1999. **Anais...** [S.l.: s.n.], 1999. p. 141–149.

ROQUE, F. M.; PEIXOTO, C. S. A. Inclusão de um módulo de inteligência na arquitetura MMIF da W3C. In: MOSTRA ACADEMICA UNIMEP, 9., 2011. **Proceedings...** [S.l.: s.n.], 2011.

ROTARU, M. **Applications of discourse structure for spoken dialogue systems**. [S.l.]: ProQuest, 2008.

SCHRÖDER, M. The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. **Adv. in Hum.-Comp. Int.**, New York, NY, United States, v. 2010, Jan. 2010.

SEBE, N. Multimodal interfaces: challenges and perspectives. **Journal of Ambient Intelligence and smart environments**, [S.l.], v. 1, n. 1, p. 23–30, 2009.

Semiótica. In: _____. **Infopédia**. [S.l.]: Porto: Porto Editora, 2003.

SHEKARPOUR, S.; NGONGA NGOMO, A.-C.; AUER, S. Question answering on interlinked data. In: **WORLD WIDE WEB**, 22., 2013. **Proceedings...** [S.l.: s.n.], 2013. p. 1145–1156.

STONE, D.; JARRETT, C.; WOODROFFE, M.; MINOCHA, S. **User interface design and evaluation**. [S.l.]: Morgan Kaufmann, 2005.

TAN, J.; DUAN, F.; INAMURA, T. Multimodal human-robot interaction with Chatterbot system: extending aiml towards supporting embodied interactions. In: **ROBOTICS AND BIOMIMETICS (ROBIO), 2012 IEEE INTERNATIONAL CONFERENCE ON**, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1727–1732.

TURK, M. Multimodal interaction: a review. **Pattern Recognition Letters**, [S.l.], v. 36, p. 189–195, 2014.

VINOSKI, S. Advanced message queuing protocol. **IEEE Internet Computing**, [S.l.], v. 10, n. 6, p. 87–89, 2006.

VO, M. T.; WOOD, C. Building an application framework for speech and pen input integration in multimodal learning interfaces. In: **ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1996. ICASSP-96. CONFERENCE PROCEEDINGS.**, 1996 IEEE INTERNATIONAL CONFERENCE ON, 1996. **Anais...** [S.l.: s.n.], 1996. v. 6, p. 3545–3548 vol. 6.

VOORHEES, E. M.; HARMAN, D. Overview of the Ninth Text REtrieval Conference (TREC-9). In: **TREC, 2000. Anais...** [S.l.: s.n.], 2000.

WALKER, J. Through the looking glass. **The Art of Human-computer Interface Design**. **New York: Addison-Wesley**, [S.l.], 1990.

WAZLAWICK, R. **Metodologia de pesquisa para ciência da computação**. [S.l.]: Elsevier Brasil, 2009.

WEISER, M. The computer for the 21st century. **Scientific american**, [S.l.], v. 265, n. 3, p. 94–104, 1991.

WIKIMEDIA. **Size of Wikipedia** - http://en.wikipedia.org/wiki/wikipedia:size_of_wikipedia. 2014.

YEH, T.; DARRELL, T. Multimodal question answering for mobile devices. In: **INTELLIGENT USER INTERFACES**, 13., 2008. **Proceedings...** [S.l.: s.n.], 2008. p. 405–408.

ANEXO A FORMULÁRIO ENVIADO AOS PARTICIPANTES DO EXPERIMENTO.

Avaliação de sistema pergunta e resposta - Dissertação de Mestrado em Computação Aplicada

Estou concluindo o curso de Mestrado em Computação Aplicada pela Unisinos e trabalhando em um projeto de um sistema para geração de respostas automáticas. Este é um trabalho de conclusão onde foi criado um protótipo para fazer essa avaliação. O protótipo não foi liberado pois ainda está em desenvolvimento e disponível apenas para dispositivos Android.

As plataformas Android e IOS possuem seus próprios aplicativos de assistentes pessoais inteligentes (Google Now e Siri). Esses programas utilizam diversos dados integrados para processar a informação, como localização, voz e gestos. O protótipo no meu trabalho tem como objetivo tentar deixar a utilização desses assistentes (e sistemas similares) mais robusta. Para isso preciso fazer um primeiro teste preliminar. Sendo assim, gostaria que você contribuisse com o meu trabalho preenchendo um formulário conforme as instruções abaixo. Não leva mais do que cinco minutos.

No primeiro quadro, escreva uma pergunta. Como o programa ainda é um protótipo, peço que utilize perguntas curtas e diretas (quem foi, como foi, onde foi) sobre pessoas (políticos, escritores, atores e empresários) ou cidades grandes (Nova Iorque, Paris), por exemplo. Depois, no segundo quadro, escreva a resposta esperada para a pergunta.

EXEMPLO:

Pergunta Direta:

* Quem escreveu Quincas Borba?

Respostas Esperadas:

* Machado de Assis

Após, peço que escreva uma pergunta relacionada ao quadro 2, de forma continuada.

EXEMPLO:

Pergunta Relacionada à Resposta Anterior:

* Qual a sua data de nascimento? (Não citar Machado de Assis).

Resposta Esperada Relacionada a Pergunta Anterior

* 21 de junho de 1839

Isso é necessário para treinar e encontrar problemas com o sistema proposto no meu trabalho.

Por último, peço que digite uma frase que considere positiva e uma frase que considere negativa. O objetivo desse ponto é analisar o humor do usuário no protótipo. Essas frases não precisam estar relacionadas entre si, nem com as questões anteriores.

EXEMPLO:

Uma frase que você considera positiva:

* Você é bonito

Uma frase que você considera negativa:

* Você é feio

Se quiser colaborar mais, preencha o formulário novamente.

O objetivo desse trabalho é entender se a percepção das respostas corretas pelos usuários é a mesma (ou similar) a gerada pelo sistema. Após reunir todas as avaliações, publico o resultado final.

Muito obrigado!

* Required

Pergunta Direta *

Exemplo: Quem escreveu Quincas Borba?

Resposta Esperada *

Exemplo: Machado de Assis

Pergunta Relacionada à Resposta Anterior *

Exemplo: Qual a sua data de nascimento? (Não citar Machado de Assis).

Resposta Esperada Relacionada a Pergunta Anterior *

Exemplo: 21 de junho de 1839

Uma frase que você considera positiva: *

Exemplo: Você é bonito

Uma frase que você considera negativa: *

Exemplo: Você é feio

Seu nome, idade, e-mail e profissão. *Exemplo: João Silva, 40, joao@terra.com.br, Gerente Administrativo
Autoriza a Unisinos a ter acesso a esse formulário? Dados não serão publicados nem divulgados! *

Sim ou Não

Submit

Never submit passwords through Google Forms.

100%: You made it.

ANEXO B DADOS RECEBIDOS DOS PARTICIPANTES DURANTE O EXPERIMENTO.

Quem é a presidente do Brasil?	Dilma Rouseff.	Qual é o seu partido?	PT
Quantos meses tem um ano?	12 meses	Qual o mês do natal?	dezembro
Em que país se localiza a cidade de Lima?	Peru	Qual seu continente ?	América do Sul
Quem escreveu O Capital?	Karl Marx	Qual sua data de falecimento?	14 de Março de 1883
Onde estão Scooby-Doo e Salsicha?	Estão dentro daquela casa assombrada	O que procuram?	Desaparecimento de uma obra de arte.
Quem escreveu Memórias Póstumas de Brás Cubas?	Machado de Assis	Em que estado nasceu o autor?	Rio de Janeiro
Quem matou Osama Bin Laden?	Rob O'Neill	Onde ele foi morto?	Na cidade de Abbottabad.
Quem é Barack Obama?	Presidente dos Estados Unidos.	Qual a sua religião?	Cristianismo
Qual a capital da Noruega?	Oslo	Quantos habitantes possuem?	593 mil habitantes
Quem é Stephen William Hawking?	Físico teórico e comólogo britânico.	Qual a sua doença?	Esclerose Lateral Amiotrófica (ELA)