UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ANÁLISE DA INFERÊNCIA HUMANA E COMPUTACIONAL DAS ASSOCIAÇÕES ENTRE MENSAGENS DE BATE-PAPO

LUIZ EDUARDO XAVIER DE CASTRO PEREIRA

Orientador

Mariano Pimentel

RIO DE JANEIRO, RJ – BRASIL

SETEMBRO DE 2015

Pereira, Luiz Eduardo Xavier de Castro.
P436 Análise da inferência humana e computacional das associações entre mensagens de bate-papo / Luiz Eduardo Xavier de Castro Pereira, 2015.
99 f. ; 30 cm

Orientador: Mariano Pimentel.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2015.

1. Processamento de linguagem natural (Computação). 2. Análise da conversação. 3. Educação. 4. Grupo de bate-papo pela Internet.

5. Mensagens instantâneas. 6. Linguística - Processamento de dados. I. Pimentel, Mariano. II. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. III. Título.

CDD – 001.535

Análise da inferência humana e computacional das associações entre mensagens de bate-papo

Luiz Eduardo Xavier de Castro Pereira

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

_____

Mariano Pimentel, D.Sc. (UNIRIO)

_____

Sean Wolfgand Matsui Siqueira, D.Sc. (UNIRIO)

_____

Hugo Fuks, D.Sc. (PUC-Rio)

Dedicatória

Dedico esse trabalho a Nele Xavier, minha vó - raiz forte da minha família.
Em sua existência, o milagre do universo.
Em minha vida e memória - meu o amor eterno.

Agradecimentos

Primeiramente gostaria de agradecer a minha família. As minhas mães Regina Xavier, Lúcia Xavier e Thereza de Castro. A jornada até esse momento foi possível porque vocês foram minhas primeiras orientadoras. Também gostaria de agradecer a Aline Thuller, por compreender minha ausência, me apoiar e também me orientar na jornada até aqui.

Também gostaria de agradecer meu eterno orientador Pimentel, por tudo que me ensinou, pela paciência, por acalmar minha ansiedade e pelos incentivos. Neste programa de mestrado eu tive a melhor experiência acadêmica até hoje. Embora muito atarefado eu consegui vivenciar ótimas experiências, incluindo a oportunidade apresentar minha pesquisa no Simpósio de sistemas colaborativos onde tive feedbacks valiosos. Nada disso seria possível sem a dedicação e a transformação que os professores da Unirio fizeram em minha vida. Assim, agradeço a todos os professores do programa de Mestrado pela oportunidade pegar um "cadinho" da sua intelectualidade, em especial a Fernanda Baião, que muito inspirou e ensinou. A Professora Flávia Santoro que me deixou apaixonado por metodologias cientifica, A Simone Bacellar que me recebeu no primeiro dia e me ajudou nos primeiros passos. Ao professor Sean, que também me recebeu no primeiro dia e me acompanhou durante esse percurso sempre me ajudando muito e fazendo acréscimos valiosos na minha pesquisa. A professora Renata Araújo, que muito inspirou em reflexões sobre a profissão e sobre a vida e professora Kate Revoredo por toda dedicação e conhecimento. Muito obrigado pela oportunidade.

SETEMBRO DE 2015

PEREIRA, Luiz Eduardo Xavier de Castro. **Inferência Humana e Computacional das associações entre mensagens de bate-papo**. UNIRIO, 2015. 103 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

Resumo

Esta dissertação investiga o fenômeno da confusão da conversação que eventualmente leva à perda do co-texto. O objetivo leva a estudar o grau que os seres humanos têm para inferir as associações entre mensagens numa sessão de bate-papo e, também, questionar em que grau um algoritmo consegue inferir corretamente as associações. A quantidade de associações corretas e incorretas é usada, nesta dissertação, para responder as questões de estudo. Para obter essas informações, foi elaborado um estudo de caso exploratório para caracterizar e avaliar a confusão da conversação. Assim foram definidas as métricas para análise: percentual de erros e acertos de inferências de associações, tempo para inferir-desistir, medir e evidenciar. Para avaliar em que grau um algoritmo pode associar mensagens, foram estudados algoritmos de mineração de textos e processamento de linguagem natural. E o resultado humano e do algoritmo foram estudados através de uma análise comparativa.

Palavras-chave: Processamento de Linguagem Natural; Mineração de Texto, Educação; Confusão da Conversação, co-texto, associações, mensagens, Bate-papo.

ABSTRACT

This dissertation investigates the conversation confusion phenomenon which eventually leads to loss of co-text. The aim leads to study the degree that human beings have to infer associations between messages in a chat session and also question to what extent an algorithm can correctly infer associations. The number of correct and incorrect associations is used in this research to answer the study questions. To obtain this information, it was designed an exploratory case study to characterize and evaluate the confusion of the conversation. Thus it was defined the metrics for analysis percentage of hits and misses associations of inferences, inferred time to give up, measure and evidence. To assess to what degree an algorithm can associate messages were studied algorithms for text mining and natural language processing. Also, the human result and the algorithm were studied through a comparative analysis.

Keywords: Natural Language Processing; Text mining, education; Confusion Conversation, co-text, associations, messaging, Chat.

SUMÁRIO

INDICE DE TABELAS

INDICE DE FIGURAS

INDICE DE TRANSCRIÇÕES

1 Introdução

Este capítulo tem por objetivo a apresentação da pesquisa documentada nesta dissertação sobre a compreensão da conversação no bate-papo, mais especificamente, sobre as associações implícitas entre as mensagens, o que precisa ser inferido pelos leitores da sessão de bate-papo. Esta pesquisa se justifica como argumentado na Seção 1.1, não só pela popularização do uso do bate-papo em sistemas de redes sociais como o Facebook, mas principalmente pelo crescimento da educação online no Brasil, que faz uso do bate-papo, dentre outros sistemas, como fórum e e-mail. A presente pesquisa se insere no contexto do projeto Portal Tagarelas, apresentado na Seção 1.2, que tem por objetivo desenvolver e investigar suporte computacional para apoiar a conversação por bate-papo no contexto educacional.

A seção 1.3 apresenta a visão geral da pesquisa, onde se aponta a confusão que ocorre em sistemas de bate-papo e a investigação sobre como as pessoas se confundem, além das dificuldades inerentes à interpretação das mensagens durante uma sessão de bate-papo. Nesta função, será possível que um algoritmo pode associar mensagens melhor do que um ser humano? Para alcançar o objetivo da pesquisa, foi necessário responder algumas questões diferentes e, neste caso, foi utilizado o método de pesquisa exposto na seção 1.4. Na seção 1.5 é apresentada a organização da escrita da dissertação.

1.1 **Motivação: a expansão da internet, da EAD e do uso do bate-papo**

"Hoje, a internet liga milhões de computadores pelo mundo e é a maior máquina inventada por humanos até agora" (IFENTHALER, 2010, p. 4). A iniciativa da internet.org[1], que inclui suporte de várias entidades públicas e privadas, em menos de um ano incluiu cerca de um bilhão de pessoas online ao redor do mundo e, com a recente parceria[2] do governo brasileiro, há expectativa de aumentar ainda mais o número de usuários no Brasil, que já em 2012 chegou a 80 milhões (IDGNOW, 2012).

[1] https://internet.org/press/one-year-in-internet-dot-org-free-basic-services
[2]http://www.brasil.gov.br/governo/2015/04/dilma-e-zuckerberg-conversam-sobre-inclusao-digital-e-conectividade-no-panama

Dentre os vários fenômenos sociotécnicos que emergiram com a popularização da internet destacamos a educação online. Em 2000, no Brasil, não havia praticamente curso algum. Uma década depois, a atividade passou a ser responsável por quase 15% das matrículas na graduação (INEP, 2010) – Figura 1.

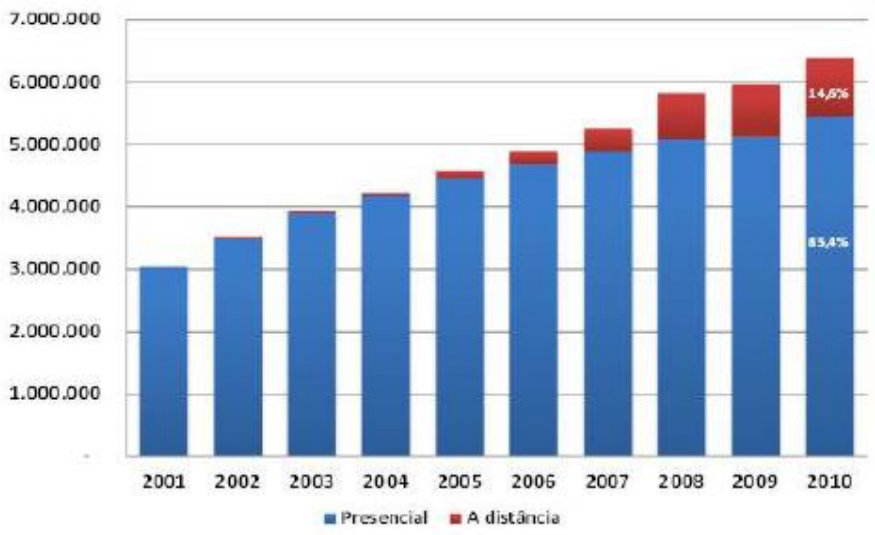


Figura 1 – Crescimento da EAD no Brasil (INEP, 2010, p.10 apud Rocha, 2013)

Nas instituições brasileiras de educação à distância, o bate-papo e o fórum são os meios de interação online mais utilizados. A Figura 2 mostra o percentual de utilização de cada meio de comunicação pelas instituições. Cerca 70% das instituições usam o Bate-papo.

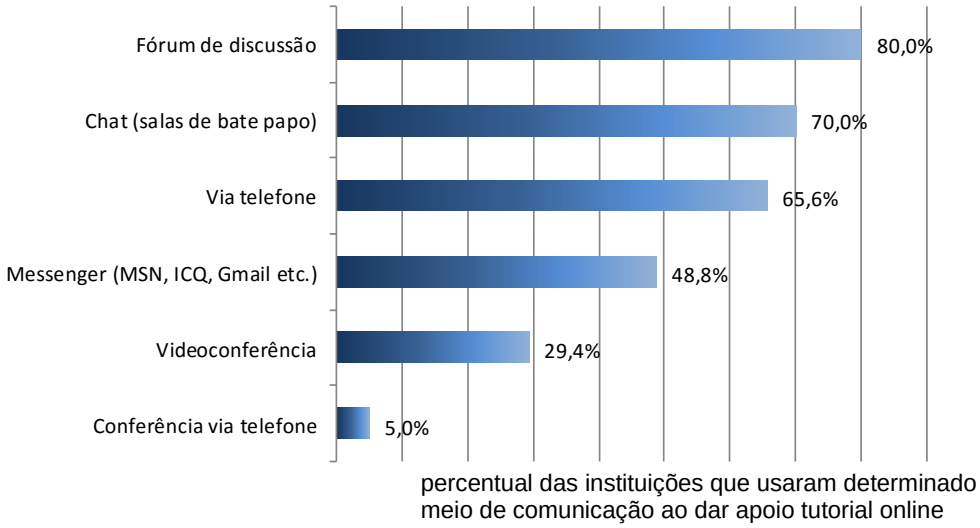


Figura 2. Recursos online adotados nas instituições de EAD (CENSOEAD.BR, 2010, p. 10 apud Rocha, 2013)

Dada à expansão da internet e no aumento da procura pela educação online, aliado ao fato dos sistemas de bate-papo serem um dos recursos mais utilizados nesta modalidade, cada vez mais pessoas se reunirão online para estudar e debater, o que justifica a realização desta pesquisa que investiga a compreensão da conversação que se realiza pelo bate-papo.

1.2 **Contexto da Pesquisa: Portal Tagarelas**

Esta pesquisa está inserida no contexto do projeto de pesquisa Portal Tagarelas (PIMENTEL & STRUC, 2012), ilustrado na Figura 3, em que se desenvolve e investiga suporte computacional para apoiar o uso de bate-papo na educação online. O suporte computacional inclui sistemas de bate-papo específicos para dinâmicas educacionais, e sistemas para auxiliar a pré-reunião (sistemas de cadastro de turmas, agenda das sessões, notificadores etc) e a pós-reunião (arquivamento, analisadores do log da sessão de bate-papo, avaliadores da participação etc).

Alguns sistemas de bate-papo mais específicos para a educação já foram desenvolvidos pelo grupo de pesquisa, tais como o *MediatedChat* (PIMENTEL, FUKS, & LUCENA, 2005), produzido para facilitar a atuação de um moderador que pode coordenar a conversação por meio de técnicas de conversação (fala um de cada vez [contribuição circular], contribuição única [cada participante só pode emitir uma única mensagem], contribuição livre ou todos bloqueados); *TabChats* (AZEVEDO, 2011), utilizado para facilitar a realização de uma aula organizada em assuntos, sendo um tema discutido separadamente em abas para discussão; *InterVIU* (NUNES R. R., 2009), que permite a realização de entrevistas pela turma e organiza a conversação em pares de pergunta-resposta; HiperDiálogo (PIMENTEL M. , 2002), Debatepapo (MORAIS, 2011) e o Debatepapo v2 (NETTO, 2014) , desenvolvidos para o debate e discussão livre, em que as mensagens podem ser associadas umas com as outras para facilitar o acompanhamento do desdobramento do discurso.

Neste contexto, principalmente avançando o conhecimento construído sobre a confusão da conversação no bate-papo e sobre a perda de co-texto (PIMENTEL, 2003; PIMENTEL e FUKS, 2009; MORAIS, 2011; NETTO, 2014), a presente contribuiu com a produção de conhecimento sobre a dificuldade para se inferir as associações entre mensagens de bate-papo, o que contribuiu para a confusão e, eventualmente, resulta na perda de co-texto (fenômenos discutidos no próximo capítulo).

Figura 3. Portal Tagarelas

1.3 **Visão Geral da pesquisa**

Os sistemas típicos de bate-papos não permitem que o participante formalize a associação de sua mensagem com a que está respondendo, a mensagem referente, o co-

texto. A ausência desta informação traz dificuldades para os leitores, pois não é possível identificar quem está falando com quem. Da mesma forma, também não é possível acompanhar os desdobramentos do discurso nas mensagens, que ficam registradas em ordem cronológica no log da conversação, mas que efetivamente estabelecem relações não-lineares umas com as outras, resultando num emaranhado de mensagens cujas associações cabem ao leitor inferir. O fato potencializa a confusão da conversação e eventualmente leva à perda de co-texto, fenômeno em que o leitor se sente perdido sem conseguir inferir a que mensagem anterior uma dada mensagem se refere. A confusão da conversação se torna mais percebida principalmente em "bate-papos sérios", como os realizados em contextos educacionais ou em reuniões de trabalho, situações onde os participantes precisam acompanhar o fluxo de mensagens da sessão (PIMENTEL & FUKS, 2009).

Quando as mensagens estão associadas, as pessoas identificam o co-texto mais corretamente e mais rapidamente, conforme já constatado em pesquisas empíricas anteriores (MORAIS, 2011). Na presente pesquisa objetiva-se investigar o complementar, conhecer a dificuldade que as pessoas têm para inferir as associações quando as mensagens não estão associadas e verificar se a computação, em comparação com os humanos, consegue inferir corretamente, através de um algoritmo, as associações entre as mensagens de bate-papo.

Esta pesquisa se realizou por meio de um estudo de caso exploratório (YIN, 2009) A questão de pesquisa foi desdobrada em questões mais específicas: com que frequência as pessoas acertam/erram as inferências que fazem sobre as associações entre as mensagens de bate-papo? As pessoas sempre fazem uma dedução, ainda que errada, ou assumem a incapacidade de inferir uma possível associação entre a mensagem? As pessoas demoram muito tempo para fazer uma inferência? Podemos dizer que o tempo que a pessoa demora para perceber uma associação significa que ela está mais perdida e que, consequentemente, tem mais chance de inferir erradamente ou de desistir? E finalmente, comparativamente ao desempenho humano, um algoritmo apresenta um desempenho equivalente, menor, muito menor, maior ou muito maior?

Com o presente estudo também foi possível definir métricas para evidenciar, medir e avaliar a confusão e a perda de co-texto no bate-papo e avaliar o desempenho de um algoritmo levando em consideração o desempenho humano.

1.4 **Método da pesquisa**

Nesta dissertação, foi realizado um estudo de caso exploratório (YIN, 2009; PIMENTEL, 2012; RECKER, 2013) que busca investigar, de forma empírica, os fenômenos num contexto real. Para isso, a pesquisa analisou como ocorre a confusão na conversa em sistemas típicos de bate-papo, investigando a sua ocorrência em uma sessão de bate-papo realizada em turmas reais de disciplinas pertencentes ao programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro (UFRJ).

O estudo teve por objetivo medir o desempenho daqueles participantes ao perceberem as associações entre as mensagens trocadas na sessão de bate-papo realizada. Em seguida, o log da sessão foi processado por um sistema que infere as relações entre as mensagens de bate-papo. Desta forma, foi possível medir o desempenho algorítmico e compará-lo com o desempenho humano.

As seguintes etapas foram realizadas para que o objetivo da pesquisa fosse alcançado:

1. Revisão da literatura e identificação do problema da pesquisa;
2. Projeto de estudo de caso e realização de um estudo piloto para validar o projeto;
3. Estudo de caso exploratório e medição dos erros das pessoas ao inferirem as associações entre mensagens;
4. Análise dos erros que as pessoas comentem ao realizarem as associações;
5. Medição dos erros de um algoritmo ao inferir as associações entre mensagens de bate-papo, e sua comparação com a performance humana.

No **primeiro passo** da pesquisa, foi feita a revisão da literatura sobre o problema da pesquisa e das técnicas de sua resolução, assim como da literatura sobre a confusão do bate-papo e a perda de co-texto. Em paralelo, foi discutida a modelagem da conversação na estrutura matemática de Grafos e, também, as técnicas algorítmicas para inferir as associações entre mensagens de bate-papo.

No **segundo e terceiro passos**, foi projetado e realizado um estudo de caso exploratório com o objetivo de analisar o grau de dificuldade das pessoas para inferirem as associações entre as mensagens de bate-papo. As turmas de alunos participaram de um debate através de um sistema típico de bate-papo (sem o registro das associações entre as mensagens). Posteriormente, os participantes tiveram que compreender a relação de cada mensagem com alguma anterior, sendo registrada a inferência estabelecida por cada participante e o tempo que demoraram para a conclusão da atividade. Como resultado deste

estudo, identificou-se que os participantes erram quase 50% das inferências. O resultado nos surpreendeu negativamente, pois consideramos muito alta a quantidade de erros: a cada duas mensagens, o participante infere erradamente o co-texto (referencial) de uma mensagem. Perante este resultado, foram estudados os erros cometidos por aqueles participantes, o que foi realizado no **quarto passo**.

No **quinto passo,** foi investigado o algoritmo proposto por Lima (2013) de classificação estatística, que usa um modelo estatístico baseado em características léxicas e estruturais das mensagens para inferir as associações entre mensagens de bate-papo e que também é baseado algumas das estratégias humanas para realizar inferências identificadas por Pimentel e Fuks (2009). O desempenho deste algoritmo então é comparado ao desempenho humano e discute-se seus resultados.

1.5 **Organização da Escrita**

Após o capítulo inicial, o presente documento possui outros cinco capítulos. O Capítulo 2 desenvolve o referencial teórico, apresentando o problema da confusão do bate-papo e perda do co-texto, bem com a revisão da literatura sobre mineração de textos e processamento de linguagens naturais. O Capítulo 3 descreve detalhadamente o desenvolvimento do estudo de caso sobre as inferências humanas sobre as associações entre mensagens e, também, os artefatos construídos que foram necessários para a condução do estudo. O Capítulo 4 descreve os padrões de erros realizados por humanos, com exemplos que mostram como e porque as pessoas erram as associações. O Capítulo 5 apresenta o algoritmo de análise do diálogo, as medições do seu resultado e a comparação com o desempenho humano. E, por fim, o Capítulo 6 expõe as condições e contribuições obtidas com esse trabalho para futuras pesquisas.

2 Referencial teórico

O objetivo deste capítulo é revisar a literatura relacionada à associação de mensagens em bate-papo. Na Seção 2.1, define-se o bate-papo e seus elementos típicos de interface. Em seguida, na Seção 2.2, são debatidas as associações entre as mensagens de bate-papo, incluindo uma discussão sobre as estratégias humanas para inferir estas associações. Na Seção 2.3, discute-se a confusão da conversação no bate-papo, focando o problema da perda de co-texto que ocorre quando um participante não consegue inferir uma associação. Alguns sistemas já foram desenvolvidos para possibilitar os próprios usuários associarem suas mensagens durante a conversação no bate-papo. Na Seção 2.4, é apresentada a estrutura árvore frequentemente usada para modelar matematicamente as associações entre as mensagens de uma conversação no bate-papo e por fim, na Seção 2.5, são apresentadas algumas técnicas computacionais para se tentar inferir, algoritmicamente, as associações entre mensagens.

2.1 **O que é um bate-papo?**

O bate-papo é um meio de conversação mediada pela computação em que os interlocutores estão conectados ao mesmo tempo e trocam mensagens textuais curtas (CALVÃO, PIMENTEL, FUKS, & LOPES, 2014). Não há comunicação face-a-face, já que os participantes estão distribuídos em locais diferentes. O bate-papo estabelece uma conversa em tempo real, isto é, a conversação é síncrona, a troca de mensagens é quase instantânea, pois uma mensagem enviada é distribuída em questão de milissegundos para todos os interlocutores conectados na sessão de bate-papo. O sincronismo da conversa requer que os interlocutores estejam conectados ao mesmo tempo, por isso, muitos sistemas de bate-papo disponibilizam algum mecanismo de percepção da presença, dispositivo que permite indicar que interlocutores estão presentes na sessão, como exemplifica a lista de participantes ilustrada na Figura 4. Por se tratar de uma forma de comunicação ligeira, os interlocutores produzem mensagens curtas, com linguagem informal e com características de oralidade, tornando o tom do discurso pessoal e emotivo.

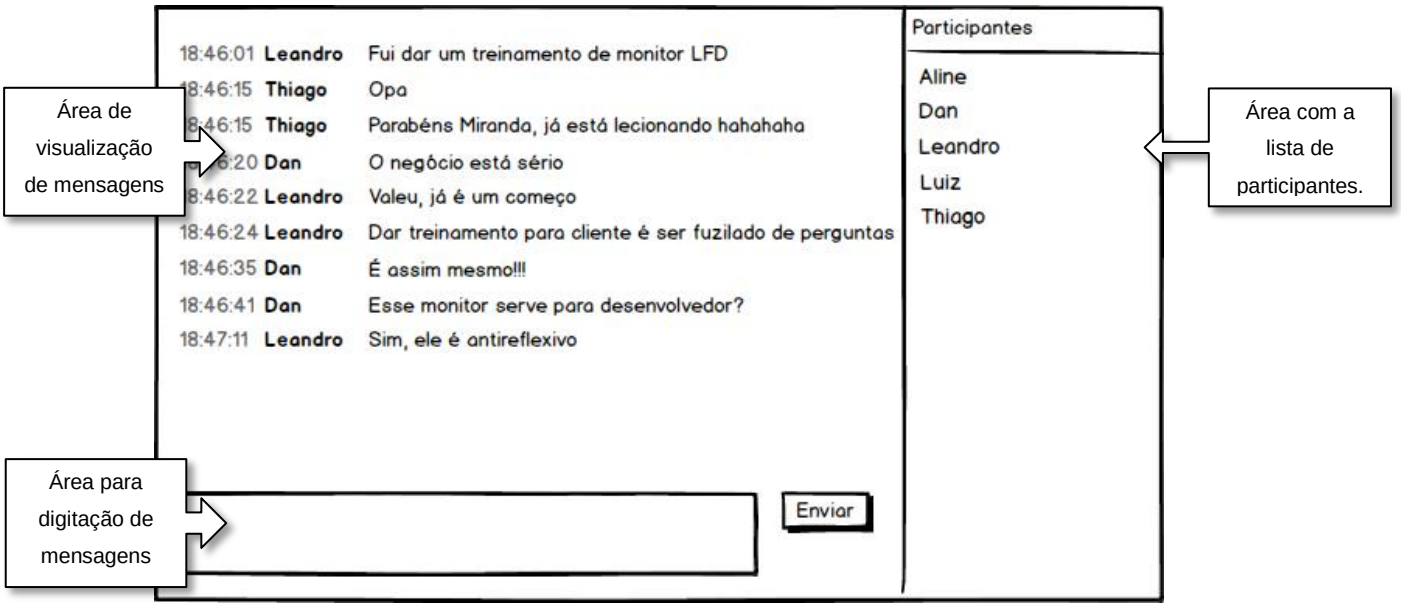


Figura 4 – Elementos típicos da interface de um sistema de bate-papo

Os sistemas de bate-papo são usados para diferentes propósitos. É comum encontrar a implementação de bate-papo como um serviço de conversação em jogos *online*, sistemas de redes sociais, sites de comércio eletrônico, sistemas de educação a distância etc. Em jogos *online*, por exemplo, a mensagem de bate-papo é frequentemente representada em forma de balão visualizado pelos personagens próximos ao do jogador (PARK *et al*, 2008). Conforme ilustrado na Figura 5, no Facebook também há uma implementação de bate-papo, além de vários outros serviços de conversação, como microblog, mensageiro instantâneo, mensagens em grupo e e-mail.

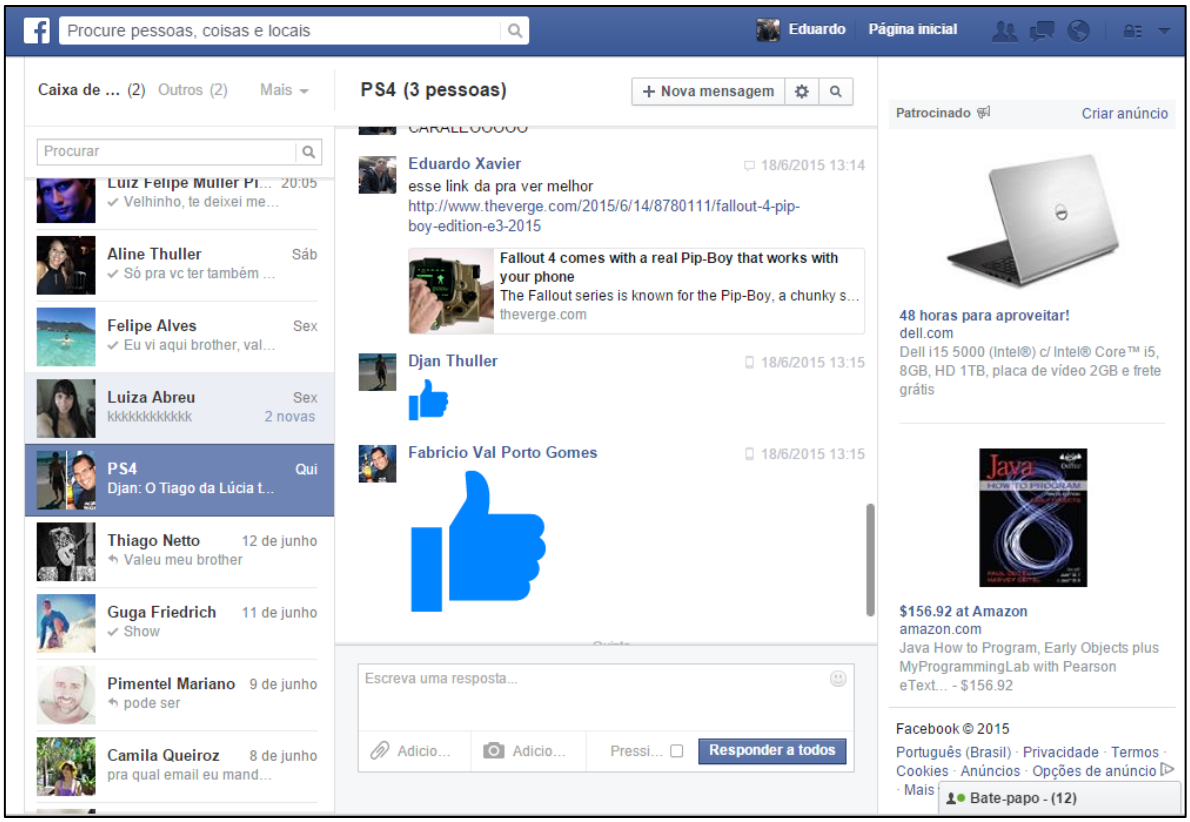


Figura 5 – Sistema de bate-papo em grupo do Facebook

2.2 Associações entre mensagens

Ao enviarem suas mensagens na sessão de bate-papo, os participantes criam um fluxo intenso de comunicações sequenciais. Compreender este fluxo de mensagens é muito difícil, e se agrava quando há diferentes tópicos sendo discutidos ao mesmo tempo (PIMENTEL & FUKS, 2009). Nesta seção discute-se a não-linearidade entre as mensagens do bate-papo, a construção de um texto não estruturado na intensa troca de mensagens e as estratégias que as pessoas usam para entender e inferir as associações entre mensagens.

2.2.1 Não linearidade entre mensagens do bate-papo

Diferente de textos tradicionais e contínuos, as mensagens publicadas nas sessões de bate-papo são organizadas por ordem de chegada, com isso, o discurso desdobrado nas mensagens não obedece à ordem visualmente linear do texto, resultando num texto intrincado (PIMENTEL & FUKS, 2009). Na conversação medida por computador, é

comum o discurso não ser estruturado linearmente, como ilustram as possibilidades de estrutura representadas na Figura 6.

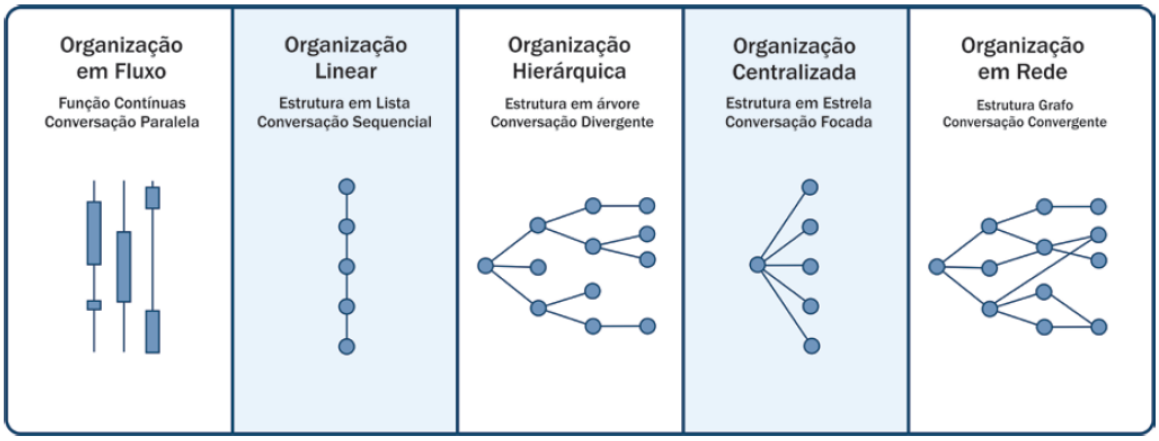


Figura 6 – Modelos de estruturação do discurso (Netto, 2014)

O Texto 1 é um exemplo com a transcrição das primeiras mensagens de uma sessão de bate-papo. Nesta transcrição, enumeramos as mensagens em função da ordem com que foi publicação na sessão. As setas entre as mensagens indicam uma associação de uma determinada mensagem com outra. Por exemplo, a mensagem 6 dá continuidade ao discurso iniciado na mensagem 3. A mensagem 7 foi publicada imediatamente após a mensagem 6, e seu conteúdo é uma resposta de Amadeus sobre a pergunta feita no final do discurso de Pinheiro na mensagem 6, sendo assim reconhecida a associação entre estas duas mensagens. Este é o caso de uma mensagem enviada em sequência que tem associação com a mensagem anterior. Já a mensagem 12 também é uma resposta para a mensagem 6 (estabelece uma associação), mas há várias mensagens entre elas, caracterizando a não-linearidade do discurso.

3	<Pinheiro>	vamos começar discutindo os aspectos positivos.
6	<Pinheiro>	O que vcs acharam que funcionou na disciplina e que deve ser mantido nur próxima edição????
7	<Amadeus>	Eu acho que a disciplina ganhou muito com as apresentações de aula por parte d alunos
8	<Edson>	Oi
9	<Edson>	Sou eu
10	<Edson>	cheguei
12	<Isa>	eu curti a parte dos feedbacks depois de cada apresentação
13	<Carol>	O que eu mais gostei foram as apresentações e os feedbacks de toda turma do Pinheiro
14	<Magno>	gostei muito do feedback que é dado no final de cada apresentação
15	<Vander>	Acho que as apresentaçõe são muitos boas nas aulas, deveria manter próxima edição.
16	<Amadeus>	sim, concordo @Carol
17	<Magno>	não vi isso em nenhuma outra disciplina e achei super válido
19	<Isa>	eu tb curti que o Pinheiro nao ficou preso soh em siscolab, ele deu orientacao sob varias outras coisas
20	<Edson>	não estou entendendo nada
21	<Leonardo>	apresentação e feedback
22	<Carol>	Gostei também do suporte dado na elaboração do trabalho pelo Pinheiro. Você n nos deixou "deixar de pensar" no trabalho
		...
58	<Edson>	acho que devemos proibir alunos especiais
59	<Magno>	uma vez que tinhamos que ler o livro para cada aula
60	<Linda>	Desisto
61	<Isa>	E cada aluno apresentou o conteúdo de um capítulo
62	<Linda>	Não consigo ler as msgs anteriores
63	<Edson>	Esses alunos especiais não colaboram
64	<Carol>	O que houve, linda?

Texto 1 – Exemplo de não linearidade em mensagens de bate papo

Primo (2007) argumenta que as ideias não ocorrem linearmente em cadeia, embora as mensagens sejam apresentadas de forma linear:

> (...) [há] distinção entre as palavras inglesas lineal e linear (ambas traduzidas em português para "linear"). O primeiro termo em matemática descreve uma relação entre variáveis cuja ligação entre suas coordenadas cartesianas resultam em uma linha reta. Já lineal refere-se a uma relação entre causas ou argumentos cuja evolução não retorna sobre si. Enquanto o oposto de linear é non-linear, lineal opõe a recursivo. Com isso em mente, Beteson nega a estrutura linear (*lineal structure*) das ideias na interação. (BATESON, 1980 *apud* PRIMO, 2007, p.107)

Na perspectiva de uma sessão de bate-papo o discurso produzido é não-linear, e as mensagens são tipicamente apresentadas em ordem cronológica de publicação.

2.2.2 **Estratégias humanas para inferir associações entre mensagens**

Para um participante entender o texto não-linear de uma sessão de bate-papo, ele precisa analisar as relações entre as mensagens e inferir associações entre elas. Para que esse processo possa ser realizado, é preciso que o usuário entenda se uma mensagem tem relação com outra enviada anteriormente. Ao buscar fazer esta associação, o participante da sessão desenvolve estratégias que buscam analisar e inferir as associações, tais como: análise de recência, coesão e coerência, sequências conversacionais, monólogos, assuntos e contexto (PIMENTEL & FUKS, 2009). Essas análises podem ser compreendidas das seguintes formas:

- **Análise da recência** – considera-se que uma mensagem tem mais probabilidade de estar relacionada com uma mensagem enviada recentemente, dificilmente a mensagem estará relacionada com alguma anterior, enviada há mais de cinco minutos. Normalmente as mensagens estão associadas a mensagens que foram postadas a não mais de 2 minutos atrás. Também considera-se a situação de mensagens que tem o tamanho de um parágrafo, é improvável que esta mensagem esteja associada a uma mensagem prévia, enviada a 10 segundos apenas. O padrão de recência considera que a relação de uma mensagem com outra anterior, tem o tempo de sua postagem – envio – variando de 10 segundos à 5 minutos.
- **Análise de coesão –** A constituição léxica da mensagem ajuda o leitor a identificar um possível relacionamento entre mensagens, então um humano analisa a estrutura léxica e gramatical das mensagens procurando semelhanças para então identificar uma associação.
- **Análise de coerência -** o participante se questiona se faz sentido uma mensagem ser resposta de outra mensagem observando sua consistência, relevância e elementos linguísticos.
- **Análise de sequências conversacionais –** considera-se que a relação entre mensagens pode ser identificada pelos pares de adjacência: uma pergunta leva a uma resposta, um convite leva a uma aceitação ou recusa, um cumprimento ou saudação leva a um cumprimento ou saudação etc.
- **Monólogos –** é comum um participante segmentar o discurso em algumas mensagens enviadas sucessivamente (uma espécie de monólogo), pois as pessoas

tendem a enviar duas ou mais mensagens curtas seguidamente para interagir mais rapidamente e não produzir uma única mensagem longa.

- A**nálise do assunto –** considera-se que mensagens relacionadas estejam discutindo um mesmo assunto.
- **Análise do contexto –** para inferir o relacionamento de mensagens, às vezes é preciso conhecer informações compartilhadas pelo grupo, mas que não estão no texto da sessão de bate-papo.

2.3 Problemas da conversação no bate-papo

A conversação que se realiza pelo bate-papo é informal, quase imediata e oferece o registro da conversa de forma não-linear. Assim, o sistema de bate-papo torna-se um ambiente antagonicamente agradável para conversar e confuso (HERRING, 2001). Nesta seção, é abordada a confusão do bate-papo, mais especificamente o problema da perda de co-texto e as pesquisas que já enfrentaram este problema.

2.3.1 A confusão no bate-papo

A confusão da conversação é menos percebida quando um grupo de usuários troca mensagens com fins de socialização (MORAIS, 2011). Porém, quando há necessidade de se acompanhar a conversação no bate-papo, por exemplo, quando o bate-papo é usado para promover um debate ou uma aula, a confusão torna-se perceptível. A partir do momento que se precisa entender o discurso, há a necessidade de se ler todas as mensagens, interpretá-las e relacionar umas com as outras, o que nem sempre é fácil de ser feito.

Hering (2001) descreve que a interação entre múltiplos interlocutores na conversa, pode gerar mensagens separadas em ordem linear da mensagem a qual ela está respondendo (ou se referindo), isso acontece por causa do surgimento de uma ou mais mensagens no mesmo momento de discussão. Numa comunicação em grupo, muitas mensagens interferem entre a mensagem de iniciação e sua resposta. Como as mensagens são apresentadas na ordem de recebimento pelo sistema, ocorre a interrupção do turno de adjacência, onde os interlocutores passam a ter dificuldade em memorizar a sequência de troca e interação de mensagens, deixando a sessão da conversação difícil de ser compreendida.

Smith e colaboradores (2000) apresentam um estudo sociológico da conversação que categoriza cinco grandes problemas relacionados aos sistemas típicos de bate-papo:

- **Falta de ligação entre as pessoas e o que elas dizem.** Não é óbvio identificar quais foram as mensagens enviadas por uma mesma pessoa na sessão de bate-papo, isso se dá em decorrência de como as mensagens são apresentadas nos sistemas de bate-papo.
- **Falta de visibilidade dos progressos da produção de turnos**. O processo de produção da mensagem é separado do processo da transmissão. Como os sistemas de bate-papo não são verdadeiramente síncronos, o turno da conversa demanda a produção completa do texto, e o receptor não percebe o processo de produção dos enunciados.
- **Falta de visibilidade do progresso-auditivo**. Os participantes do bate-papo não recebem informações momento a momento do que estão lendo, o que também aumenta a possibilidade de desentendimento.
- **Falta de controle sobre o posicionamento do turno.** Com a não-linearidade do log da sessão de bate-papo, as vezes o único jeito para entender o sentido de certas mensagens é rolar a janela para cima e ver as mensagens anteriores.
- **Falta de registro útil e contexto social.** Nos sistemas típicos de bate-papo, não se aproveita o conteúdo produzido quando a sessão termina. Não se desenvolve nenhum histórico social. Mesmo que o histórico seja mantido nos sistemas, a transcrição desse histórico é muitas vezes difícil de ser compreendida, esbarra-se em sequências conversacionais rompidas e confusas, o que torna o histórico confuso e ambíguo.

Morais (2011) enumera alguns fatores que podem promover a confusão no bate-papo: os próprios participantes (falta de memória, de conhecimento ou de interesse pelo assunto), o grupo (quantidade de participantes ou falta de coordenação), o sistema usado (interface ruim ou falta de organização), a conversação (quantidade de mensagens ou complexidade da conversa) e a não-linearidade das mensagens. Esses fatores combinados contribuem para o surgimento de problemas como a interrupção da dinâmica, sobrecarga de mensagens, descontextualização e, principalmente, a perda de co-texto:

> "A interrupção da dinâmica ocorre numa sessão de bate-papo, quando um usuário mediador está coordenando uma dinâmica e outros usuários entram com mensagens desnecessárias, que obstruem o fluxo da dinâmica. A sobrecarga de mensagem ocorre numa sessão de bate-papo quando um usuário não consegue ler todas as mensagens durante a conversa. Isso ocorre quando um elevado número de mensagens de todos os participantes é exibido em um curto espaço de tempo. A descontextualização ocorre quando um usuário entra no meio de uma sessão de bate-papo e os outros usuários já estão engajados na discussão, então, o usuário que entrou depois pode encontrar alguma dificuldade para conseguir entender a conversa. A perda de co-texto ocorre numa sessão de bate-papo quando um usuário não identifica qual mensagem anterior está sendo referenciada numa determinada mensagem mais recente, não conseguindo estabelecer o encadeamento da conversação". (MORAIS, 2011, pp. 15-16).

Nesta seção foram apresentados diversos problemas que ocorrem na conversação pelo bate-papo. O foco da presente dissertação está no problema da perda de co-texto, por isso este problema é aprofundado na subseção a seguir.

2.3.2 **Perda de co-texto**

Durante a sessão de bate-papo, cada participante deve analisar cada mensagem para inferir quem está falando com quem. Esse é um processo mental que acontece com base na análise do discurso e no fluxo de mensagens. Quando o participante não consegue identificar a relação da mensagem com alguma anterior, ocorre o fenômeno de perda do co-texto.

> O termo **co-texto** designa texto ao redor, o que está escrito antes ou após um enunciado e que fornece elementos para compreendê-lo. Difere-se de contexto que designa fatores externos ao texto, também necessários para a compreensão do texto.
>
> **Perda de co-texto** é o termo elaborado nesta pesquisa para designar o fenômeno que ocorre, numa sessão de bate-papo, quando o participante não consegue identificar a mensagem anterior que fornece elementos necessários para a compreensão de uma determinada mensagem. (PIMENTEL M. , 2002, p. 75).

Pimentel (2003) mostra alguns exemplos de perda de co-texto. No Texto 2, para compreender a mensagem 30 de Liane, é necessário identificar que ela estava contra argumentando a mensagem 26 anterior. Humberto não identificou esta associação e manifestou sua perda de co-texto na mensagem 31: "Contrário de que Liane, me perdi".

24	<Liane>	Directo, até onde eu sei é um software de autoria e não Groupware
26	< Pablo >	No meu entendimento software de autoria contribui para um groupware
30	< Liane >	Acredito que é o contrario, groupware pode ajudar no processo de autoria pois po facilitar o processo de comunicação entre os componentes da equipe
31	< Humberto :	Contrario de que Liane, me perdi

Texto 2 - Perda de co-texto manifestada na mensagem 31

148	<Liane>	Eu particularmente acho que ainda não consiguimos "alinhar as idéias"
163	< Humberto :	Respondendo a Liane lá no alto: Vai demorar muito até alinharmos as nossas ideias
166	< Liane >	Concordo...
167	<Marcelo>	com o que, Liane?

Texto 3 – Perda de co-texto manifestada na mensagem 167

No Texto 3, a mensagem 166 mostra que Liane concorda com o argumento apresentado na mensagem 163, no entanto, ela poderia estar concordando com diversas outras declarações anteriores também. Em seguida, Marcelo manifesta sua dificuldade para identificar com quem, especificamente, Liane estava se referenciando na mensagem 167, explicitando a sua perda de co-texto.

2.3.3 **Associação de mensagens nos sistemas de bate-papo**

Herring (2001) argumenta que a confusão do bate-papo tem origem nas limitações impostas pelos sistemas de conversação. Alguns pesquisadores propuseram soluções para associar mensagens visando diminuir o problema da confusão do bate-papo.

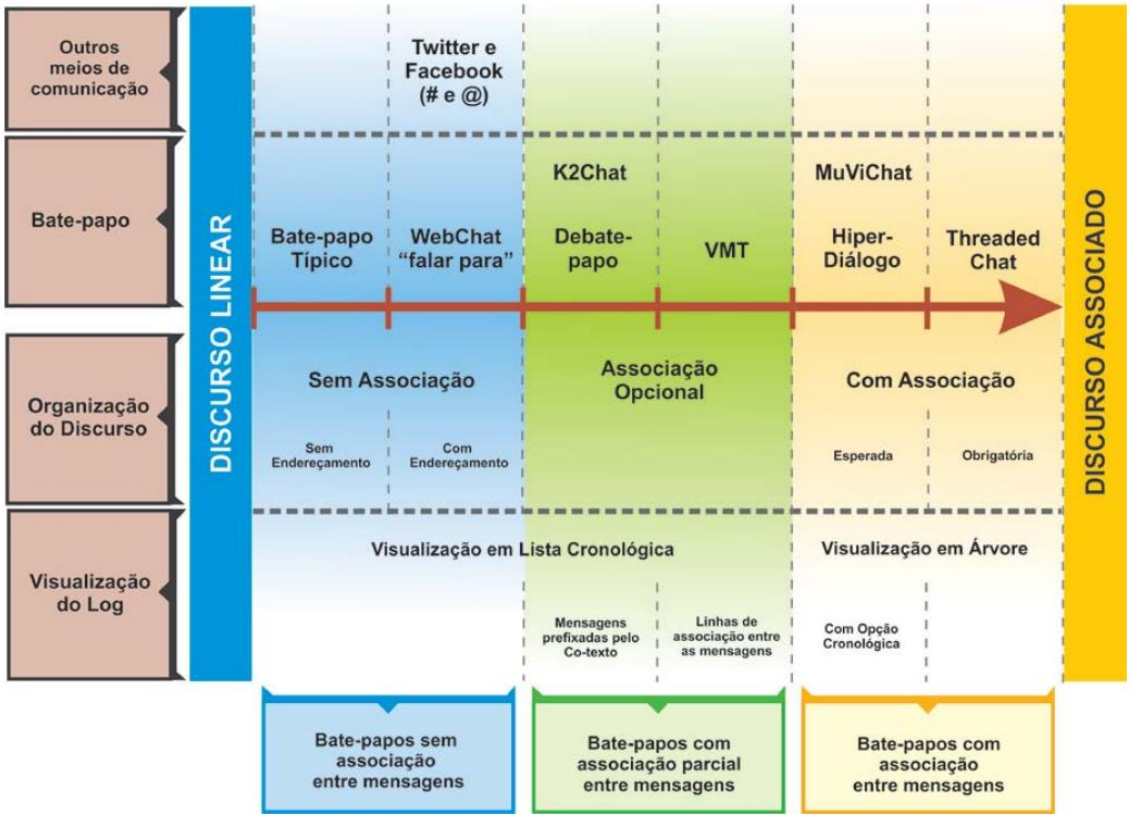


Figura 7 – Diferentes estruturas de associação do discurso em bate-papo (NETTO, 2014, p.63)

Netto (2014) identificou variações em sistemas de bate-papo quando se trata da organização linear das mensagens. As variações são decorrentes do sistema de estruturação do discurso que implementa o bate-papo. Conforme esquematizado na Figura 7, foram identificadas seis variações em função da estruturação do discurso, desde a ausência de associação, como nas implementações de bate-papo típico (discurso estruturado linearmente), até a associação obrigatória (discurso estruturado em árvore), como nas implementações dos sistemas *ThreadedChat* (SMITH, JJ, & BURKHALTER , 2000), Hiperdiálogo (PIMENTEL M. , 2002) e *MuViChat* (HOLMER, LUKOSCH, & KUNZ, 2009), que obrigam os usuários a estabelecerem o encadeamento em todas as mensagens que enviavam, gerando uma visualização das mensagens em formato de árvore.

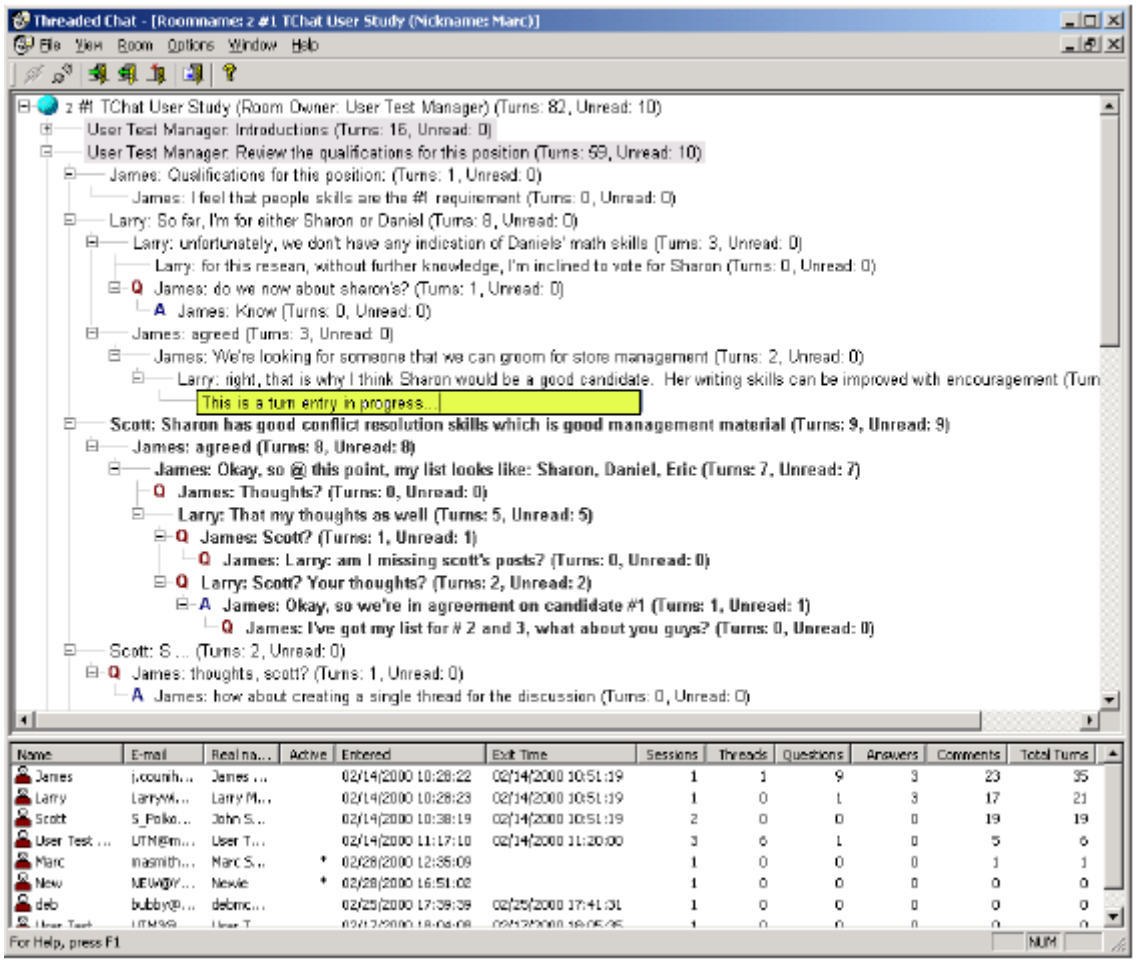


Figura 8 – Interface do ThreadedChat: exemplo de chat onde o usuário é obrigado a estabelecer o encadeamento durante o envio da mensagem

A Figura 8 ilustra a interface do *ThreadChat*, que propõe o encadeamento obrigatório entre mensagens. Nesse sistema, o usuário, para enviar uma mensagem, necessariamente precisa associar a nova mensagem com alguma anterior. Ao longo da sessão da conversa, o histórico do bate-papo fica com estrutura semelhante a de um fórum de discussão.

Já nos sistemas Entrevist@ (PESSOA, 2002), *ConcertChat* (MÜHLPFORDT & WESSNER, 2005), K2Chat (NUNES, UGULINO, GONCALVES, & SANTORO, 2008) e InterVIU (NUNES R. R., 2009) o estabelecimento da associação ocorre na própria lista cronológica das mensagens, sem apresentar visualização em árvore e com encadeamento opcional.

A Figura 9 ilustra a interface gráfica do Debate-papo (MORAIS, 2011), posteriormente aperfeiçoada resultando no Debate-papo v.2 (NETTO, 2014), que contém a estrutura básica de um sistema de bate-papo, mas disponibiliza um recurso onde o usuário pode responder diretamente as mensagens que foram recebidas. Quando o usuário

responde diretamente a uma mensagem, sua resposta é enviada aos demais participantes da conversa com o co-texto anexado, localizado na parte superior da mensagem. Este sistema possibilita a associação de mensagens de forma opcional, diferente do *ThreadedChat*.

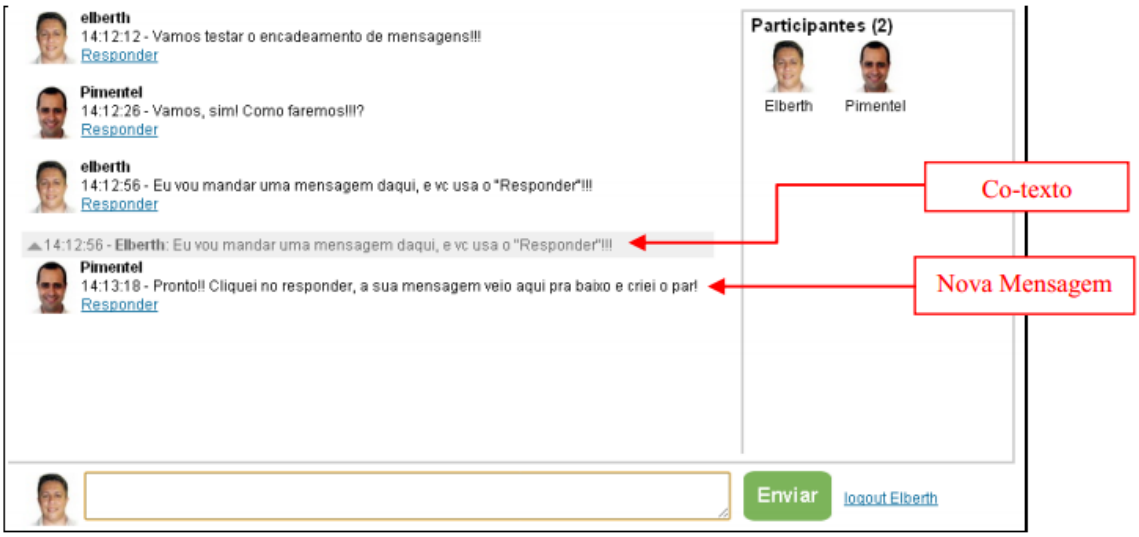
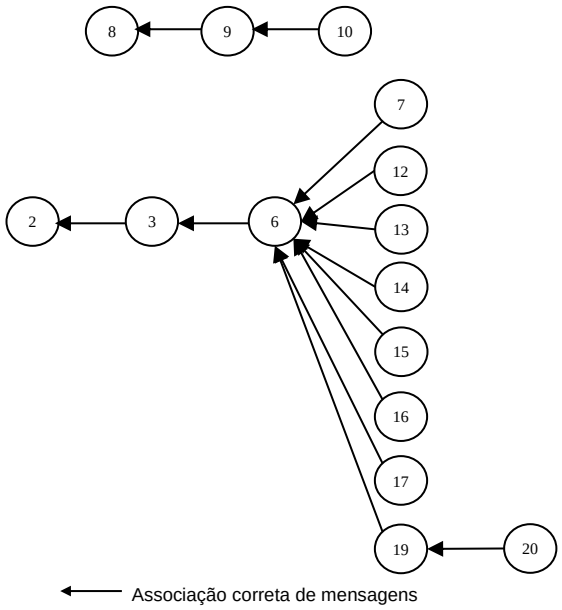


Figura 9 – Interface do Debate papo; Exemplo de chat onde o usuário é estabelece opcionalmente o encadeamento durante o envio da mensagem

2.4 Associação entre mensagens: modelagem em árvores

Nesta seção, é discutida a modelagem da conversação no bate-papo na estrutura em árvore (PIMENTEL M. , 2002). A título de ilustração, o Texto 1 encontra-se representado em árvore na Figura 10. Cada mensagem está identificada apenas pelo seu número-identificador no vértice, e as setas (arestas direcionadas) representam as associações entre as mensagens inferidas pelo pesquisador analista de discurso.



Representação gráfica da conversação no bate-papo transcrita no Texto 1

V = { 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 19, 20, 21}

A = { (2,3), (6,3), (7,6), (8,7), (9,6), (10,8), (12,6), (13,6), (14,6
(15,2), (16,13), (17,14), (19,6), (20,19), (21,6) }

V (Vértices) = conjunto de mensagens

A (Arestas) = conjunto de associações entre mensagens

Figura 10 – Grafo de representação do conjunto de mensagens associadas

Na modelagem da conversação do bate-papo, cada mensagem foi associada a uma única mensagem anterior, o que resulta num tipo particular de grafo denominado árvore. Grafos não só possibilitam uma forma conveniente para visualizar informações como também possibilitam o emprego de uma estrutura matemática para resolver problemas (EPP, 2011, p. 626). A mensagem que não possui associação com nenhuma anterior é denominada raiz e dá origem a uma nova árvore. Um conjunto de árvores é denominado floresta. A mensagem que não desencadeia relações posteriores com outras mensagens é denominada folha. Nessa pesquisa, nós acrescentamos outras denominações para facilitar a atividade de análise de dados. Quando há uma associação entre duas mensagens, a primeira é caracterizada como mensagem precedente e a segunda é chamada descendente. Na existência de duas ou mais mensagem descendentes, cada uma será caracterizada como irmã ou adjacente da outra. Na Tabela 1 – são listados os termos adotados nesta dissertação e na Figura 11 são aplicados estes termos na estrutura apresentada na Figura 10.

Tabela 1 – Conceitos da teoria de grafos aplicados a análise da conversa

Vértice	Uma mensagem
Aresta	Uma associação entre duas mensagens
Raiz	Mensagem que dá origem a uma árvore
Folha	Mensagem sem relações posteriores
Precedente	Mensagem que antecede a próxima de reciproca
Descendente	Mensagem posterior que forma o par estímulo-resposta
Irmã/adjacente	Ênupla de mensagens descendentes - não singular.
Árvore	Conjunto de mensagens associadas
Floresta	Conjunto de árvores

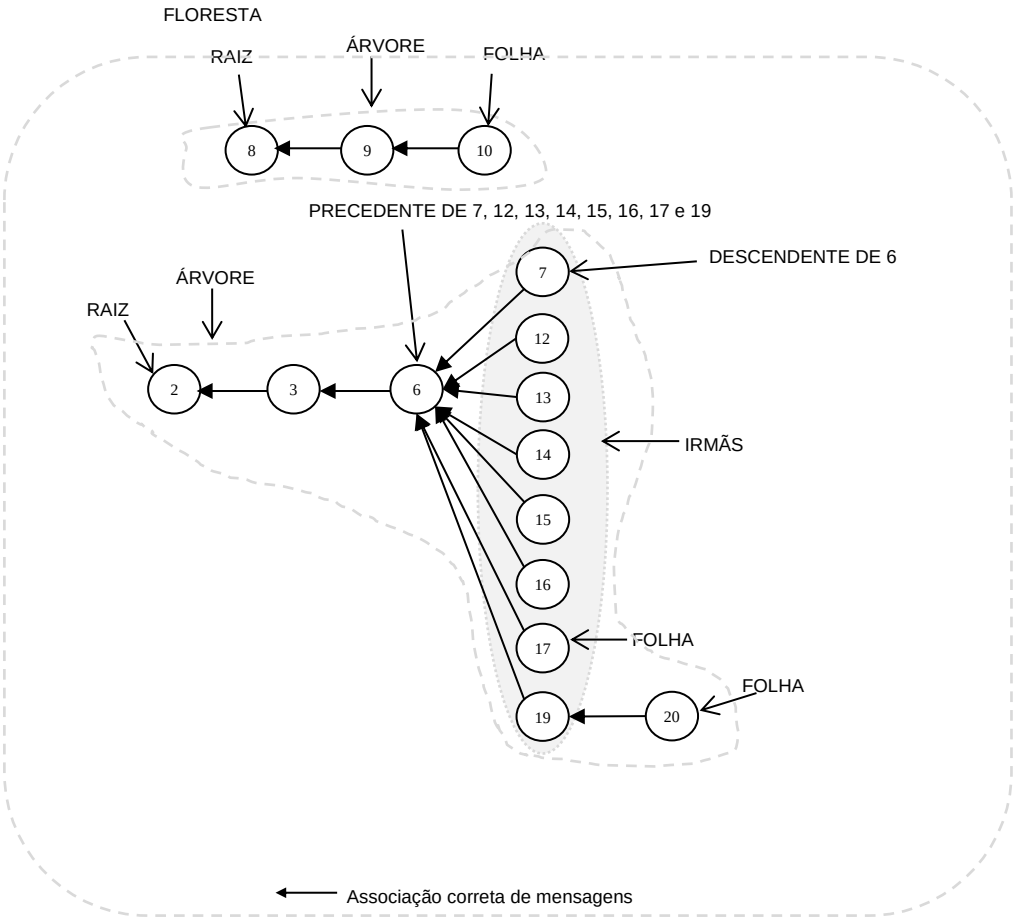


Figura 11 – Gráfico que caracteriza os tipos de algumas mensagens

Na Figura 11 é apresentada uma floresta com duas árvores. A mensagem 2 e 8 são as raízes das árvores. as mensagens 20, 10 e 7 foram marcadas na figura para ilustrar o conceito de folhas, no entanto, as mensagens 16, 17, 12 e 15 também são folhas. Na mensagem 6, Pinheiro pergunta “O que vocês acharam que funcionou na disciplina e o que deve ser mantido na próxima edição?”. Quando os alunos respondem esta mensagem, ela

passa a se caracterizar como uma mensagem precedente de respostas 12, 13, 14, 15 e 19, já estas mensagens-respostas são caracterizadas como descendentes da mensagem 6. As mensagens descendentes de uma mesma mensagem são denominadas irmãs uma das outras.

2.5 **Técnicas algorítmicas para inferir associação entre mensagens**

O objetivo desta seção é apresentar as técnicas computacionais que foram usadas nesta pesquisa para inferir as associações entre mensagens de bate-papo.

2.5.1 **Mineração de textos**

Os avanços em hardware e software em plataformas de redes sociais promoveram a rápida criação de grandes volumes de dados diferentes, grandes volumes de conteúdo gerado em formato texto produzido por pessoas. Os sistemas de bate-papo entram nesse contexto porque são sistemas onde os humanos geram textos através das sessões de conversação. Para analisar essas sessões, suas mensagens e as associações entre as mensagens da sessão, são adotadas estratégias da área de mineração de textos e processamento natural de linguagem.

A área de estudo de mineração de textos providencia técnicas com base em construção de modelos e algoritmos de aprendizado (ZHAI, 2012), no entanto, a atividade de minerar textos é complexa. Feldman e Sanger (2006) argumentam que a maioria dos sistemas de mineração de textos objetiva a descobertas de padrões em coleções de documentos, correlacionando e mapeando relacionamentos complexos ou mesmo identificando tendências que podem parecer impossíveis de descobrir. A Figura 12, mostra a arquitetura geral de um sistema de mineração de textos onde há o recebimento de documentos (*input*) e produção de padrões, mapas de conexões e tendências (*output*).

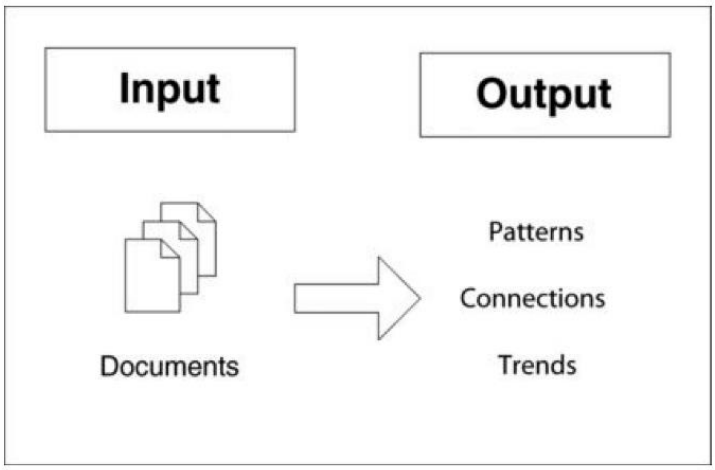


Figura 12 – Arquitetura geral de sistemas de mineração de texto (Feldman e Sanger, 2006)

Em mineração de textos, um documento é a menor unidade discreta básica de dados de textos. Não necessariamente é um arquivo de relatório, e-mail, artigo, etc. Um documento é definido como menor unidade dentro de um contexto particular e que pertencente a uma coleção de documentos. Por exemplo, o histórico de uma sessão de bate-papo contém várias mensagens, onde cada mensagem pode ser considerada como um documento. Como essa mensagem pertence à sessão de conversação, ela então pertence a uma coleção de outros documentos – que são mensagens.

No entanto, ainda na visão dos autores, os algoritmos de descoberta de conhecimento não reconhecem textos escritos por humanos em sua linguagem natural. Estes textos são considerados como documentos em formato desestruturado, mesmo que contenha uma estrutura linguística que por sua vez emprega semântica, estrutura sintática, elementos tipográficos (como pontuação, numerais, caracteres especiais, etc).

A estrutura linguística não é suficiente para que os algoritmos de descoberta de conhecimento trabalhem e é preciso transformar os elementos contidos na linguagem natural em elementos com estrutura de representação explícita, passível de entendimento pelo algoritmo. Para Miner e colegas (2012) a forma mais popular de transformação e representação de um texto é a criação de um modelo de representação em espaço vetorial. Esse modelo pode ser construído a partir de características como, por exemplo, a quantidade de ocorrências de uma mesma palavra num documento. Feldman e Sanger (2006) aprofundam o argumento mostrando que pode haver uma transformação que represente e identifique um subconjunto de outras características para representar o documento num modelo. As características mais comuns são:

- **Caracteres** - componente individual, exemplos: letra, numeral, caractere especial. Característica mais fundamental que pertencente ao alto nível semântico: palavras, termos e conceitos.
- **Palavras** - nível básico de riqueza semântica, exemplos; uma palavra de apenas um *token* linguístico. Já frases, expressões de palavras compostas não constituem uma característica de nível de palavra.
- **Termos** - podem ser apenas uma palavra ou composição de palavras selecionadas diretamente de um corpus de um documento nativo. Exemplos: "Presidente Lula" e "Rainha Elizabeth".
- **Conceitos** - característica gerada para um documento de forma manual, estatística, baseada em regras ou com metodologias híbridas de categorização. Exemplos: uma palavra ou expressões com múltiplas palavras. Melhor nível de características para lidar como sinônimos ou polissemia, hipônimos[3] e hiperônimos[4].

Funcionalmente a arquitetura de um sistema minerador de textos emprega quatro operações clássicas:

1. **Pré-processamento**, que inclui todas as rotinas, processos e métodos necessários para preparar e estruturar o documento, convertendo informações do formato original num formato entendido;
2. **Operação central de mineração**, em que se realiza a descoberta de padrões, melhora o conhecimento gerado etc.;
3. **Camada de apresentação**, onde se gera um artefato visual para facilitar a análise dos resultados;
4. **Refinamento**, que inclui método para filtrar dados redundantes, fazer agrupamentos (clusterização), etc.

As operações 1 e 2 são as mais importantes por se desdobrarem numa de processos dentro da visão generalizada de uma arquitetura de um sistema de mineração de textos.

[3] Hipônimos são palavras de sentido cujos significados hierarquicamente mais específicos do que de outras. Exemplo: Maçã e morango são hipônimos de fruta. (PEREZ, 2015)
[4] Hiperônimos são palavras de sentido genérico, cujos significados são mais abrangentes do que os hipônimos. Exemplo: Animais é hiperônimo de cachorro e cavalo. (PEREZ, 2015)

2.5.2 **Processamento de Linguagem Natural**

Processamento Linguagem Natural (NLP, em inglês) é um campo de estudo onde as técnicas linguísticas são usadas para realizar análise fonética, morfológica, sintática e semântica do texto (JURAFSKY & MARTIN, 2000, pp. 2-4). São propostas técnicas voltadas para tratar o grande volume de dados em formato texto gerados pelos humanos em linguagem natural (dos humanos).

A área da mineração de texto abrange diferentes campos de pesquisa, incluindo a NLP, a Recuperação de Informação, a Mineração da WEB, entre outras (JIANG, 2012). No entanto, a etapa de pré-processamento é a parte mais importante e comum entre elas, pois diferentes técnicas podem ser aplicadas, dependendo do objetivo de estruturar dados não estruturados ou semiestruturados.

As técnicas de NLP em geral são usadas em textos longos, mas algumas pesquisas atuam em textos curtos, com sentenças que contém entre 10 e 20 palavras, sem necessariamente ter sentenças gramaticais completas.

Aplicações de textos curtos tem proximidade com a associação e encadeamento de mensagens devido a estrutura reduzido do texto, no entanto, este campo de pesquisa possui alguns desafios para gerar similaridades entre textos pequenos. OLIVA e colaboradores (2011) apontam que um dos maiores problemas dos métodos para gerar similaridade entre textos curtos é que eles ainda são adaptações das técnicas usadas em textos longos.

FURLAN *et al* (2013) apresentam duas maneiras de determinar a similaridade de textos curtos. A primeira, usa topologia estatística para gerar similaridade a cada palavra e a segunda usa modelos que contém informações sobre conceitos e suas interconexões. Próximo desta abordagem está o trabalho de Lima (2013), que apresentou um modelo de análise de diálogo onde se aplicam as técnicas de pré-processamento em conjunto com o modelo específico dentro do domínio de associação entre mensagens de bate-papo. O algoritmo proposto por Lima é objeto de investigação no Capítulo 5 da presente dissertação.

3 Análise humana das associações entre mensagens de bate papo

Para investigar como as pessoas associam as mensagens e as dificuldades que enfrentam para estabelecer as associações, foi feito o estudo de caso documentado neste capítulo. Para organizar o estudo, foram seguidos os passos esquematizados na Figura 13 que caracterizam um processo linear e iterativo de pesquisa (RECKER, 2013).

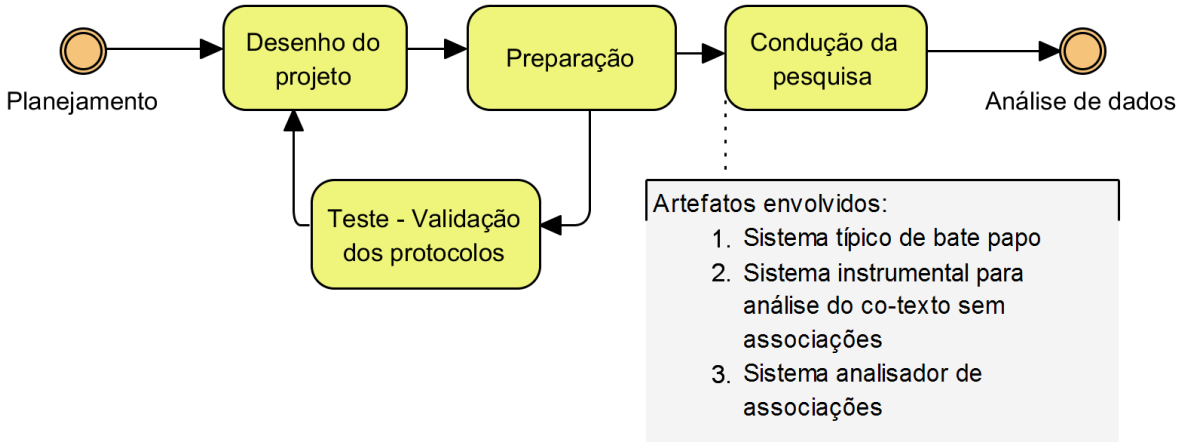


Figura 13 – Processo de pesquisa usado na condução do caso de uso

Na Seção 3.1, são apresentados o planejamento e as questões de estudo. Na Seção 3.2, são descritas as unidades de análise, os protocolos e artefatos desenvolvidos para conduzir a pesquisa com qualidade. Na Seção 3.3, é discutido como foram testados os protocolos e o planejamento. Na Seção 3.4, é relatado como o estudo foi conduzido. Por fim, na Seção 3.5, são apresentados os resultados e as análises dos dados produzidos.

3.1 Objetivo e questões do estudo

O estudo de caso (YIN, 2009) apresentado neste capítulo, teve por objetivo investigar, de maneira exploratória, a capacidade que o ser humano tem em inferir associações de mensagens de bate-papo: qual o grau de dificuldade que os humanos têm em inferir as associações de mensagens de bate-papo? É necessário responder esta questão para ser possível relativizar a performance de um algoritmo ao inferir, automaticamente, as

associações entre as mensagens de bate-papo. Digamos, por exemplo, que um algoritmo acerte 75% das associações, essa seria uma performance boa ou ruim? Esta resposta só pode ser dada quando fazemos uma comparação com outro algoritmo, ou, como aqui proposto, com seres humanos. Se as pessoas, em média, acertam 95% das inferências, e um algoritmo acerta apenas 75% é possível dizer que a performance é ruim, que ainda pode ser melhorada. Mas digamos que as pessoas, em média, acertem apenas 50% das inferências, então concluiríamos que o algoritmo apresenta desempenho extraordinário.

Para investigar o grau de dificuldade que as pessoas têm ao inferir as associações entre as mensagens de bate-papo (objetivo do estudo), buscou-se respostas para as seguintes questões de pesquisa:

a. Com que frequência as pessoas erram as inferências que fazem?
b. As pessoas sempre fazem uma inferência, ainda que errada, ou assumem a incapacidade de deduzir uma possível associação entre as mensagens?
c. As pessoas demoram muito tempo para fazer uma inferência?

A pesquisa exploratória sobre cada uma destas questões ajuda a melhorar a compreensão sobre a confusão da conversação, decorrente da dificuldade para se inferir uma associação entre mensagens e, eventualmente, a consequente perda de co-texto. Portanto, o presente estudo, além de ter o objetivo de definir um parâmetro para que se possa avaliar o desempenho de algoritmos propostos para inferir associações entre mensagens de bate-papo, também tem por objetivo conseguir a definição de melhores métricas para que seja possível evidenciar, medir e avaliar a confusão e a perda de co-texto no bate-papo, o que tem sido perseguido por nosso grupo de pesquisa desde o trabalho seminal nesta linha, pois:

> Não se deve concluir, a partir destes dados, que a perda de co-texto é um fenômeno esporádico. A freqüência aqui obtida parece indicar apenas ‘a ponta de um iceberg’ – a perda de co-texto é um fenômeno cognitivo e a manifestação textual é apenas uma medida indireta deste fenômeno; nem toda perda de co-texto é manifestada textualmente. (PIMENTEL, 2003, p.97)

Métricas e parâmetros, como o percentual de erros de inferências e o tempo para realizar as inferências, podem medir e caracterizar melhor o fenômeno da confusão no bate-papo e da perda de co-texto.

3.2 **Desenho do projeto**

Para ser possível alcançar o objetivo desta pesquisa, que é investigar o processo humano de inferir associações entre mensagens de bate-papo, foi projetado um estudo em que uma turma participa de uma sessão de bate-papo e depois cada um infere as associações entre as mensagens daquela sessão. Desta forma, o "gabarito das associações" é definido pelo próprio grupo (não é dependente da análise de um analista do discurso experiente, pois ainda que experiente, é sujeito a erros), uma vez que todos inferem as associações de todas as mensagens, cada participante indica a que mensagem estava respondendo em cada mensagem que ele(a) próprio(a) enviou, e, assim, o participante produz o gabarito de suas próprias mensagens. Ao se juntar o gabarito de todos os participantes consegue-se o de toda a sessão de bate-papo. Este modelo é fundamental para que seja possível classificar as demais inferências como certas ou erradas, o que possibilita avaliar o grau de acerto ou erro das pessoas ao entenderem associações. O tempo que cada pessoa demora para inferir a associação de cada mensagem é outra variável importante neste estudo, e precisa ser registrada. Desta forma, foi projetado um estudo de caso exploratório[5] com produção de dados quantitativos baseados em acertos e tempo. As unidades de análises são os participantes da sessão de bate-papo, que ocorre no contexto de uma turma real. Portanto, o projeto se caracteriza como um estudo de caso múltiplo (YIN, 2010), como esquematizado na Figura 14.

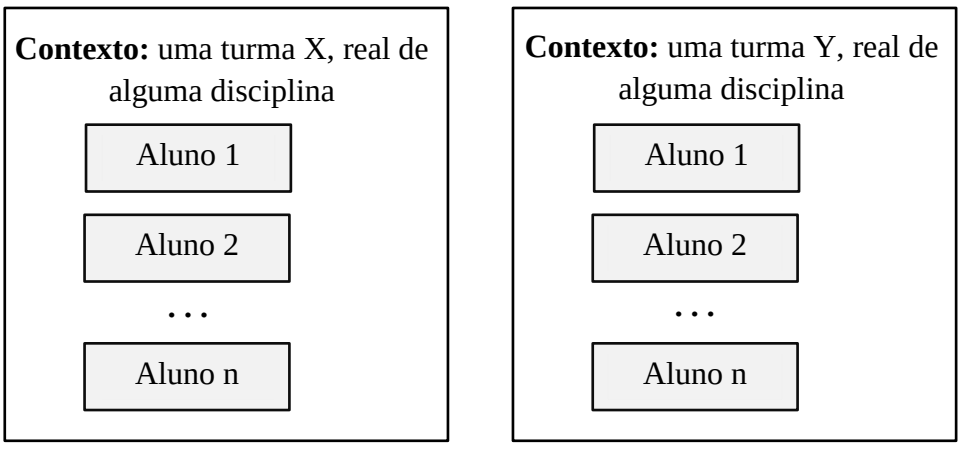


Figura 14 – Esquema do estudo com múltipla unidade de análises

[5] Pesquisas de sistemas de informação podem ser classificadas como interpretativas quando elas assumem que nosso conhecimento da realidade é obtido através da construção social como a língua, consciência, sistemas e outros artefatos. Produzir conhecimento do contexto do sistema de informação é um processo onde o sistema de informação influencia e é influenciado pelo seu contexto (KLEIN e MAYERS, 1999)

3.2.1 **Procedimentos do estudo**

Para a realização do estudo de caso foi definido o processo ilustrado na Figura 12, organizado em três etapas: planejamento do debate da turma, condução do debate, análise da conversação e produção do banco de dados.

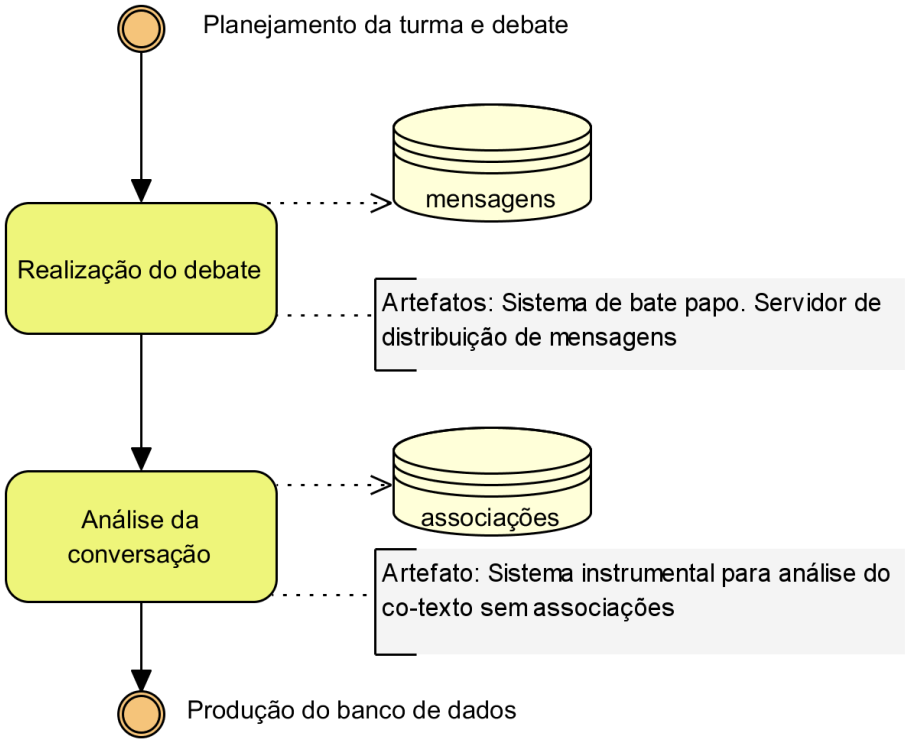


Figura 15 – Procedimentos definidos do projeto deste caso de uso

No primeiro procedimento, **planejamento do debate na turma**, é preciso definir o grupo, o assunto a ser debatido numa sessão de bate-papo e o que produzirá o log a ser posteriormente analisado pelos próprios participantes. A turma precisa ser real, não pode ser uma escolha aleatória de pessoas, mas sim alunos de estejam realmente cursando uma disciplina, que se conheçam e tenham real interesse em discutir o assunto da sessão – isto porque no Estudo de Caso "*o enfoque está sobre um fenômeno contemporâneo no contexto da vida real*" (YIN, 2010, p.22). Os alunos também precisam assinar um termo de consentimento (Apêndice I). Deve ser selecionado um sistema que implemente a interface típica de bate-papo, sem possibilitar qualquer tipo de associação entre mensagens.

O segundo procedimento, **condução do debate**, deve garantir que os alunos só se comuniquem pelo bate-papo durante a sessão, sem se comunicar por outros meios virtuais

ou reais. O pesquisador não pode participar nem influenciar o debate, apenas orientar o professor da turma sobre os procedimentos a serem seguidos.

No terceiro procedimento, **análise da conversação**, os próprios alunos participantes da sessão devem analisar o histórico das mensagens, inferindo a associação entre elas. Além das associações estabelecidas por cada um, deve-se registrar também o tempo que o sujeito demorou para inferir a associação em cada mensagem. Estes dados são armazenados num banco de dados para posterior análise.

3.2.2 **Artefatos usados nos procedimentos**

Para viabilizar este estudo, foram desenvolvidos os sistemas apresentados a seguir. Para manter a uniformidade entre o sistema de bate-papo e o sistema para as pessoas registrarem suas inferências das associações entre as mensagens, optou-se pelo desenvolvimento de ambos os sistemas. Um terceiro modelo também foi desenvolvido para automatizar as análises dos dados produzidos.

3.2.2.1 Sistema típico de bate papo

Foi desenvolvido um sistema de bate-papo típico seguindo a arquitetura de comunicação estabelecida dentro dos padrões do protocolo XMPP [6]. A adoção deste protocolo abriu muitas possibilidades para apoiar a pesquisa em cima de outros projetos *open source* e garantir o sucesso da condução da pesquisa. Por exemplo, não foi preciso desenvolver um servidor de distribuição de mensagens. Na presente data de publicação desta pesquisa é possível encontrar os seguintes servidores que implementam o protocolo XMPP: *ejabberd*, *Jabberd*, *OpenIM*, *WPJabber*, *iChat Server*, *Metronome*, *Isode M-Link*, *MongooseIM* e o *OpenFire*.

Para melhor atender essa pesquisa foi adotado o servidor *Openfire*[7]. A Figura 16 mostra sua interface administrativa com um menu para administração o servidor, seus usuários e grupos, as sessões de conversação e administração as salas. A seção “Salas de

[6] XMPP é um protocolo aberto para comunicação real-time que emprega recursos de mensageiro instantâneo, presença, conversa em grupo, voz e chamadas de vídeo, *middleware leve*, distribuição de conteúdo e roteamento de dados em formato XML. Extraído em http://xmpp.org/about-xmpp/ em 16/07/2015

[7] O *Openfire* é um servidor de colaboração em tempo real, licenciado pela *Open Source Apache License*. Ele foi construído com base no protocolo aberto para mensageiros instantâneos, XMPP). Extraído de http://www.igniterealtime.org/ em 16/07/2015

Conferência" lista todas as salas disponíveis e informações respectivas a cada uma das salas como quantidade de usuários, possibilidade de edição, gravação das mensagens em banco de dados e possibilidade de apagara sala.

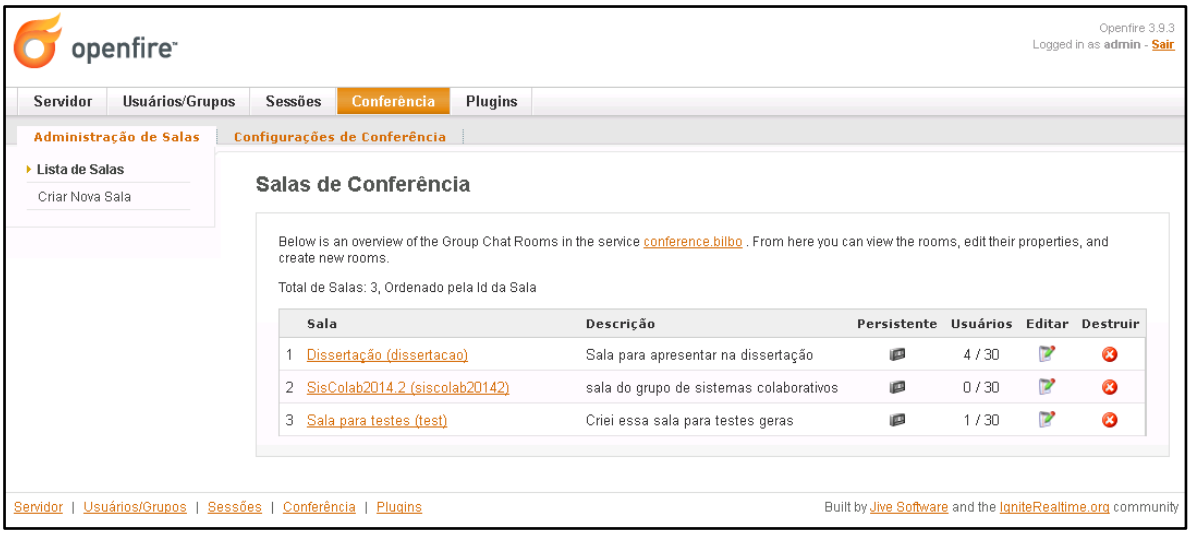


Figura 16 – Interface gráfica do sistema Open Fire

A interface do sistema típico de bate-papo foi desenvolvida para se comunicar com o servidor de mensageria usando as linguagens HTML, CSS e JavasScript. Na Figura 17, ilustra o sistema típico mostrando a lista de usuários online à esquerda, e ao centro, uma área de conversação onde é exibida a lista de mensagens enviadas pelos participantes.

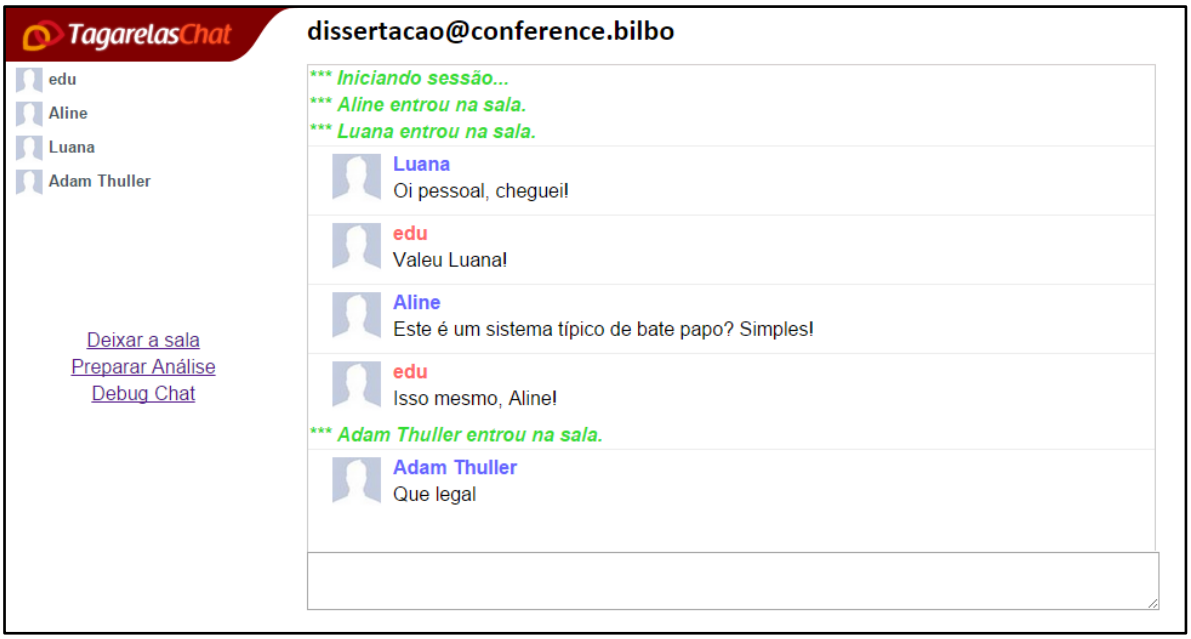


Figura 17 – Interface gráfica do sistema típico de bate papo implementado na pesquisa

No entanto, para o sistema enviar e receber mensagens com apoio do servidor, ao rigor do protocolo XMPP, foi necessário utilizar a biblioteca *Strophe.js*[8]. A Figura 18 ilustra a arquitetura de comunicação apresentando o sistema típico de bate papo e o servidor *OpenFire* como componentes da comunicação. O *Strophe.js* é usado como uma interface entre os dois componentes.

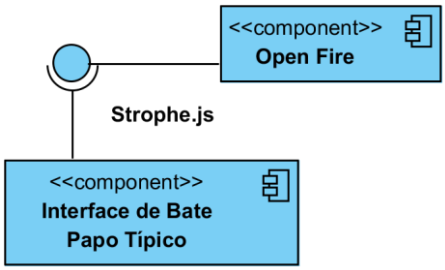


Figura 18 – Diagrama de componentes do sistema típico de bate-papo

Combinando o *OpenFire* e o *Strophe.Js* no projeto de arquitetura de comunicação foi possível construir arquitetura robusta e proporcionar uma boa experiência para os participantes na atividade de conversar no sistema típico de bate-papo.

3.2.2.2 Sistema instrumental para análise do co-texto sem associação

O objetivo deste sistema é ajudar o usuário na atividade de associar mensagens e na produção de um banco de dados de mensagens associadas por cada usuário. Ele apresenta, aos poucos, o histórico de mensagens ao usuário (mensagem a mensagem). Nesse modelo, a ação transfere ao usuário a sensação de que ele está numa sessão de bate-papo.

Quando uma mensagem do log é exibida, o usuário pode identificar o co-texto escolhendo uma das mensagens anteriores presentes na área de visualização das mensagens. No entanto, se o usuário percebe que a mensagem não tem co-texto, ele pode usar o botão “sem associação”. Quando o usuário não encontra o co-texto da mensagem ou não sabe dizer se a mensagem tem ou não um co-texto, ele pode usar a opção “Desisto”.

[8] O *Strophe.js* é uma biblioteca desenvolvida em linguagem Javascript para facilitar o desenvolvimento de sistemas de internet baseados em XMPP. Extraído de http://strophe.im/strophejs/ em 16/07/2015

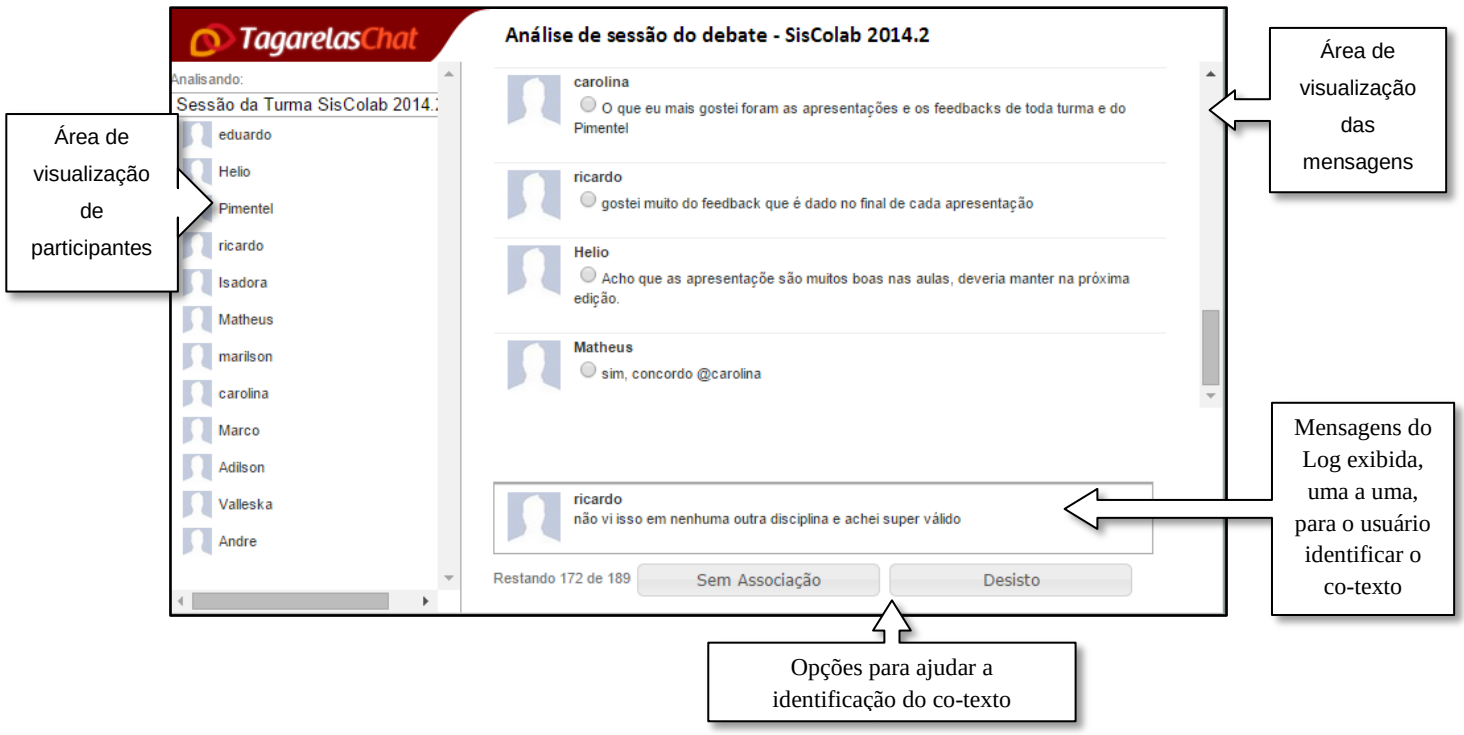


Figura 19 – Evolução do sistema instrumental para análise do co-texto sem associação Netto (2014, p.71)

Este sistema foi desenvolvido com base no sistema instrumental para análise do co-texto de Netto (2014, p.32). Na versão implementada na presente pesquisa, o recurso para associar mensagens foi alterado: As "Mensagens do Log" são exibidas uma a uma como se tivessem sendo digitadas pelo usuário. No entanto, a próxima mensagem exibida surge imediatamente após o estabelecimento da associação entre a mensagem exibida e uma mensagem existente na área de visualização das mensagens. O estabelecimento da associação ocorre quando o usuário indica clica com o mouse em cima de uma das mensagens que é mostrada na área de visualização. Assim, ele deixa explícita sua inferência, estabelecendo a associação entra a nova mensagem e uma mensagem existente.
O usuário pode usar o botão “Sem associação” para indicar que a mensagem exibida não tem associação com nenhuma outra mensagem presente na área de visualização. Ele também pode sinalizar desistência usando o botão “Desisto“ para deixar explícito que ele não conseguiu inferir uma associação.

3.2.2.3 Sistema analisador de associações

O sistema analisador de associações instrumentaliza a proposição do estudo, por organizar os dados relevantes e permitir a análise dos resultados. Para descrever o sistema, vamos ilustrar os componentes com a Figura 20, onde nós temos como componente principal o analisador de associações. Ele é responsável por ler o formato de dados gerados pelo Sistema instrumental para análise do co-texto sem associação. Ele também é encarregado de gerar os arquivos resultantes da análise.

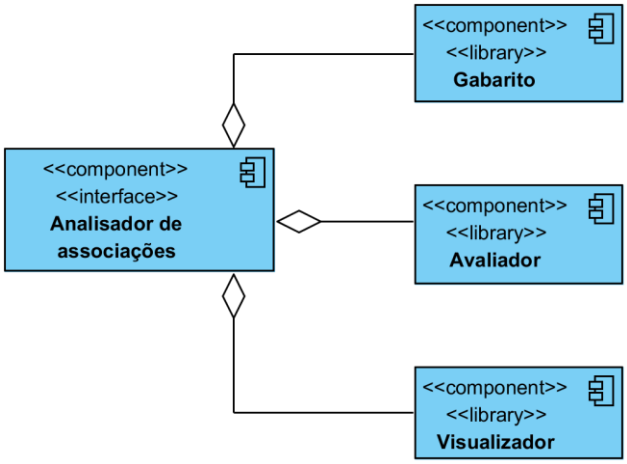


Figura 20 – Componentes do analisador de associações

O componente "**gabarito**" é responsável por extrair o padrão das associações entre as mensagens. Quando o usuário infere uma associação de uma mensagem ao qual ele é o remetente, esta ligação é considerada a correta e fica registrada no gabarito da sessão. Ao se analisar as inferências de todos os participantes do log da sessão, consegue-se extrair o gabarito das associações entre as mensagens da sessão do bate-papo.

O componente "**avaliador**" usa o gabarito para analisar as associações inferidas por cada participante, identificando os erros e acertos individuais e do grupo. Nesta análise, são desconsideradas as associações que o usuário indica nas mensagens em que ele mesmo é o emissor, pois estas associações foram consideradas "o gabarito", isto é, o usuário não está realizando uma inferência (sujeita a erro ou acerto), mas sim uma indicação de qual é, de fato, a mensagem associada com aquela que ele havia enviado.

O componente "**visualizador**" apresenta o resultado de erros e acertos de associações de cada participante e de cada mensagem. Este componente também é responsável em gerar em arquivos contendo dados do resultado por participante, o

gabarito, resultado das apurações de erros por mensagens e usuários e gerar um grafo do gabarito.

A Figura 21 apresenta o “resultado geral da análise” que diz a quantidade os erros e acertos de associações do grupo inteiro de participantes. “O resultado individual de cada participante” exibe a lista de todos os participantes do grupo, exibindo a apuração de erros e acertos, já desconsiderando as associações indicadas nas mensagens emitidas pelo próprio usuário.

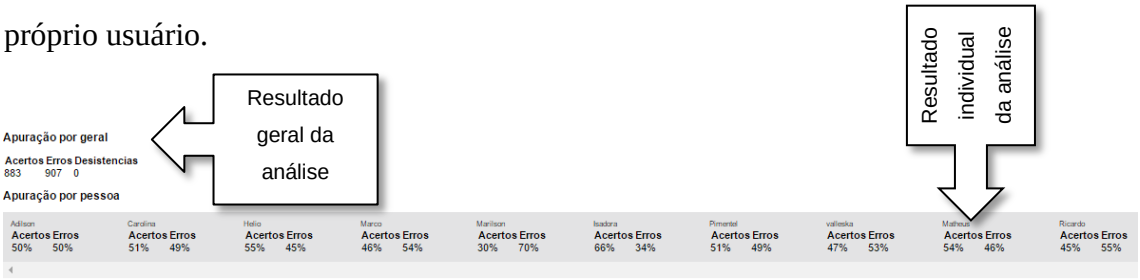


Figura 21 – Apuração geral e por participante, realizada pelo sistema “analisador de associações”

O gabarito do das associações também pode ser extraído em um arquivo ou visualizado num grafo, como ilustrado na Figura 22. Essa visualização permite aumentar ou diminuir a profundidade da visualização.

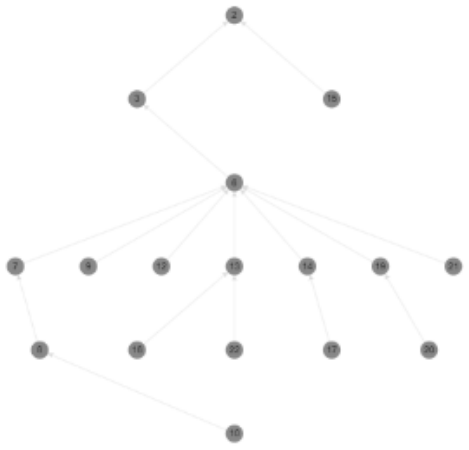


Figura 22 – Gabarito representado como grafo

A analisa o resultado por mensagem, mostrado o desempenho de cada participantes em tentar associá-la com outra mensagem anterior. Nesta figura, é possível visualizar o resultado da mensagem 15. Os participantes Vanessa e Helmo acertaram a associação. Já Marildo, errou ao indicar que a mensagem 15 estava associada com a 8.

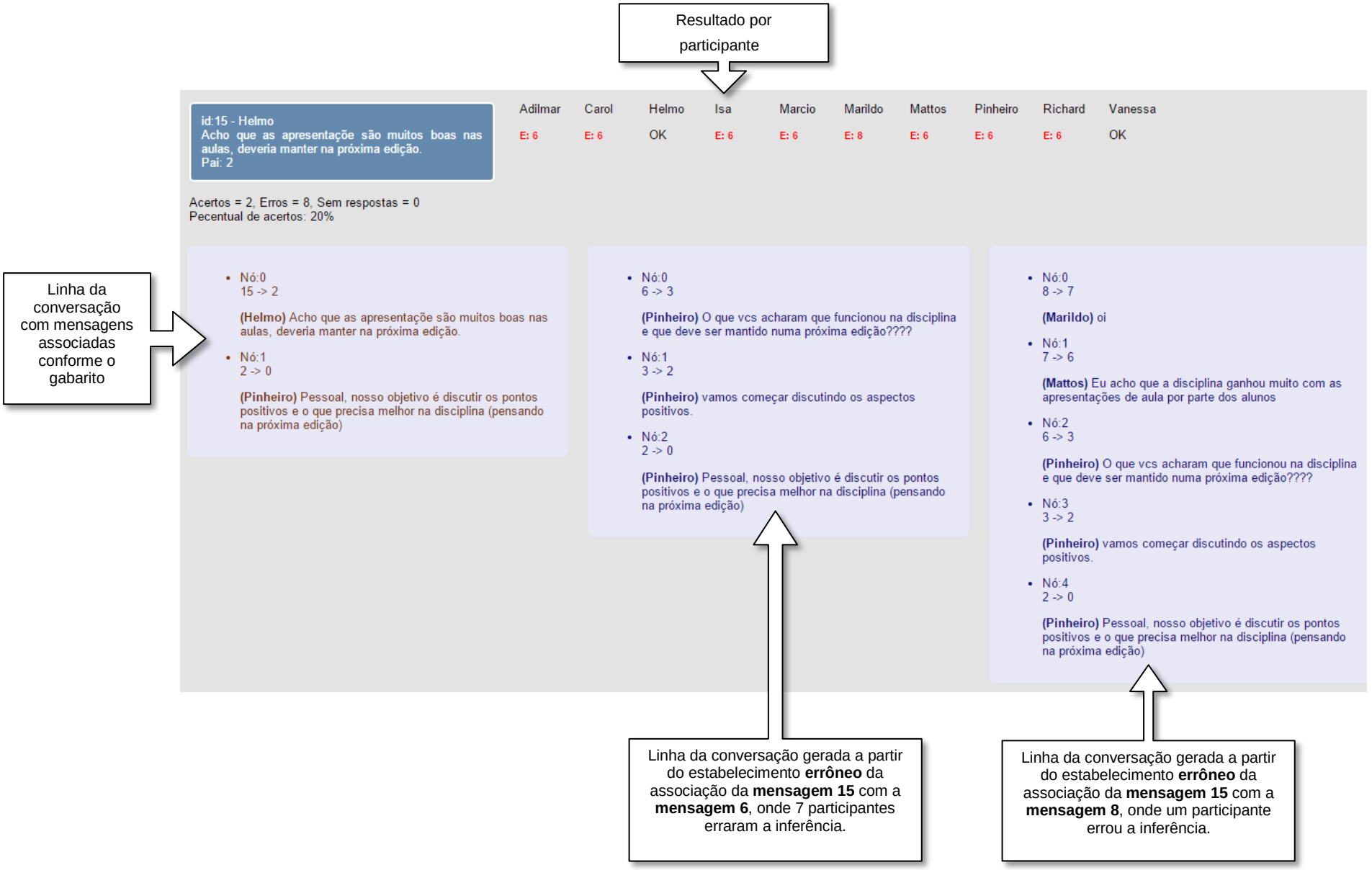


Figura 23 – Fragmento do módulo de visualização que mostra a apuração de erros e acertos por mensagem

Todas as mensagens são analisadas mas nessa figura somente a mensagem 15 é exibida. A figura mostra o “Resultado por participante“, onde mostra-se a apuração de erros e acertos dos participantes em relação a mensagem. Nesta área, os resultados individuais de participante são exibidos mostrando “OK”, quando o participante acertou ou quando o participante erra, exibindo o texto “E: X”, onde X é o número da mensagem que o participante associou erroneamente.

Ainada analisando a figura, a “Linha da conversação com mensagens associadas conforme o gabarito” mostra a uma cadeia de mensagens empilhadas, exibindo o conteúdo de cada mensagem obedecendo estritamente as associações definidas no gabarito. Já na “Linha da conversação gerada a partir do estabelecimento errôneo da associação”, se exibe também a cadeia de mensagens empilhadas e o conteúdo de cada mensagem, no entanto a primeira mensagem da linha da conversação é a mensagem que foi estabelicida errôneamente e a partir desta mensagem, é construída a linha da conversação conforme o determinado no gabarito.

Na figura, os outros sete participantes também indicaram, errôneamente, que a mensagem 15 estava associada a mensagem 6. A partir desta inferência coletiva incorreta, é possível visualizar a linha da conversação e analisar sua relação com a linha da conversação do gabarito.

Este módulo foi importante para se poder analisar e categorizar os tipos de erros discutidos no Capítulo 4, onde se estudou os padrões de erros de associações comparando a linha da conversação gerada pelo gabarito com a linha da conversação gerada a partir do erro da associação.

3.2.2.4 Sistema de bate papo com associações explícitas

O sistema típico de bate-papo usado no primeiro estudo foi transformado num sistema que exige que usuário se comunique fazendo ações explícitas, como por exemplo, enviar mensagem para todos e enviar mensagem para uma pessoa específica. Esse sistema foi gerado para avaliar o método de criação do gabarito usado no sistema analisador de associações, descrito no item 3.2.2.3.

Na Figura 24, a ação “Falar com todos” é exibida de forma inédita. Quando o usuário aciona esta ação, o sistema permite que ele escreva seu texto e envie para todos os participantes da sessão de bate-papo. No entanto, quando o usuário deseja falar com uma pessoa apenas. É necessário clicar na mensagem da pessoa para então aparecer a área de edição para que ele escreva sua mensagem em resposta a outra.

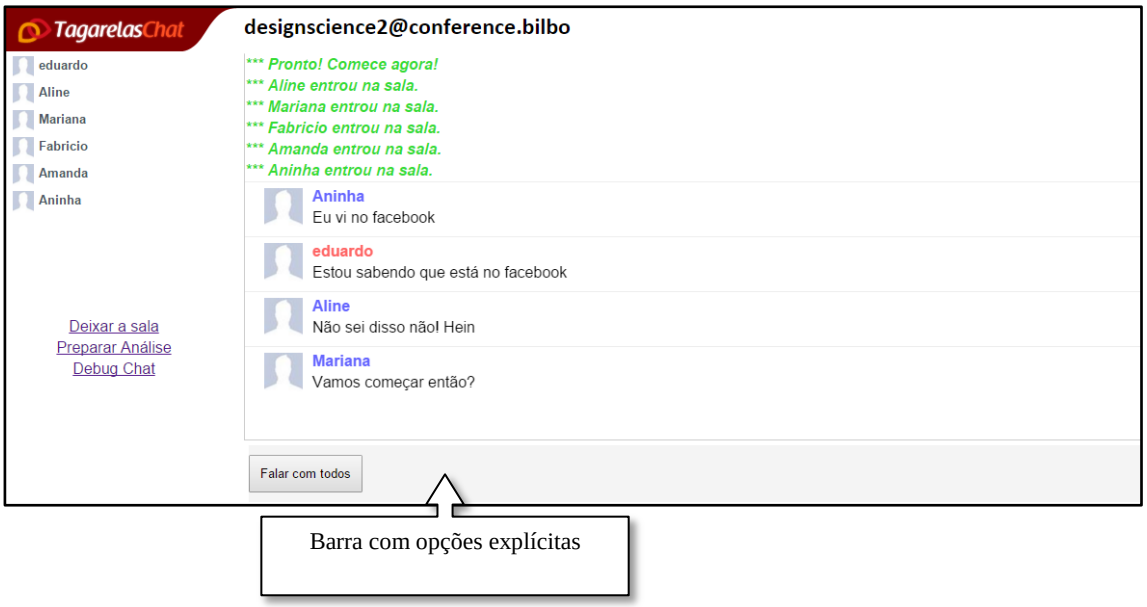


Figura 24 – Interface gráfica do sistema de bate papo com ações explícitas

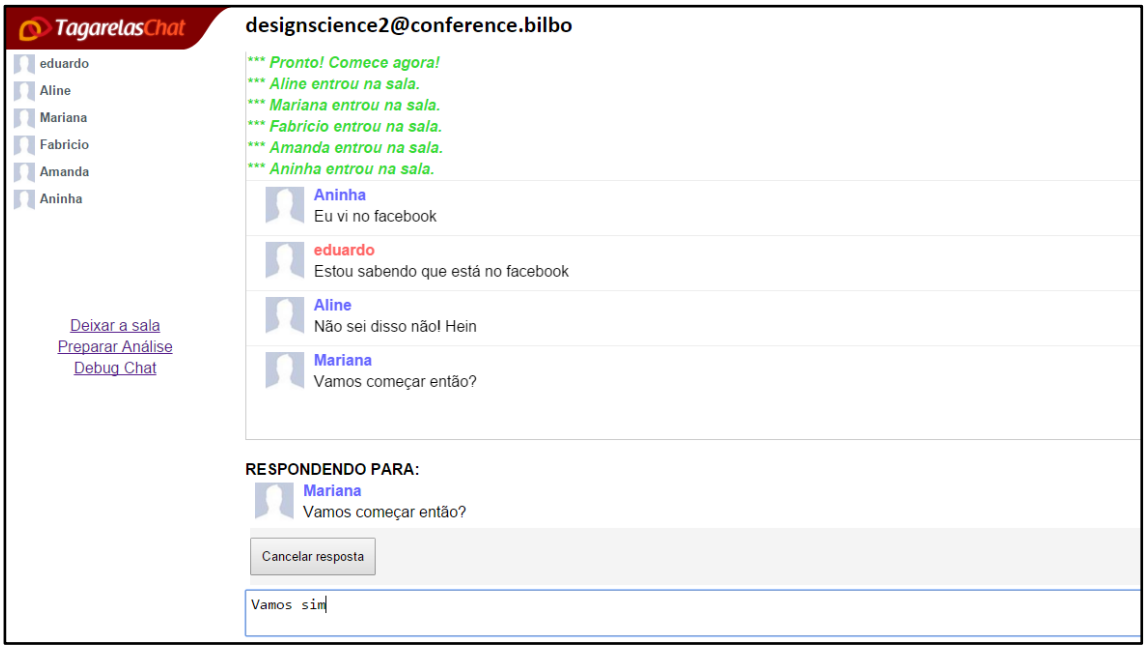


Figura 25 – Exemplo de envio de mensagens para pessoas especificas com identificação do co-texto durante a digitação da mensagem

Na Figura 25 é exibido um exemplo de envio de uma mensagem para um usuário específico. Nessa situação, Eduardo responde diretamente a Mariana. O co-texto é identificado na área “Respondendo Para”. Caso o usuário decida cancelar a ação, ele pode acionar o botão “Cancelar Resposta”. O sistema exibe o botão “Falar com todos” novamente.

3.3 **Estudo de caso piloto**

Um estudo de caso piloto foi realizado a fim de testar os procedimentos definidos e artefatos desenvolvidos. O estudo mostrou algumas falhas. Nem todos os alunos conseguiram completar a atividade. Alguns faltaram, outros desistiram devido ao grande número de mensagens para analisar. Alguns problemas de implementação atrapalharam o uso do sistema intrumental de análise das associações e, durante a fase de análise dados, foi percebido que houve falhas na captura do tempo entre as ações. Como resultado deste estudo piloto, foi possível refinar os critérios para a definição do debate, da turma e de todos os sistemas desenvolvidos.

3.4 **Realização do estudo de caso**

Após a realização dos ajustes necessários identificados com o estudo piloto, foi possível realizar o estudo de caso planejado. A análise foi realizada com duas turmas de Sistemas Colaborativos do programa de Pós-Graduação da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Uma turma com um professor e nove alunos e outra turma com um professor e cinco alunos. Como tema da sessão de bate-papo, o professor definiu discutir os pontos positivos e o que precisava ser feito melhor na disciplina para a próxima edição. Foi utilizado o Sistema típico de bate papo e o Sistema de bate papo com associações explícitas. As sessões de bate-papo duraram cerca de uma hora. Após a sessão os participantes receberam um vídeo, nele, o pesquisador mostra como realizar a análise das mensagens, utilizando o sistema instrumental para o registro das inferências das associações entre mensagens da sessão de bate-papo. Todos os participantes realizaram a **análise da conversação** ao longo de três semanas. A análise dos dados produzidos pelos participantes é apresentada na seção a seguir.

3.5 **Análise de dados**

As unidades de análise do estudo foram os alunos e as sessões de bate-papo geraram mensagens e inferências de associações entre essas mensagens. A Tabela 2 quantifica, em cada turma, o número de participantes, número de mensagens produzidas na sessão em cada log de bate-papo e o número de inferências analisadas nesta pesquisa. Cada inferência analisada tem até três possibilidades de resultado: a primeira, quando o participante reconhece e estabelece uma associação entre duas mensagens relacionadas. A segunda, de entender que não é possível

estabelecer uma associação entre duas mensagens e a terceira, a de não conseguir distinguir se uma mensagem tem ou não uma associação, levando o participante a desistir de associar.

Tabela 2 – Sessões de bate-papo analisadas

Turma	Participantes	Mensagens	Total	Inferências Corretas	Incorretas
SisColab2014.2	10	179	1.611	704	907
SisColab2015.2	6	88	441	213	228
Total	**16**	**261**	**2.318**		

O resultado da análise das ações foi analisado de forma estrita. Foram consideradas corretas as inferências corretas conforme determina o gabarito de respostas de associações de mensagens. Por exemplo, o estabelecimento de uma associação é correto quando o gabarito diz que uma mensagem deve se associada com outra mensagem e o usuário avaliador estabelece a associação conforme diz o gabarito. Se o gabarito diz que uma mensagem não tem relação com nenhuma outra, o usuário pode desistir de associar ou deixar explícito que a mensagem não tem associação. O participante pode desistir de associar uma mensagem por não se sentir seguro ou mesmo ter segurança para afirmar que determinada mensagem não tem associação. Quando o gabarito diz que uma mensagem tem associação com outra e o usuário desiste ou diz que não há associação, é considerada uma inferência incorreta.

A seguir, as três perguntas serão respondidas com base nos dados produzidos e documentados nas tabelas Tabela 3 e Tabela 4.

1. Com que frequência as pessoas erram as inferências que fazem?
2. As pessoas sempre fazem uma inferência, ainda que errada, ou assumem a incapacidade de inferir uma possível associação entre mensagem?
3. As pessoas demoram muito tempo para fazer uma inferência?

Essas perguntas serão respondidas considerando a realização de inferências de associações quando participante não é o próprio remetente da mensagem.

Tabela 3 – Resultado detalhado das unidades de análise, mostrando médias e correlações. Turma SisColab2015.2

	Resultado dos dados	Adilmar	Carol	Helmo	Isa	Marcio	Marildo	Mattos	Pinheiro	Richard	Vanessa
	Média de tempo para associar as mensagens dos outros	9	9	11	8	8	1	8	14	11	8
	Média de tempo para associar as próprias mensagens	12	7	6	15	8	3	9	28	7	7
	Percentual de desistências	0,04	0,00	0,01	0,00	0,00	0,78	0,05	0,06	0,00	0,00
	Correlação de tempo e distância sobre mensagens inferidas	0,53	0,53	0,18	0,26	0,50	0,07	0,36	0,51	0,26	0,33
TOTAL	Total de mensagens	179	179	179	179	179	179	179	179	179	179
TOTAL	Número de desistências	7	0	1	0	0	140	9	11	0	0
TOTAL	Inferências de Associações Corretas	89	91	98	119	82	53	96	91	80	84
TOTAL	Inferências de Associações Erradas	90	88	81	60	97	126	83	88	99	95
DESISTÊNCIAS	Média do tempo (em segundos)	17		8			0	8	18		
DESISTÊNCIAS	Maior tempo (em segundos)	57		8			22	28	53		
DESISTÊNCIAS	Menor tempo (em segundos)	2		8			0	4	3		
DESISTÊNCIAS	Quantidades corretas	5		1			39	2	8		
DESISTÊNCIAS	Quantidades erradas	2		0			101	7	3		
INFERÊNCIAS	Média do tempo (em segundos)	9	8	10	8	8	7	8	14	11	8
INFERÊNCIAS	Maior tempo (em segundos)	39	35	246	125	70	22	52	121	240	82
INFERÊNCIAS	Menor tempo (em segundos)	2	2	1	0,87	0,671	2	2	2	2	1
INFERÊNCIAS	Quantidades corretas	84	91	97	119	82	14	94	83	80	84
INFERÊNCIAS	Quantidades erradas	88	88	81	60	97	25	76	85	99	95
INF. CERTAS	Média do tempo (em segundos)	9	8	9	8	8	3	8	15	13	8
INF. CERTAS	Maior tempo (em segundos)	31	31	42	96	26	22	52	121	240	29
INF. CERTAS	Menor tempo (em segundos)	2	2	1	0,87	2	0,152	2	3	2	2
INF. CERTAS	Maior distância	59	48	59	48	59	13	48	48	59	59
INF. ERRADAS	Média do tempo (em segundos)	10	9	13	9	8	1	8	14	9	8
INF. ERRADAS	Maior tempo (em segundos)	57	35	246	125	70	17	36	68	25	82
INF. ERRADAS	Menor tempo (em segundos)	2	2	2	0,974	0,671	0,152	2	2	2	1
INF. ERRADAS	Maior distância	36	63	36	18	26	35	51	67	52	35

Tabela 4 – Resultado detalhado das unidades de análise, mostrando médias e correlações. Turma SisColab2014.2

	Resultado dos dados	Carlos	Lara	Maita	Pinheiro	Robson	Andrezza
	Média de tempo para associar as mensagens dos outros	16	15	10	13	28	7
	Média de tempo para associar as próprias mensagens	20	15	17	9	7	7
	Percentual de desistências	0	0	0	0,011	0,011	0
	Correlação de tempo e distância sobre mensagens inferidas	0,13	0,03	0,16	0,03	-0,08	0,12
TOTAL	Total de mensagens	88	88	88	88	88	88
	Número de desistências	0	0	0	1	1	0
	Inferências de Associações Corretas	44	48	46	53	43	0
	Inferências de Associações Erradas	43	39	41	35	45	0
DESISTÊNCIAS	Média do tempo (em segundos)				53	17	
	Maior tempo (em segundos)				53	17	
	Menor tempo (em segundos)				53	17	
	Quantidades corretas				1		
	Quantidades erradas					1	
INFERÊNCIAS	Média do tempo (em segundos)	16	15	11	12	24	7
	Maior tempo (em segundos)	69	74	74	99	760	48
	Menor tempo (em segundos)	0,5	1,4	0,9	0,2	2	1,5
	Quantidades corretas	45	49	47	53	44	48
	Quantidades erradas	43	39	41	35	44	41
INF. CERTAS	Média do tempo (em segundos)	14	12	12	9	9	6
	Maior tempo (em segundos)	61	63	74	32	40	17
	Menor tempo (em segundos)	1	1	1	1	2	1,5
	Maior distância	40	40	71	27	40	45
INF. ERRADAS	Média do tempo (em segundos)	18	18	11	17	39	9
	Maior tempo (em segundos)	69	74	27	99	760	49
	Menor tempo (em segundos)	0,5	1.4	1	1,3	3	1,5
	Maior distância	57	41	78	14	41	42

a. Com que frequência as pessoas erram as inferências que fazem?

Para responder essa pergunta, iremos usar os dados da Tabela 5 referente à turma SisColab2014.2 e também usaremos os dados da Tabela 6 referente a turma Sicolab2015.2. Em todas as tabelas observa-se que a quantidade de associações corretas é inferior a quantidade de associações incorretas. A Tabela 5 identifica a média de percentual de acerto em 43%, onde, aproximadamente a cada duas mensagens, um participante não sabe corretamente a que mensagem está se referindo. Na outra turma, identificada pela Tabela 6, a média percentual de acerto está em 48%, o que não é considerado uma diferença expressiva.

Tabela 5 – Resultado geral das unidades de análise mostrando associações corretas e erradas comparadas estritamente com o gabarito. Turma SisColab2014.2

	Associações				
Participante	Total	Acertos	%	Erros	%
Adilmar	171	81	47%	90	53%
Carol	156	68	44%	88	56%
Helmo	158	77	49%	81	51%
Isa	161	101	63%	60	37%
Marcio	163	66	40%	97	60%
Marildo	138	12	9%	126	91%
Mattos	165	82	50%	83	50%
Pinheiro	168	80	48%	88	52%
Richard	170	71	42%	99	58%
Vanessa	161	66	41%	95	59%
Média	**161**	**70**	**43%**	**91**	**57%**

Tabela 6 – Resultado geral das unidades de análise mostrando associações corretas e erradas comparadas estritamente com o gabarito. Turma SisColab2014.2

	Associações				
Participante	Total	Acertos	%	Erros	%
Carlos	79	36	46%	43	54%
Lara	73	34	47%	39	53%
Maita	73	32	44%	41	56%
Pinheiro	72	37	51%	35	49%
Raissa	74	48	65%	26	35%
Robson	70	26	37%	44	63%
Média	**74**	**36**	**48%**	**38**	**58%**

Em geral, os participantes analisaram mais as mensagens enviadas por outras pessoas do que suas próprias mensagens. Em todas as tabelas, basta observar o total de mensagens do grupo

e subtrair do número de associações que se inferiu. Observando a média de erros é validou-se que o problema da confusão e da perda de co-texto é grave e por isso ele se tornou objeto de estudo desta pesquisa.

Mas alguém apresentou um comportamento muito diferente dos demais? Para dar esta resposta, considerou-se o comportamento normal do grupo, isto é, verificou-se se a distribuição da quantidade de acertos se caracteriza como uma distribuição normal. Como ilustrado na Figura 26, observando os dados da turma SisColab2014.2, a mediana foi de 89 erros, com desvio padrão de 16,6 erros. Para verificar se alguém é discrepante do grupo, deve-se avaliar se há "ponto fora da curva" (FARIAS & CÉSAR, 2003), isto é, verificar se os valores extremos não estão adequados à distribuição da curva normal dos valores da amostra.

Do lado esquerdo da média, na Figura 26, está o grupo de pessoas que errou menos. Do lado direito, está o grupo de pessoas que errou mais. Foi identificado que 20% dos participantes ficaram em pontos extremos: Isa está muito acima da faixa do desvio padrão do grupo, pois ela acertou 66% das inferências que fez; Marildo, por outro lado, está muito abaixo da média, acertou apenas 30%**.**

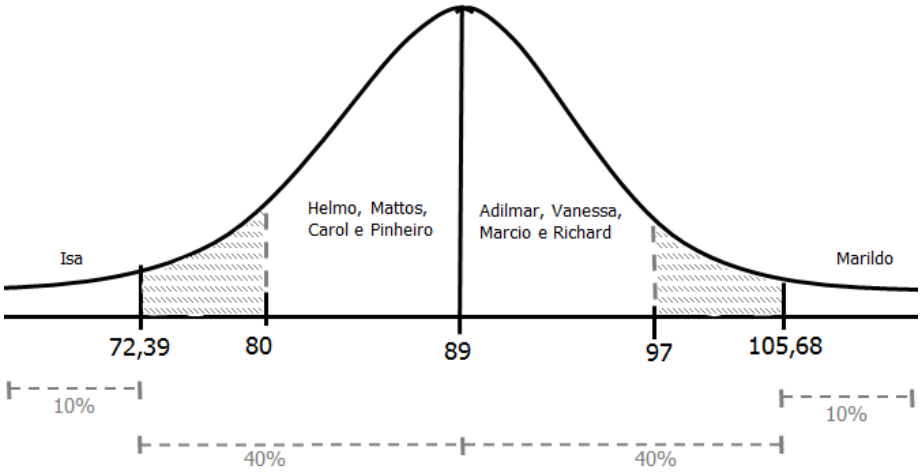


Figura 26 – Turma SisColab2014.2 - Distribuição normal dos erros de associações entre mensagens.

Observando a distribuição da turma SisColab2015.2 na Figura 27, é possível identificar a participante Raissa errou menos, caracterizando-se como um ponto fora da curva. A mediana de erros do grupo foi de 40 e o desvio padrão de 7 erros. Na turma SisColab2015.2, cerca de 80% das pessoas também se mantiveram próxima a mediana. Mas nessa distribuição de erros, o ponto 43 é a metade do desvio padrão. Entre 40 e 43 é a região onde se encontra a maioria das pessoas. Observamos que Carlos teve 43 erros (ponto que representa a metade do desvio padrão) e

Robson teve 44 erros, um ponto além. Embora visualmente a figure separe as pessoas, o fato é que numericamente elas estão muito próximas.

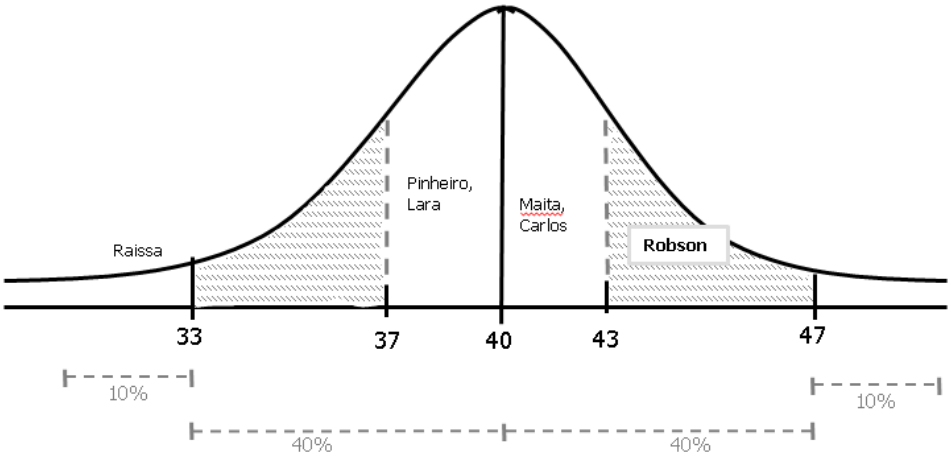


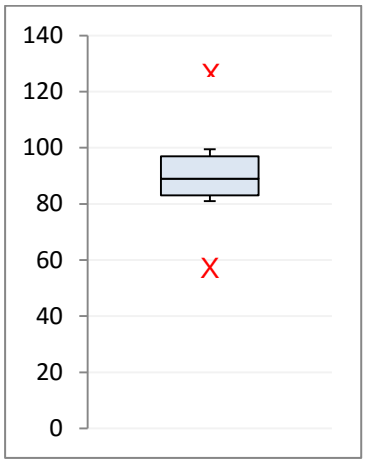
Figura 27 – Turma SisColab2015.2 - Distribuição normal dos erros de associações entre mensagens

Uma observação relevante entre a distribuição de erros dos dois grupos está na diferença da quantidade de participantes. A primeira turma com o dobro de participante gerou quase que o dobro de mensagens e consequentemente errou mais associações que o grupo menor. O primeiro grupo teve dois pontos foras da curva e o segundo grupo, praticamente teve apenas um. Mas como validar esses pontos fora da curva? Foi adota um segunda forma de análise (OUTLIER Calculator, 2015) para realizar um teste (GRAPHPAD, 2015) com diagrama de caixas.

A Figura 28 exibe o diagrama de caixas da turma SisColab2014.2. Neste gráfico, é possível identificar os pontos fora da curva (ou valores extremos) (DIAMOND & JEFFERIES, 2001). Observe a posição geométrica de Isa, que um desempenho muito superior aos demais participantes – por errar menos. Ela está fora do normal. Mas Marildo teve um desempenho muito abaixo dos demais participantes.

Outra característica importante visível na Figura 28, é que ela exibe uma caixa com distribuição não simétrica e levemente enviesada, onde na parte abaixo da média está representado o grupo de participantes. Na linha que divide o quadrado, representa a mediana em 89. O grupo classificado como abaixo da mediana, apresentou uma variação de erros entre 81 e 88 e com baixa representatividade de erros. Já o quarto acima da linha da mediana, possui espaço maior, com quantidade de erros entre 90 e 99, uma variação maior do que o quarto inferior. Entretanto o quarto de participantes abaixo da mediana está mais concentrado que o quarto

superior, pois seu lado da caixa está menor. Os dois pontos extremos em Isa (com 60 erros) e Marildo (com 126 erros) são identificados como os pontos fora da curva que, dependendo da análise do que se deseja fazer, devem ser desconsiderados.

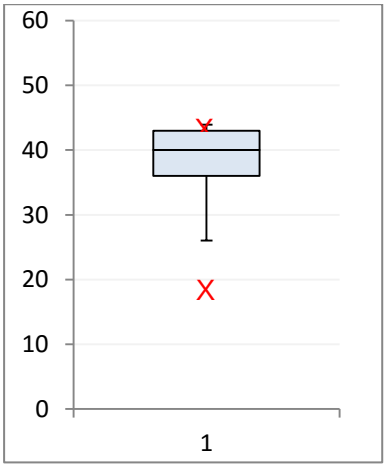


Min = 81, Q1 = 83, Med = 89, Q3 = 97, Máx = 99

Pontos fora da curva/Outliers = 126 e 60

Figura 28 – Visualização pontos extremos em 60 e 126 erros

A Figura 29 tem as mesmas características da Figura 28, mas mostra que a densidade de erros está mais concentrada no grupo inferior, abaixo da linha da mediana. O limite do terceiro quartil é 43, e difere muito pouco do ponto extremo em 44. Nessa figura podemos concluir que Robson embora seja um ponto extremo, ele não é um ponto fora da curva como a participante Raissa é, por ter errado bem menos que o grupo.



Min = 26, Q1 = 35, Med = 40, Q3 = 43, Máx = 44

Pontos fora da curva/Outliers = 26

Figura 29 – Visualização pontos extremos em 26 e 44 erros

b. As pessoas sempre fazem uma inferência, ainda que errada, ou assumem a incapacidade de inferir uma possível associação entre mensagem?

Para ser possível investigar esta questão, foi adicionado o botão "desisto" no sistema instrumental. O resultado do comportamento dos participantes da SisColab2014.2 em relação à desistência ao se tentar fazer inferências, é apresentado na Tabela 7.

Tabela 7 – Análises das desistências da atividade de associar mensagens

	Qual o percentual de desistência?	Quanto mais longe, mais se demora para descobrir a relação?	Demora-se muito para desistir?	Acertar é mais rápido ou demorado que errar?	Inferir as próprias mensagens é mais rápido?
Adilmar	4%	Não	Não	Indiferente	Não
Carol	0%	Não	Não	Indiferente	Sim
Helmo	1%	Não	Não	Indiferente	Sim
Isa	0%	Não	Não	Indiferente	Não
Marcio	0%	Não	Não	Indiferente	Não
Marildo	78%	Sim	Não	Não	Não
Mattos	5%	Não	Não	Indiferente	Não
Pinheiro	6%	Não	Não	Indiferente	Não
Richard	0%	Não	Não	Indiferente	Sim
Vanessa	0%	Não	Não	Indiferente	Sim

Neste estudo, 50% dos participantes não desistiram de inferir qualquer associação entre mensagens. A cada duas pessoas, uma sempre irá inferir (ainda que erradamente) a relação de todas as mensagens do bate-papo, sem nunca desistir. Nos resultados apresentados na Tabela 5, é possível perceber na linha de percentual de desistências, que metade não desistiu de associar mensagens, ou seja, eles não reconheceram a possibilidade de que uma mensagem pode não estar associada à outra mensagem. Isso evidencia o problema da conversação em sistemas típicos. Também é interessante notar que os demais participantes não costumam desistir muito. Com exceção do participante Marildo, que apresentou um percentual de desistência de 78% e tem um comportamento fora do normal (em termos de desistência, também se caracteriza como um ponto fora da curva). Seu comportamento demonstra falta de habilidade para participar de uma sessão de bate-papo, ou dificuldade (ou paciência) para realizar a atividade de analisar as mensagens e inferir as associações entre elas. Analisando o tempo para desistir, percebe-se que

ele é o mais rápido do grupo e conclui-se que ele encerrou a atividade realizando ações rápidas de desistência, sem interesse em fazer as associações corretamente.

Mesmo excetuando o participante Marildo, considerado um caso extremo e atípico, metade dos demais participantes desistiu de tentar inferir 4% das mensagens. Este é um dado que indica um certo grau de dificuldade para acompanhar a conversação. Eles não inferiram errado, a situação é ainda pior, eles desistiram de tentar descobrir a que mensagem anterior a nova mensagem estava se referindo.

No estudo com a turma SisColab2015.2 apenas duas pessoas desistiram de associar e apenas uma vez cada.

c. As pessoas demoram muito tempo para fazer uma inferência?

Conforme dados apresentados na Tabela 3, os participantes da turma SisColab2014.2 demoraram em média 8,7 segundos para estabelecer uma inferência de associação entre as mensagens enviadas por outros remetentes. Desconsiderando a participação do Marildo pelo o comportamento anormal, o tempo médio de participantes comprometidos é de 9,5 segundos. Usando os dados da Tabela 4, os participantes da turma SisColab2015.2 demoraram em média 12,5 segundos para inferir uma associação entre mensagens. De acordo com os dados da pesquisa de Rocha (2014), as pessoas demoram um tempo de leitura da mensagem proporcional à quantidade de caracteres da mensagem, podendo ser aproximado pela Equação 1:

$$\overline{T}_L = 1 + (0{,}04 * \text{número médio de caracteres por mensagem})$$

Equação 1. Tempo médio de leitura de uma mensagem de bate-papo (ROCHA, 2013, p.27)

Considerando que nesta sessão de bate-papo a quantidade média de caracteres por mensagem foi de 75, podemos afirmar que os participantes demoravam aproximadamente 4 segundos para ler a mensagem. Com este parâmetro, é possível relativizar o tempo para inferir a associação entre mensagens, que foi maior do que o dobro. Ou seja, os participantes demoram muito mais tempo tentando inferir a que mensagem do log uma dada mensagem se refere, do que lendo a mensagem propriamente dita. Este também é um dado importante, pois evidencia que a inferência de mensagens provoca realmente uma grande sobrecarga cognitiva.

4 Por que as pessoas erram?

No estudo apresentado no capítulo anterior, as pessoas erraram 43% das inferências sobre as associações entre as mensagens de bate-papo. Este resultado foi muito surpreendente para o nosso grupo de pesquisa, pois não esperávamos que as pessoas errassem tanto –tínhamos uma expectativa inicialmente de que o percentual de erro estaria entre 5% a 20%. Esse estranhamento nos motivou a investigar por que as pessoas erram tanto ao inferir as associações. Buscou-se identificar os padrões de erros e levantar alguns motivos para que o equívoco ocorra. Escolheu-se a teoria fundamentada em dados, por se tratar de um método de pesquisa interpretativista que explica a realidade a partir dos significados (MELLO; CUNHA, 2003). Neste capítulo, primeiro são apresentados os padrões de associação entre mensagens, na Seção 4.1. Estes padrões fundamentam o modelo para analisar os padrões de erros identificados e discutidos na Seção 4.2.

4.1 Estruturação do discurso e a modelagem da conversa em grafo

Esta seção apresenta os padrões de associação entre mensagens que ocorrem na sessão de bate-papo. Estes padrões formam a base para entender os padrões de erros discutidos na seção seguinte.

4.1.1 Estrutura linear do discurso

Uma sequência linear ocorre quando uma mensagem está associada com a anterior na ordem cronológica com que foram registradas na sessão, como no exemplo abaixo:

1. Pinheiro – Bom dia, alunos. Vamos iniciar nossa aula?
2. Richard – Bom dia professor, vamos sim!
3. Isa – Bom dia, Richard, estamos no mesmo grupo?

Texto 4 – Exemplo sequencial de mensagens

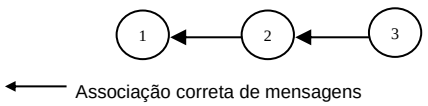


Figura 30 – Representação em grafo da sequência

No fragmento de bate-papo transcrito no Texto 4 e representado na Figura 30, a mensagem 1 é iniciada por Pinheiro, que é o professor da turma. A mensagem 2 é enviada por Richard como uma resposta à mensagem 1 do professor. Em seguida, Isa fala diretamente com Richard, o que indica que a mensagem 3 está associada à mensagem 2. Essas mensagens estão associadas com a anterior imediata no log, estabelecendo uma estrutura linear do discurso (não apenas cronológica).

4.1.2 **Estrutura em mensagens irmãs**

Identifica-se a estrutura de mensagens irmãs na situação em que uma mensagem dá origem a várias mensagens descendentes, que dão uma resposta para a mensagem precedente ou tratam do assunto contido nela.

1. Pinheiro **–** O que vcs acharam que funcionou na disciplina e que deve ser mantido numa próxima edição????
2. Richard **–** gostei muito do feedback que é dado no final de cada apresentação
3. Isa **–** eu curti a parte dos feedbacks depois de cada apresentação

Texto 5 – Exemplo sequencial de mensagens irmãs

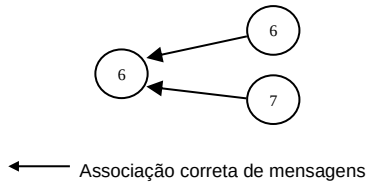
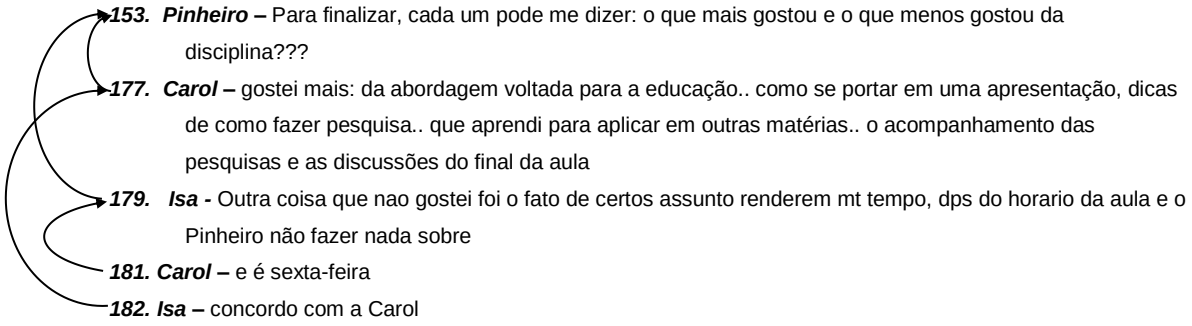


Figura 31 – Representação de mensagens irmãs

No fragmento de bate-papo transcrito no Texto 5 e representado na Figura 31, a mensagem 6 contém uma pergunta do professor aberta para todos os alunos, o que dá origem às mensagens descendentes 6 e 7, em que Richard e Isa enviam mensagem respondendo a mensagem 6 do professor. Este exemplo ilustra o discurso estruturado em mensagens irmãs.

4.1.3 **Estrutura da mensagem raiz**

A mensagem raiz é a primeira mensagem que precede todas as outras, desencadeadas na mesma árvore de conversação.



Texto 6 – Exemplo de mensagem raiz

No fragmento de bate-papo transcrito no Texto 6 e representado na Figura 32, a linha de conversação é iniciada pela mensagem 153. Nesta mensagem Pinheiro pede para encerrar a discussão e pergunta sobre o que as pessoas gostaram ou não na disciplina. Carol apresentou sua visão sobre os pontos positivos na mensagem 177. Isa expõe na mensagem 179 mais uma situação que ela avaliou negativamente sobre a condução da disciplina. Perceba que as mensagens 153, 177 e 179 apresentam uma estrutura do discurso em mensagens irmãs, mas estas, por sua vez, são continuadas pelas mensagens 181 e 182, aprofundando a linha da conversação.

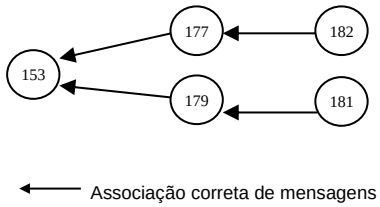
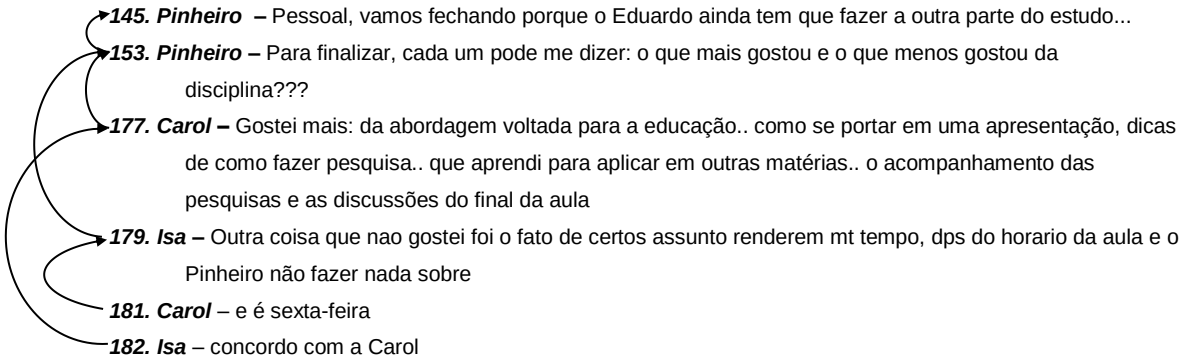


Figura 32 – Representação em árvore da sequência de mensagens: raiz e dois ramos

Isa responde a pergunta de Pinheiro com a mensagem 179 e, posteriormente, envia a mensagem 182 para dizer que concorda com a mensagem 177 de Carol - ela continua dentro da mesma árvore de conversação porque referenciou a Carol que, por sua vez, falava com a mensagem raiz. Independe do tamanho da linha da conversação, a mensagem 153 é a raiz, a mensagem comum que deu origem a toda esta árvore.

4.1.4 **Estrutura em monólogo**

Um monólogo é caracterizado como uma linha de conversação onde mensagens são enviadas seguidamente pelo mesmo remetente que dá continuidade ao discurso contido na mensagem precedente (PIMENTEL, 2003).



Texto 7 – Exemplo de uma linha de conversação com monologo de Pinheiro

O fragmento de bate-papo transcrito no Texto 7 difere do Texto 6 por incluir a mensagem 145. Identifica-se que Pinheiro inicia o argumento na mensagem 145, para então questionar a opinião dos alunos sobre a disciplina na mensagem 153. Essa linha de mensagens é identificada como um monólogo de Pinheiro, independentemente do número de mensagens que existam entre as mensagens 145 e 153.

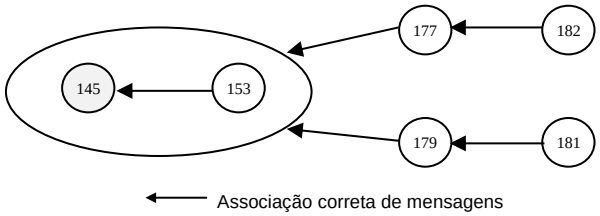
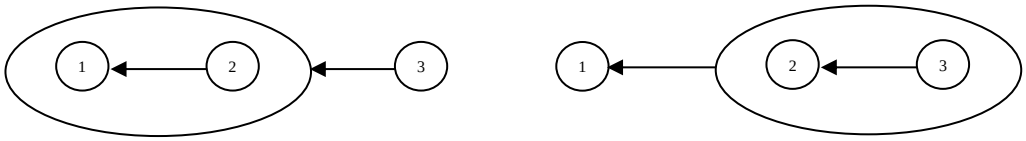


Figura 33 – Representação do monólogo como grupo

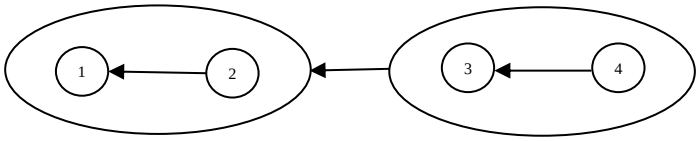
Na Figura 33, as mensagens 145 e 153 estão contidas num grupo que representa uma única unidade, o monólogo. Neste caso, é aceitável reconhecer que as mensagens 177 e 179 estão associadas ao grupo de mensagens que constituem todo o monólogo, e não apenas à mensagem 153, como ilustrado na Figura 32. Vale ressaltar, contudo, que esta modelagem em grupo não é possível em grafo. Grafo não admite a representação de grupos e só admite aresta entre vértices.

4.1.5 **Estrutura de associação com monólogos**

Considerando o agrupamento de um monólogo, conforme esquematizado na Figura 34, identifica-se que uma mensagem pode estar associada a um monólogo, ou um monólogo pode estar associado a uma mensagem precedente, ou mesmo que um monólogo pode estar associado a outro monólogo.



a) Uma mensagem fala para monólogo. b) Monólogo fala para uma mensagem



c) Um monólogo para um monólogo

Figura 34 - Modelos de interação com monólogos

4.2 **Erros de inferências das associações**

Nesta seção, são apresentados os padrões identificados dos erros cometidos pelas pessoas ao inferirem as associações entre mensagens no estudo de caso relatado no capítulo anterior.

4.2.1 **Dificuldade da inferência aumenta com o volume de mensagens**

Com o passar do tempo da sessão, é possível observar o aumento progressivo da quantidade de mensagens enviadas. O fato aumenta a complexidade para inferir as associações entre mensagens, já que o conjunto de postagens possíveis que a nova mensagem pode estar associada pode crescer consideravelmente. “Durante a conversação, a floresta de mensagens vai crescendo; algumas árvores vão encorpando, outras vão surgindo” (PIMENTEL M. , 2002, p. 141) como ilustrado na Figura 35.

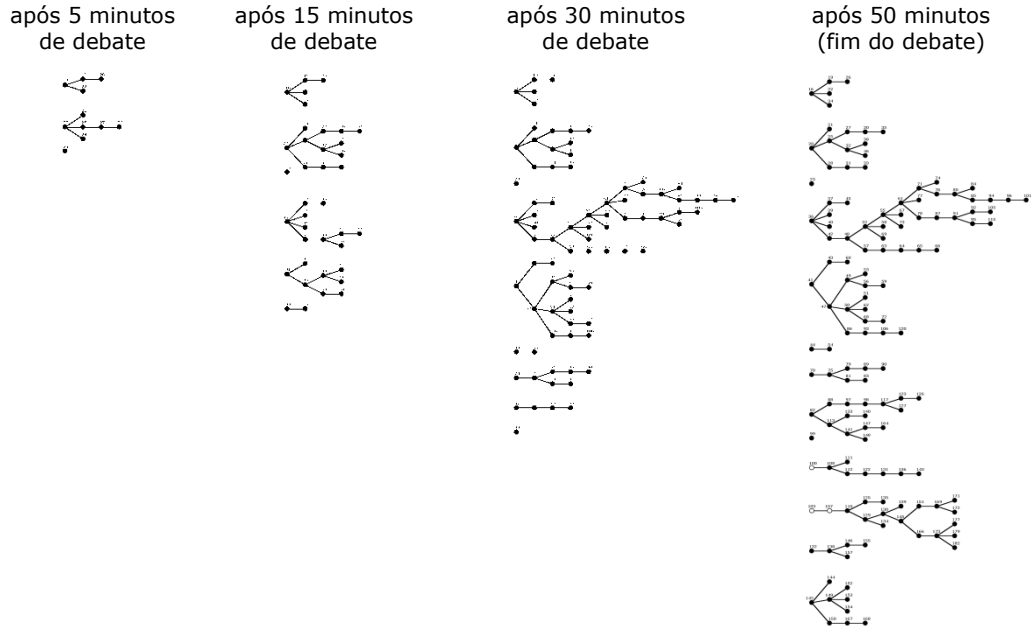


Figura 35 - Evolução temporal da floresta de mensagens [IINE, debate 4] (Pimentel, 2003)

No início da sessão, há menos mensagens e o processo de inferência de associação entre elas é facilitado. No entanto, quanto mais mensagens, mais as pessoas inferem erradamente as associações. Isso acontece devido ao aumento de possibilidades para se estabelecer a associação. Se a associação fosse um processo totalmente aleatório, a probabilidade de acertar a mensagem associada seria 1/(*n* mensagens anteriores), isso nos mostra que, quanto mais mensagens acumuladas, menor é a probabilidade de acertar a mensagem-referente.

4.2.2 **Erro por apontamento de irmãs**

Um tipo de padrão de erro identificado, é aquele que ocorre quando o sujeito infere erroneamente uma associação para uma das mensagens irmãs da mensagem-referente (a que contém o co-texto).

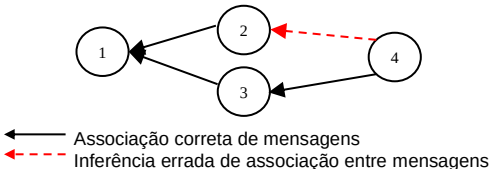
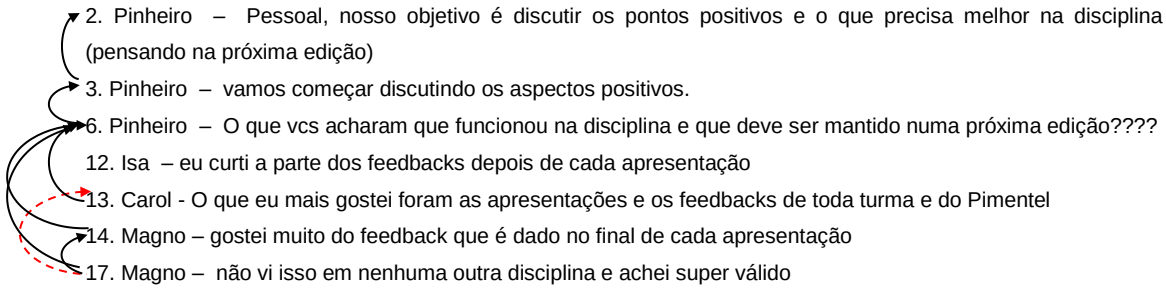


Figura 36 – Ilustração do erro de associação de mensagens irmãs

Na Figura 36 é apresentado um exemplo em que o sujeito conclui erradamente que a mensagem 4 está associada a mensagem 3, no entanto, o correto seria inferir que a mensagem 4

está agregada com a mensagem 2. Geralmente as mensagens irmãs possuem certa semelhança de conteúdo e, por isso, esta estrutura do discurso é mais suscetível a erros de inferências.



Texto 8 – Exemplo de erro de associação de mensagens irmãs

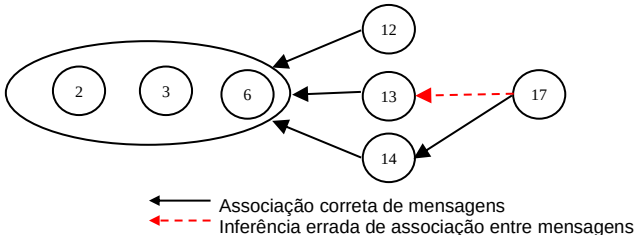


Figura 37 – Representação em árvore do erro de associação de mensagens irmãs

No fragmento de bate-papo transcrito no Texto 8 e representado na Figura 37, a mensagem 17 deveria estar associada com a mensagem 14, no entanto o sujeito inferiu que a mensagem 17 estava associada com a mensagem 13 que é irmã da mensagem 14. A situação de gostar do "feedback" era comum entre Carol (mensagem 13) e Magno (mensagem 14) e o fato de Magno não ter visto isto em nenhuma outra disciplina (mensagem 17), tornou difícil para que o sujeito percebesse que Magno estava continuando um monólogo (entre as mensagens 14 e 17) invés de responder diretamente a Carol.

4.2.3 **Erro por escolha de ramificações diferentes**

Quando a estrutura de conversação se ramifica, pode acontecer do sujeito inferir erradamente uma mensagem que está numa subdivisão diferente, já que elas compartilham alguma mensagem ancestral comum e, por isso, possuem alguma semelhança no assunto entre as mensagens descendentes.

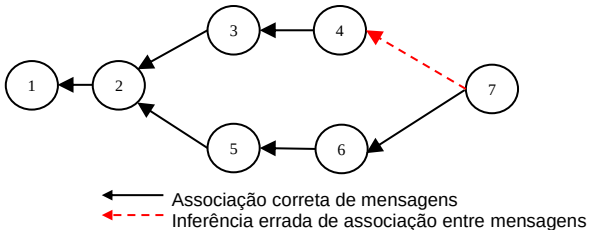


Figura 38 – Ilustração por erro de escolha de ramificações diferentes

A Figura 38 mostra uma linha de conversação iniciando com a mensagem 1 - raiz, em seguida uma estrutura de mensagens irmãs 3 e 5. A mensagem 4 é uma mensagem descendente da mensagem 3 assim como a mensagem 6 é descendente da mensagem 5. A última mensagem a compor a linha da conversação é a 7. Esta, deveria ser associada a mensagem 6, no entanto ela foi associada a mensagem 4. Perceba que quando acontece a estruturação das mensagens irmãs 3 e 5, também surge outras ramificações caracterizada pelas mensagens dependentes. A mensagem 4 pertence a uma ramificação diferente da mensagem 6 mensagem 4, mas ambas ramificações pertencem a mesma linha da conversação.

O fragmento de bate-papo transcrito no Texto 9 e representado na Figura 39, mostra a mensagem 85 como raiz, uma estrutura de mensagens irmãs 120 e 142, sendo que as mensagens 134, 140 e 141 são descendentes da mensagem 120. A mensagem 144 é descendente da 142. A mensagem 150 está associada à mensagem 144, pois Vanessa indicou que estava se referindo a Matteus. No entanto, um sujeito inferiu erradamente que Vanessa estava continuando a mensagem 141, também enviada pela Vanessa. Ou seja, o usuário erradamente inferiu que Vanessa estivesse realizando um monólogo.

85. Marildo – ou seja, os alunos especiais não agregam muito para o curso!

104. Pinheiro – Ah, Marildo, agora que entendi. Achei que vc estava brincando... o aluno especial (ouvinte), como o Vander, foi bom. O Silas não apareceu hoje, mas chegou a tumultuar? Me parecem pouco confiáveis, pois não estão sob a avaliação formal do curso, mas vc achou que atrapalhou?

120. Vanessa – perdemos 2 semanas de aula

129. Pinheiro – é, ele furou 2 vezes... no passado eu já impedi qualquer aluno especial. Dessa vez resolvi assumir o risco. Não tenho muita clareza ainda.

134 Carol – não acho q deveria cortar não

140 Leonardo – acho que pode acontecer com qualquer um, acho que ter um ou dois alunos especiais não atrapalha

141 Vanessa – ele contribuiu muito , o Anderson

142. Vander – Não sei Pinheiro, talvez se você pudesse fazer uma entrevista com os alunos especiais que querem entrar, assim como eu, seria bom...

144. Mattos – concordo, o problema não é ter aluno especial. É comprometimento

150 Vanessa – trouxe uma visao externa pra turma... visao de mercado e algumas criticas interessantes

Texto 9 – Exemplo do erro por escolha de ramificações diferentes

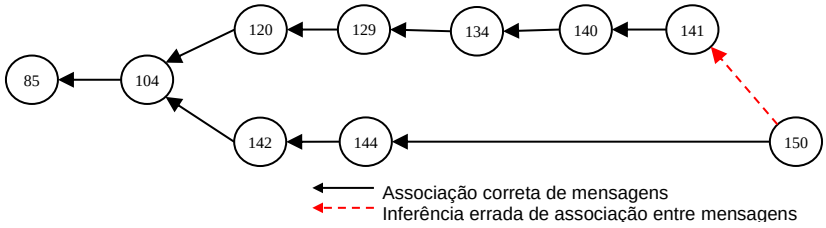


Figura 39 – Representação do grafo do erro de escolha de ramificações diferentes

Este exemplo mostra a complexidade sobre decidir a associação quando a linha da conversação é ampliada. A maioria dos participantes errou esta associação e, analisando o texto das mensagens 144 e 150, é possível identificar as características que definem o monólogo, pois Vanessa parece continuar o argumento da contribuição quando exemplifica com as visões que o aluno ouvinte trouxe para a turma. No entanto, conforme explicitado pela própria Vanessa, ela indicou estar dando continuidade à mensagem 144.

4.2.4 **Erro de associação para monólogos**

Este tipo de erro ocorre quando o sujeito identifica uma associação para uma mensagem que pertence a um monólogo presente na linha da conversação. Nesta situação, embora o sujeito não estabeleça a associação para a mensagem correta, ele ao menos a associa com uma das mensagens no monólogo originado pelo autor da mensagem correta. Este é o erro ilustrado na Figura 40, em que as mensagens 1 e 2 constituem um monólogo, e a mensagem 3 está associada à mensagem 2, no entanto, um sujeito infere erradamente que está associada à mensagem 1.

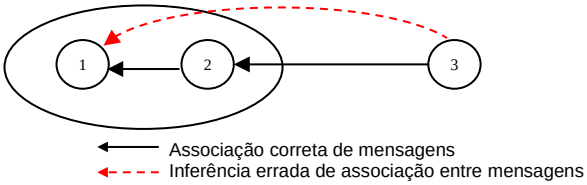


Figura 40 – Ilustração do erro de associação para monólogos

2. Pinheiro - Pessoal, nosso objetivo é discutir os pontos positivos e o que precisa melhor na disciplina (pensando na próxima edição)
3 Pinheiro **-** vamos começar discutindo os aspectos positivos.
6 Pinheiro **-** O que vcs acharam que funcionou na disciplina e que deve ser mantido numa próxima edição????
19 Isa - eu tb curti que o Pinheiro nao ficou preso soh em siscolab, ele deu orientacao sobre varias outras coisas

Texto 10 – Exemplo de associação para monologo

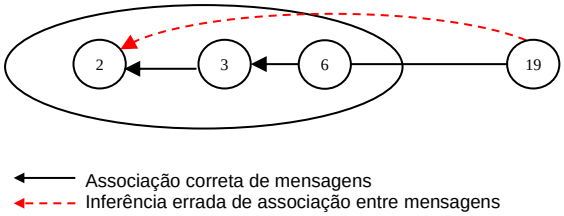


Figura 41 – Visualização do erro de associação para monólogos

O fragmento de bate-papo transcrito no Texto 10 e ilustrado na Figura 40, mostra o monólogo {2,3,6}. Na mensagem 6, Pinheiro questiona sobre o que funcionou bem na disciplina e a postagem 19, enviada por Isa, era uma resposta para esta mensagem. No entanto, um sujeito inferiu erradamente que a mensagem 19 estava associada a mensagem número 2, que inicia aquele monologo.

4.2.5 **Erro por associação para a mensagem raiz**

Este tipo de erro ocorre, como ilustrado na

Figura 42, quando o sujeito identifica uma associação para a mensagem raiz da linha da conversação da mensagem analisada. Nesta situação, embora o sujeito não estabeleça a associação para a mensagem correta, ao menos a associa com a linha de conversação correta. Isso acontece porque as mensagens daquela linha tratam de um mesmo assunto (há perda de co-texto, mas não do assunto).

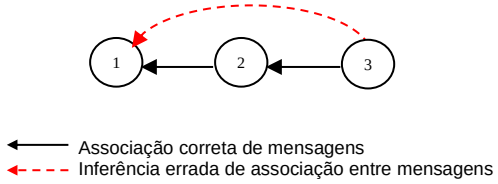


Figura 42 – Ilustração do erro de associação para a mensagem ancestral

O fragmento de bate-papo transcrito no Texto 11 e representado na

Figura 42, contém o monologo {2, 3 e 6} de Pinheiro e as irmãs 23, 19 e 27, descendentes da mensagem 6. A mensagem 32 está associada a 27, contudo, um sujeito inferiu erradamente que a mensagem 32 estava associada com a mensagem raiz daquela linha da conversação.

2. Pinheiro - Pessoal, nosso objetivo é discutir os pontos positivos e o que precisa melhor na disciplina (pensando na próxima edição)
3. Pinheiro **-** vamos começar discutindo os aspectos positivos.
6. Pinheiro **-** O que vcs acharam que funcionou na disciplina e que deve ser mantido numa próxima edição????
19. Isa - eu tb curti que o Pinheiro nao ficou preso soh em siscolab, ele deu orientacao sobre varias outras coisas
23. Adilmar - a monografia ser efetivamente uma pesquisa, achei que foi muito válido e importante.
27. Marcio - o suporte na monografia foi bem legal mesmo
32. Isa **-** a parte de metodologia cientifica foi sensacional

Texto 11 – Exemplo de associação para mensagem raiz

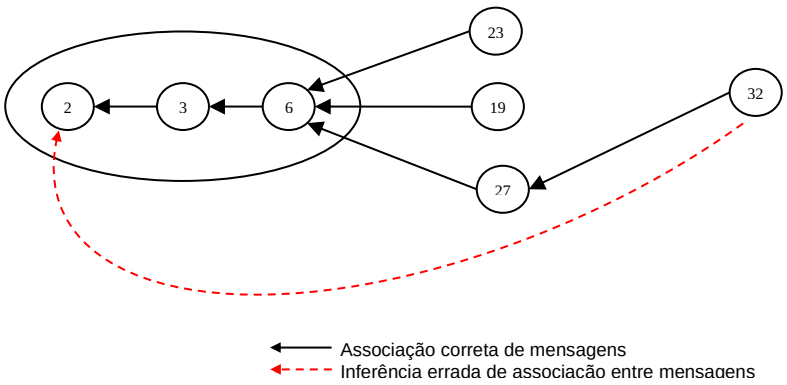


Figura 43 – Representação do grafo do erro de escolha da mensagem raiz

Estes foram os principais padrões de erros identificados na análise das inferências estabelecidas pelos participantes do estudo de caso relatado no capítulo anterior. Estes casos são mais fáceis de compreender porque as mensagens, de alguma forma, possuem relação na conversa.

5 Comparação entre a inferência humana e a computacional

O objetivo deste capítulo é medir a performance de um algoritmo ao inferir as associações entre mensagens da sessão de bate-papo em estudo e relativizar o resultado com a performance humana, conforme resultado apresentado no estudo de caso relatado no Capítulo 3 desta dissertação. Na Seção 5.1, é apresentado o algoritmo utilizado para a produção automática das inferências. Na Seção 5.2 é apresentado o projeto do presente estudo e na Seção 5.3 são relatados os resultados.

5.1 **Analisador de diálogo**

Dentre outras funcionalidades, O sistema de bate papo *AidChat* (LIMA, 2013), possui um módulo analisador de diálogo capaz de inferir associações entre mensagens durante a sessão do bate-papo. No entanto, o projeto *AidChat* tem o objeto recomendar conteúdos, depois de analisar o conteúdo das mensagens recebidas e identificar dúvidas e situações de falta de conhecimento dos participantes no assunto que estão discutido, apoiando o diálogo e seus usuários.

O módulo de análise do diálogo foi usado nessa pesquisa porque é responsável em realizar as inferências de associações entre mensagens. Essas inferências de associações são realizadas através de classificação estatística a partir do processo de Markov (RABINER, 1989). Como um algoritmo supervisionado, o aprendizado de máquina é análogo ao aprendizado humano que acontece a partir de experiências passadas para então gerar novos conhecimentos, melhorando sua habilidade em realizar operações no mundo real (LIU, 2011). Históricos de bate papo anteriores são usados neste aprendizado e com eles, é possível montar um modelo paras se aprender associar uma mensagem com outra com base em características estruturais e léxicas. Essas características estão listadas a seguir:

Características estruturais

- Número de mensagens entre A e B

Características léxicas

- Número de palavras em A
- Número de palavras em B

- Número de palavras em A que também existem em B
- Número de *n-grams* presentes em A e B (n entre 4 e 6)
- Primeira e última palavra de A
- Primeira e última palavra de B
- A contém o nome do usuário que enviou a mensagem B?
- B contém o nome do usuário que enviou a mensagem A?

O *AidChat* desenvolve o modelo com base nessas características estruturais e léxicas em cada mensagem e também adota algumas estratégias humanas para inferir associações entre mensagens (PIMENTEL & FUKS, 2009). Por exemplo, a característica estrutural que aponta o número de mensagens entre A e B é baseada na estratégia humana de Análise da recência, qual considera que uma mensagem B tem mais probabilidade de se associar com outra mensagem A se a diferença entre o envio da mensagem A e B estiver entre 10 segundos e 5 minutos. No entanto, o algoritmo não usa o tempo para identificar a probabilidade da relação entre as mensagens, é usada a quantidade de mensagens existentes entre A e B, limitando essa análise às 20 mensagens anteriores[9] da mensagem B, candidata a associação com outra existente.

A estratégia humana de análise de coesão considera a constituição lexical da mensagem, observando aspectos de sua estrutura e gramática como, por exemplo, a interseção de palavras similares entre uma mensagem A e uma mensagem B. No entanto, essa análise de coerência é feita de forma moderada porque nesse algoritmo não é possível identificar coesão textual utilizando o conceito semântico e suas relações de sentidos empregados no texto, apenas é caracterizado a superfície textual (ARAÚJO, 2015), com palavras e frases em sequencia linear. O *AidChat* possui analisador semântico apenas para realizar as recomendações mas essa estratégia não é empregada no analisador de diálogo.

A última estratégia humana empregada no módulo de Análise do diálogo é a Análise de sequências conversacionais, que considera a relação entre as mensagens em pares de adjacência, ou seja, se uma pergunta leva uma resposta ou um convite leva uma aceitação ou recusa, etc. O algoritmo analisa a semelhança entre a primeira palavra e última palavra de um par de mensagens, e também, se uma mensagem A contém o nome do usuário que enviou a mensagem B e vice versa.

[9] O trabalho escrito considera o limite máximo de 15 mensagens. No entanto, a implementação em java do algoritmo considera 20 mensagens – o último foi considerado na abordagem desta pesquisa.

A estratégia humana de identificação de monólogos não é explicitamente aplicada. É possível identificar um monólogo acidentalmente usando as estratégias que identificam a sequência conversacional, no entanto, a identificação do monologo exige heurísticas características e por isso não se identificou relação com essa estratégia humana. Já as análises do assunto e do contexto possui dependência, em alto grau, das estratégias de análise semântica do texto, que também não é empregado no analisador do diálogo.

O analisador do diálogo limita-se a atender 4 das 7 estratégias humanas e por isso ele foi escolhido na presente pesquisa. Na Figura 44 representa-se a arquitetura do "Módulo de Análise do Diálogo". É possível identificar os submódulos de pré-processamento, identificação de pares de adjacência e o comunicografo. A identificação de dúvidas e recomendação de conteúdo não interfere nos resultados do presente estudo e foram desconsiderados.

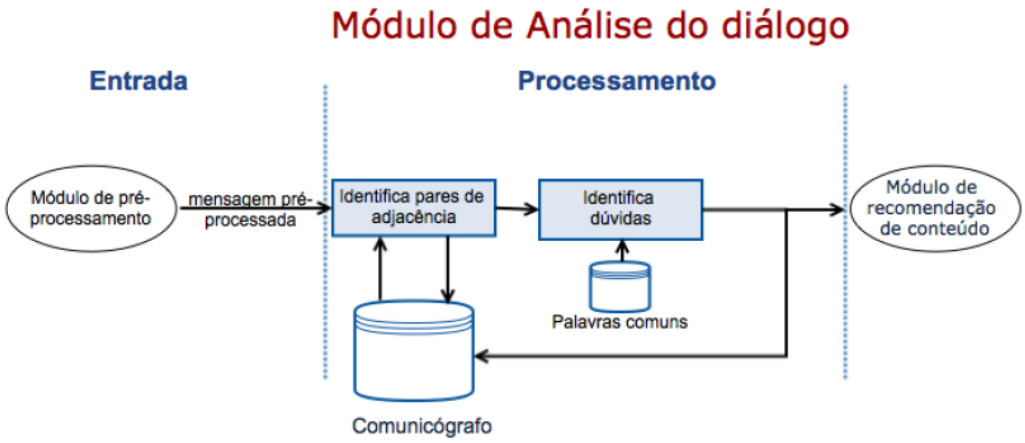


Figura 44 – Representação do grafo do erro de escolha da mensagem raiz

Na etapa "Entrada", é realizada a atividade de pré-processamento, onde são realizadas as atividades de remoção de *stopwords*, *Stemming* ou radicalização (em português), correção ortográfica das palavras.

Na etapa de processamento, realiza-se a identificação dos pares de adjacência, o que aqui caracterizamos como associação entre mensagens. Em seguida, monta-se o comunicografo, o banco de dados com as características estruturais e léxicas dos pares de adjacência de cada mensagem com as outras anteriores, conforme ilustrado na Figura 45.

Até aqui, investigou-se o Analisador do Diálogo de forma holística e seguiremos adiante detalhando-o um pouco mais para mostrar como tivemos que alterá-lo e usar na presente pesquisa.

A Figura 45 ilustra o processo em dois passos complexos. O primeiro tem a ver com a construção do comunicografo de treinamento. Nesses dados, deverá constituir o gabarito da associação entre as mensagens. No segundo passo é construído um segundo comunicografo considerando os dados do experimento, que deverá ter suas mensagens associadas.

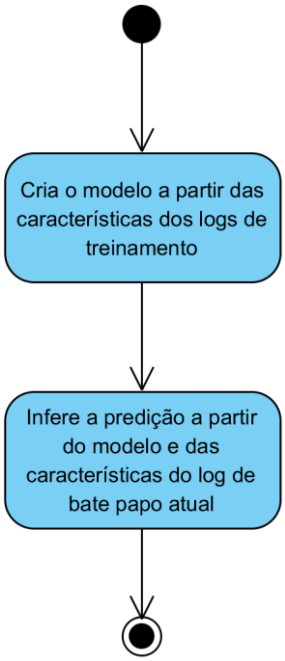


Figura 45 – Visão geral do processo de associação de mensagens

A Figura 46 apresenta o diagrama de classes usadas na implementação da presente pesquisa. Como o processo de inferir as associações acontece em dois passos concretos: treinar e experimentar. Foi identificado que ambos os passos compartilham funcionalidades parecidas. Logo, criou-se o a classe *Comunicografo*. Nesta classe, estão presente os recursos para gerar o banco de dados do comunicografo a partir dos que são fornecidos pela propriedade *DiretorioLog*. A propriedade *ArquivoCaraceristica* localiza o caminho físico do arquivo do comunicografo e a propriedade *ArquivoModelo* localiza o caminho do arquivo do modelo estatístico. O método privado *GerarCaracterísticas* é usado pela função básica *Executar*, junto com a função abstrata *ExecutarCore*, Como a classe *Comunicografo* é abstrata, as classes Treinamento e Experimento herdam suas características e implementam especificamente o método abstrato *ExecutarCore.*

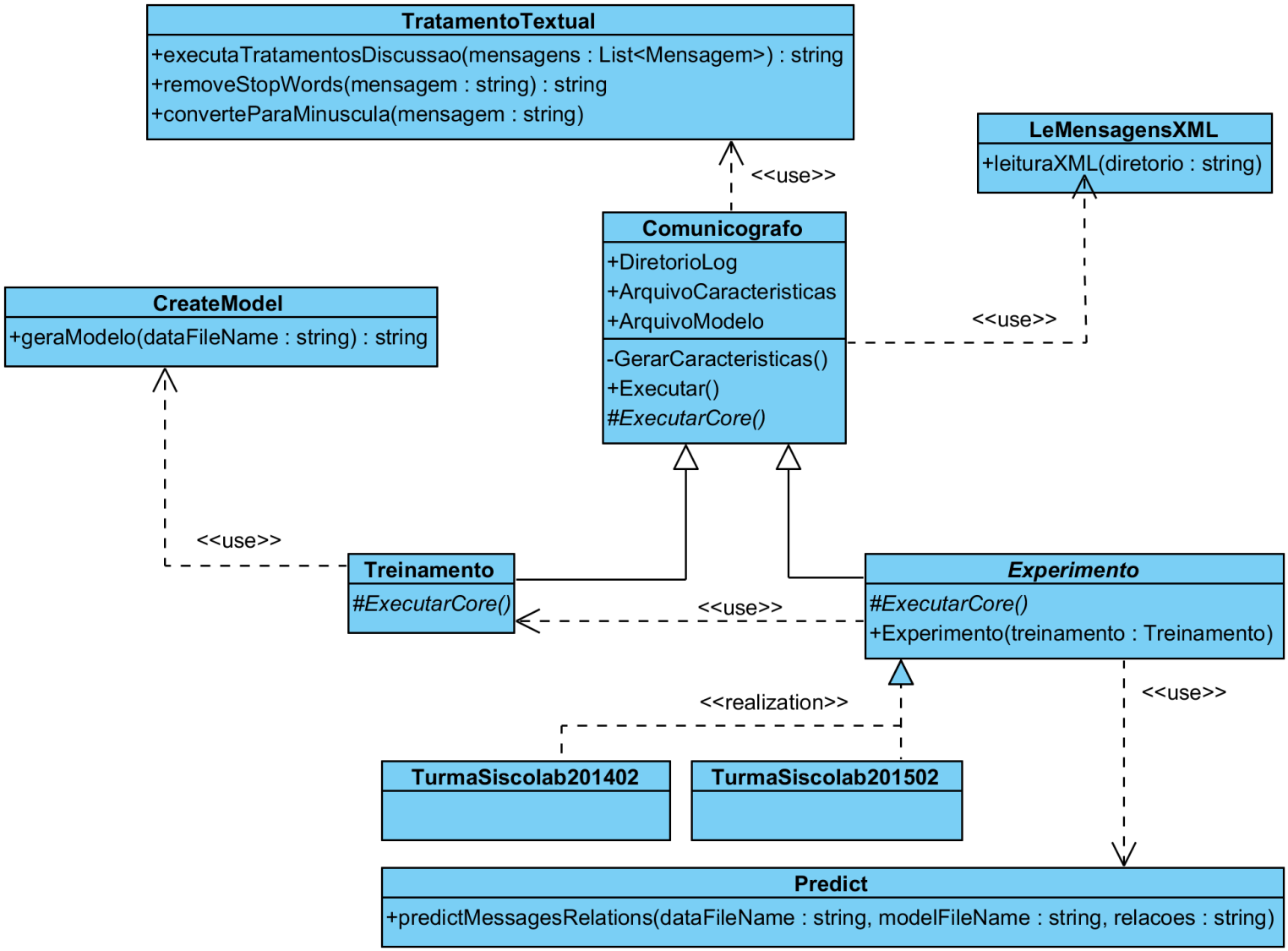


Figura 46 – Diagrama de classes do algoritmo de associação de mensagens

O Fato é que na classe *Treinamento* o método *ExecutarCore* tem uma implementação específica. Nesse processo, é usado a funcionalidade de gerar características (o arquivo de comunicografo) e em seguida o modelo estatístico é criado através da invocação do método *geramodelo* contido na classe *CreateModel*.

A classe de *Experimento*, que também herda a classe *Comunicografo*, implementa o método *ExecutarCore*, de uma forma mais complexa. Primeiramente, considera-se que para instanciar essa classe é requerido uma instância de treinamento com a propriedade *ArquivoModelo* devidamente preenchida. Seu processo de execução também gera as características (o arquivo de comunicografo) e em seguida, utiliza o método *predictMessagesRelations* da classe *Predict*, considerando o modelo estatístico criado pela instância de treinamento.

No entanto, a classe *Experimento* também é abstrata. Sua realização só pode ocorrer através das classes concretas que refletem o experimento do estudo: as classes TurmaSiscolab201402 e TurmaSiscolab201502.

```
public class Core
{
    public static void main(String[] args) throws FileNotFoundException, IOException
    {
        Treinamento treinamento = new Treinamento();
        treinamento.Executar();

        List<Experimento> experimentos = new ArrayList<Experimento>();
        experimentos.add(new TurmaSiscolab201402(treinamento));
        experimentos.add(new TurmaSiscolab201502(treinamento));

        for (Experimento experimento : experimentos)
        {
            experimento.Executar();
        }
    }
}
```

Figura 47 – Implementação do processo de execução do treinamento e experimento

Para melhor ilustrar o processo descrito junto com a descrição das responsabilidades das classes no diagrama e apresentado pela Figura 45 – Visão geral do processo de associação de mensagens usaremos Figura 47, que concretiza de forma visualmente simples a implementação desse processo complexo. No trecho de código apresentado, é instanciada a classe de treinamento para gerar o modelo. Em seguida, é montado um vetor de experimentos das turmas, passando dados treinados. Em seguida, demanda-se a execução de cada experimento.

5.2 **Projeto do estudo**

O objetivo do presente estudo é identificar o percentual de acerto das inferências produzidas computacionalmente pelo algoritmo analisador de dialogo e comparar o resultado com as inferências produzidas pelos próprios participantes da sessão de bate-papo, conforme relatado no Capítulo 3. Os mesmos logs de bate-papo sobre o qual foi produzido um gabarito pelos participantes daquele estudo, também foi usado para o algoritmo inferir as associações entre mensagens. Como resultado, os seguintes dados serão coletados:

- A quantidade inferências no total;
- A quantidade de inferências corretas;
- A quantidade de inferências incorretas.

5.3 **Análise dos dados**

Considerando que o módulo analisador de diálogo é um algoritmo de aprendizado, então é desejado estimar o quão eficaz o modelo preditivo é na prática. Como o algoritmo foi usado na pesquisa anterior para inferir associações entre mensagens, foi entendido que se poderia usar o mesmo modelo para analisar os logs desta pesquisa. Assim, o algoritmo foi rodado duas vezes, uma considerando os dados de treinamento e modelo da pesquisa anterior, com uma base de dados de 1936 mensagens de bate-papo já associadas, e outra considerando o próprio log da sessão de bate-papo em estudo.

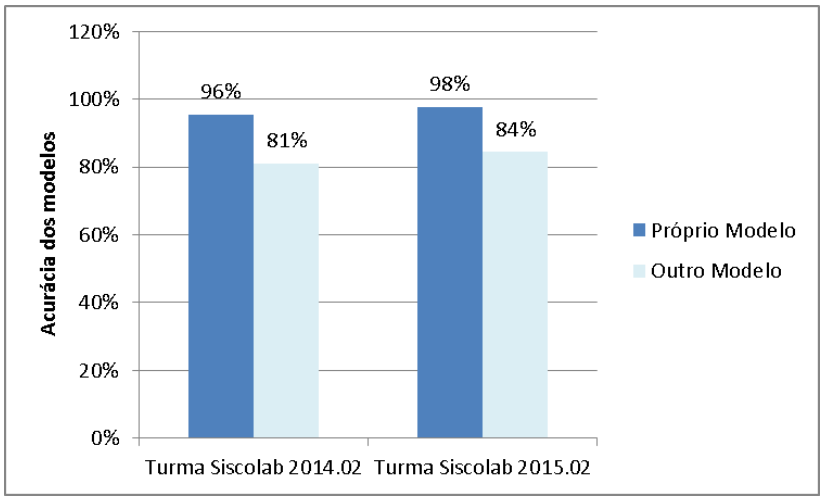


Figura 48 – Comparação da acurácia entre os modelos usados

Na Figura 48 são apresentados os percentuais de acerto do algoritmo considerando a escolha dos dados de treinamento para geração do modelo estatístico. A série "Próprio Modelo" apresenta o resultado do algoritmo quando o mesmo é rodado utilizando o a própria sessão de bate-papo como dados de treinamento. Então nas turmas Siscolab 2014.02 e SisColab 2015.02 os resultados de inferências de associações corretas são superiores a 90%. Já na série "Outro Modelo", apresenta-se o resultado do algoritmo quando mesmo é rodado utilizando 1290 mensagens (2/3 dos dados coletados nas sessões de bate-papo estudadas no AidChat). Assim, as turmas Siscolab 2014.02 e SisColab 2015.02 apresentam resultados semelhantes em torno de 80% de acertos.

Tabela 8 – Apuração das inferências realizadas pelo algoritmo

Descrição dos dados	Turma Siscolab 2014.02	Turma Siscolab 2015.02
Total de mensagens da sessão	179	90
Total de mensagens que devem ser associadas	116	46
Total de acertos do algoritmo para mensagens sem associação	63	44
Total de acertos do algoritmo para mensagens com associação	82	32
Total de acertos do algoritmo (com associação e sem associação)	145	76
Percentual de acertos (com associação e sem associação)	81%	84%
Total de mensagens que deveriam ser associadas (erro do algoritmo)	34	14
Percentual de acertos (apenas mensagens com associação)	71%	70%

A Tabela 8 apresenta a apuração detalhada dos resultados gerados pelo algoritmo quando ele utiliza dados de treinamento criados a partir de outras sessões de bate-papo – o que já fora mencionado com a série "Outro Modelo" na Figura 48. No entanto nesta tabela é apresentado resultados mais detalhados. O "total de mensagens da sessão" mostra a quantidade de mensagens totais da sessão considerando mensagens que tem associação e que não tem associação. O "Total de mensagens que devem ser associadas" mostra a quantidade mensagens que tem associações e não considera as mensagens que não tem associação. As linhas "Total de acertos do algoritmo para mensagens sem associação" e "Total de acertos do algoritmo para mensagens com associação" separa conjunto específicos de mensagens e a linha "Percentual de acertos (com associação e sem associação)" representa a situação que considera tudo que o algoritmo deveria acertar para dizer que uma mensagem tem associação com outra e quando a mesma não tem. Já a linha "Percentual de acertos (apenas mensagens com associação)" representa o percentual de acertos do algoritmo considerando o conjunto de mensagens que só tem associação. Analisar esses dois percentuais de forma diferente evidencia que o algoritimo

tem competência em associar mensagens e principalmente reconhecer quando mensagens não tem associação. Nos resultados estudados, o algoritmo sempre acertou 100% das situações em que ele entende que uma mensagem não tem associação. Nas situações onde as mensagem tem associação, o algoritmo erra em torno de 30%. Esse resultado é exibido nos totais na linha "Total de mensagens que deveriam ser associadas (erro do algoritmo)".

Como o propósito deste trabalho é comparar o resultado de inferências entre humanos e o algoritmo, não foi considerado as técnicas de validação como por exemplo "cross-validation", já que na pesquisa anterior já fora exercitado.

Para realizar a comparação do desempenho humano com o desempenho do algoritmo na questão de inferência de associações entre mensagem, foi usado o modelo estatístico criado a partir dos logs de outras sessões de bate-papo.

No capítulo 3 investigou o resultado da análise humana das associações. Em todas as turmas analisadas, a quantidade de associações corretas foi inferior a quantidade de associações incorretas. O percentual médio de acerto dos humanos foi de 48% considerando os logs analisados. Quando o algoritmo analisa o mesmo log, ele obtém um percentual médio de acertos de 83%, o que é 55% superior ao que um ser humano consegue inferir.

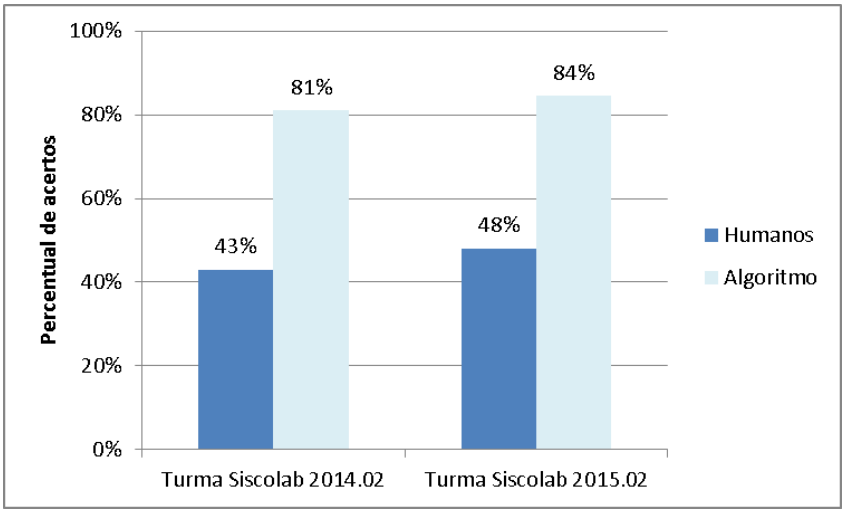


Figura 49 – Comparação do percentual de acertos entre humanos e algoritmo

Para chegar nesses resultados, foi considerado o modelo estatístico gerado a partir dos logs de outras sessões de bate-papo. A Figura 49 ilustra a comparação do resultado de cada turma em relação o desempenho do algoritmo. Analisando as associações da turma Siscolab 2014.02 os humanos tiveram um resultado médio de 43% de acertos, já o algoritmo teve 81%.

Na análise da turma Siscolab 2015.02 os humanos tiveram resultado médio de 48% de acertos contra 84% de acertos do algoritmo.

Os humanos têm dificuldade para inferir associações entre mensagens de forma correta e acertam, aproximadamente, uma a cada duas mensagens. O algoritmo estudado tem um desempenho superior de 55% em relação aos humanos e apresentou média de 83% de acerto na análise dos logs.

Por combinar algumas das estratégias humanas para inferir associações entre mensagens e técnicas de um algoritmo supervisado, o modelo estatístico do comunicografo apresentou ótimo desempenho para o contexto pesquisado. No entanto, há algumas possibilidades de melhorá-lo, tanto aumentando a quantidade de dados de treinamento quanto que desenvolvendo um modelo baseado nas heurísticas dos padrões de erros mais comuns para associar mensagem, que fora apresentado no Capítulo 4.

Outro fator a ser considerado, é que se os próprios participantes indicassem a mensagem que estão respondendo ao enviar uma nova mensagem, ou os demais participantes teriam uma boa chance de inferir erradamente com quem aquele sujeito está falando. Como consequência, parece que o mais adequado é desenvolver boas interfaces para o usuário estabelecer a associação da mensagem e boa interface para apresentar a mensagem referente (co-texto). Este é um objetivo que nosso grupo de pesquisa vem perseguindo há anos, junto com outros pesquisadores. O resultado obtido com a presente pesquisa indica que a busca por boas interfaces deve ser continuada em trabalhos futuros.

6 Conclusão

Esta dissertação apresentou uma pesquisa sobre a dificuldade humana para inferir as associações entre as mensagens de bate-papo, comparando a desempenho humana com o desempenho de um algoritmo. Este trabalho se justifica devido o aumento da popularização do uso de bate-papo em sistemas de redes sociais como o Facebook e também pelo crescimento da educação online que usa, dentre outros sistemas como fórum e e-mail, os sistemas de bate-papo.

Trabalhos anteriores tentaram resolver o problema da confusão da conversação e a perda de co-texto e esta pesquisa avança esses estudos investigando, empiricamente, os fenômenos da confusão em contexto real, medindo o desempenho dos participantes na atividade de associar mensagens das sessões que eles participaram.

A revisão da literatura caracterizou a questão da não linearidade das mensagens em sistemas típicos de bate-papo, os problemas de conversação como a confusão do bate-papo (falta de ligação entre as pessoas e o que elas dizem, falta de visibilidade dos progressos da produção de turnos, falta de visibilidade do progresso-auditivo, falta de controle sobre o posicionamento do turno, falta de registro útil e contexto social), também a situação da perda de co-texto. Mostramos como as pesquisas anteriores tentaram soluções diferentes para indicar estruturas de associação do discurso que se produz nas mensagens de bate-papo. Foi definido um modelo matemático baseado na teoria de grafos, que é usado para caracterizar e modelar os padrões erros de associação cometidos por seres humanos. Esses padrões foram discutidos e categorizados após analisar o erro das associações.

O estudo na área de processamento de linguagem na natural foi importante por causa da natureza do algoritmo que é baseado tanto nas características de um algoritmo supervisionado como também é baseado nas estratégias adotadas que os humanos adotam na hora de inferir uma associação entre mensagens.

Assim, nossa investigação procedeu apurando a situação de erros e acertos de associações cometidos por humanos. No primeiro experimento do estudo com a primeira turma, utilizou-se um sistema típico de bate-papo, em seguida as pessoas utilizaram o sistema instrumental para análise do co-texto. Assim, os participantes da sessão de bate-papo também indicaram, em outro momento, todas as associações das mensagens em que eles foram remetentes e também suas inferências de associações de mensagens que eles não foram remetentes. Com esse banco de

dados de associações, nosso estudo produziu um gabarito a partir agrupando todas as associações de mensagens que seus remetentes indicaram sua associação.

No segundo experimento do estudo, foi utilizado um sistema de bate-papo mais sofisticado, onde o participante da sessão passou a indicar o destinatário de suas mensagens de forma explícita. Ou as mensagens estavam sendo enviadas para todos ou para uma pessoa especificamente. Pelo fato do remetente dizer explicitamente com ele estava se referindo, o gabarito de mensagens associadas foi construído neste ponto, diferentemente do estudo anterior onde o gabarito foi criado por outro sistema, num momento posterior a realização da sessão de bate-papo.

Sobre a forma de criação dos gabaritos foi possível identificar que não houve diferença percentual sobre os erros e acertos por causa da escolha e do momento em que o gabarito fora criado. A frequência de erros e acertos das turmas analisadas é a quase a mesma. Na primeira turma foi 43% de acerto contra 48% da segunda turma (50% menor).

Na primeira sessão de bate-papo analisada, havia 10 participantes e a segunda turma 6 participantes. Na primeira sessão, foram criadas 179 mensagens contra 88 da segunda. O número de mensagens impactou na atividade de associar mensagens. 50% dos participantes da primeira sessão desistiram de associar as mensagens, é considerado que nessa turma teve mais mensagens e mais pessoas. Na segunda turma, houve duas desistências de duas pessoas e a situação de desistência se tornou irrelevante.

A distância entre duas mensagens candidatas a ter uma associação não influencia no tempo que o usuário demora para inferir uma associação. Tão pouco o tempo impacta na ação de desistir da associação, assim como é possível concluir que não é mais rápido acertar do que errar uma associação (e vice versa).

O algoritmo pesquisado infere associações entre mensagens através de classificação estatística a partir do processo de Markov. Como um algoritmo supervisionado, ele cria um modelo estatístico baseado em informações do passado e por isso foi considerado todo o histórico de sessões de bate papo realizadas em outro contexto de pesquisa. Assim, foi possível treinar o modelo estatístico, que tem como base a análise características estruturas e léxicas das mensagens que intrinsicamente considera quatro das sete estratégias humanas para inferir uma associação, sendo elas a análise da recência, análise de coesão textual, análise de coerência e análise de sequências conversacionais.

A apuração das inferências realizadas pelo algoritmo mostrou um resultado médio de acerto em 80%, o que é 55% superior ao resultado obtido pela média dos humanos.

A definição do modelo a ser escolhido para realizar inferências de associações é extremamente importante. Para treinar o algoritmo foi usado um histórico com 1936 de mensagens de bate-papo já associadas.

No entanto, qual a expressividade dessa diferença de resultado entre humanos e o algoritmo? No capítulo sobre a análise dos erros dos humanos foi estudado as categorias de erros. Cada categoria revela o padrão do erro, ao mesmo tempo revela uma oportunidade de flexibilização deste erro. Quão errado é? Não seria necessário diferenciar e ter o gabarito de forma estrita e flexibilizada? Nem todas as mensagens de bate-papo são direcionadas a uma pessoa especificamente também e o modelo estatístico usado pelo algoritmo não prevê isso. Então, ele chegaria a 100% no futuro? Dentre essa e outras respostas e perguntas, estudos futuros, poderiam diferenciar a forma como se indicou se o humano errou de fato através da flexibilização do erro.

Foi discutido que dificuldade em inferir uma associação aumenta proporcionalmente com volume de mensagens e daí identificou-se categorias de erros comuns, como o erro de associação para monólogos, erro por associação para a mensagem raiz, erro por escolha de ramificações diferentes e erro por apontamento de irmãs. Essas categorias mostram a complexidade que o diálogo ganha a medida que ele se desenvolve.

As próximas seções desta capítulos discutiram nossas contribuições, limitações e trabalhos futuros.

6.1 **Contribuições**

Com o uso cada vez mais difundido dos sistemas de bate-papos, principalmente no contexto da educação online, a presente pesquisa teve por objetivo investigar a dificuldade para inferir as associações entre mensagens de bate-papo – tanto das pessoas quanto de algoritmos.

A presente pesquisa constatou que o índice de erros por parte das pessoas é grande. Cerca de 43% das inferências que realizam sobre as associações entre mensagens são erradas. Esta grande quantidade de equívocos nos motivou estudar os padrões de estruturação do discurso (mensagens irmãs, raiz, monólogo) e os tipos de erros que as pessoas cometem ao inferir as associações entre as mensagens.

Além dos resultados já mencionados, a presente pesquisa também apresenta outras contribuições:

- Métricas (percentual de erros e acertos de inferências de associações, e tempo para inferir-desistir) para medir, evidenciar, caracterizar e avaliar a confusão da conversação no bate-papo que eventualmente resulta na perda de co-texto;
- Parâmetro para que se possa avaliar o desempenho de outros algoritmos para inferir as associações entre mensagens, tendo como base o desempenho dos humanos e do próprio algoritmo já investigado.

6.2 **Limitações e trabalhos futuros**

Os resultados obtidos na presente pesquisa se baseiam nas deduções realizadas pelos participantes em duas sessões de bate-papo. Em trabalhos futuros, é de interesse repetir o estudo em outras turmas e outros contextos, para que seja possível configurar um percentual preciso das associações entre as mensagens de uma sessão de bate-papo inferidas pelos seres humanos.

Os resultados obtidos na presente pesquisa, também se baseiam no uso de um único algoritmo para realizar as inferências das associações entre mensagens de bate-papo. Em trabalhos futuros, é de interesse avaliar a performance de outros algoritmos ou evoluir o algoritmo atual, levando em consideração o parâmetro humano e o desempenho comparativo entre os algoritmos em si, dessa forma, será possível buscar melhores soluções para a realização destas inferências.

Como trabalho futuro, pretende-se contribuir com a busca por um algoritmo que tenha um desempenho melhor do que o algoritmo estudado para esta dissertação.

A presente pesquisa avançou com o conhecimento sobre a confusão existente em conversações realizadas em bate-papos. Elas estão diretamente relacionadas, ao emaranhado de mensagens disponíveis no log onde os usuários precisam inferir sobre que mensagem se refere a anterior, fato que, eventualmente, provoca a perda de co-texto (desistência de inferência) ou gera inferências equivocadas. A presente pesquisa contribui com uma série de pesquisas anteriores, realizadas por Pimentel (2003), Fuks e Pimentel (2009), Morais (2011) e Netto (2014).

Ao todo, são 15 anos de pesquisa nesta linha, tempo em que o nosso grupo acumulou conhecimento suficiente para produzir bons artigos para divulgação em conferências e periódicos importantes do setor, como CSCW e ijCHS, atividades que serão objetivadas após a defesa desta dissertação.

7 Bibliografia

(2008). *ABNT NBR ISO 9001. (2008) Sistemas de Gestão da qualidade – Requisitos.* ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, Rio de Janeiro.

ANTON, H., & BUSBY , R. (2003). Algebra Linear Contemporânea. São Paulo, BR: ARTMED.

ARAÚJO, E. M. (2015). *O processo da coesão referencial em textos*. Acesso em 21 de 07 de 2015, disponível em Revista Lingua Portuguesa: http://linguaportuguesa.uol.com.br/linguaportuguesa/gramatica-ortografia/30/artigo219546-1.asp

AZEVEDO, V. (2011). *TabsChat: organização da conversação de um bate-papo em abas de discussão.* Rio de Janeiro: UNIRIO.

CALVÃO, L. D., PIMENTEL, M., FUKS, H., & LOPES, M. (2014). *Do email ao Facebook: uma perspectiva evolucionista sobre os meios de conversação da internet.* Rio de Janeiro: UNIRIO.

DIAMOND, I., & JEFFERIES, J. (2001). Beginning Statistics: An Introduction for Social Scientists. Londres – Reino Unido.

EPP, S. S. (2011). *Discrete Mathematics With Applications* (4ª ed.). Boston – US: Brooks/Cole CENGAGE Learning.

FARIAS, A., & CÉSAR, C. (2003). Introduçao à estatística. Rio de Janeiro – RJ: RTC.

FELDMAN, R., & SANGER, J. (2006). The text mining handbook : advanced approaches in analyzing unstructured data. New York, USA.: Cambridge University.

FURLAN, B., BATANOVIĆ, V., & NIKOLIĆ, B. (June 2 de 2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems, 55 - Issue 3*, 710-719.

GRAPHPAD. (2015). Acesso em 25 de 07 de 2015, disponível em QuickCalcs: http://graphpad.com/quickcalcs/Grubbs1.cfm

HERRING, S. (2001). Computer-mediated discourse. *The Handbook of Discourse Analysis*, pp. 612-634.

Holmer, T., Lukosch, S., & Kunz, V. (2009). Diminishing chat confusion by multiple visualizations. *Journal of Universal Computer Science* , pp. 3139-3157.

IDGNOW. (2012). *Anais do IV Seminário da Educação a distância: tão longe, tão perto*. Acesso em 18 de 12 de 2013, disponível em IGNOW INTERNET: http://idgnow.com.br/internet/2012/04/10/numero-de-internautas-no-brasil-chega-a-quase-80-milhoes/

IFENTHALER, D. (2010). *Learning and Instruction in the Digital age.* London, UK: Springer.

JIANG, J. (2012). *Mining Text Data – Information Extraction from Text.* Londres – Reino Unido. : Springer.

JURAFSKY, D., & MARTIN, J. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* (U. S. River, Ed.) Prentice Hall.

KLEIN, H. K., & MYERS, D. M. (03 de 1999). A set of principles for conduction and evaluating interpretive field studies in Information Systems. *Society for Information Management and The Management Information Systems Research Center, 23*(1), 67-93.

LIMA, B. (2013). *Uma abordagem de dados abertos visando solucionar problemas de comunicação textual.* Programa de pós-graduação em Informática, Instituto de Matemática, Institudo Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

LIU, B. (2011). Web Data Mining - Exploring Hyperlinks, Contents and Usage Data - Second Edition. Berlin: Springer-Verlag Berlin Heidelberg.

MARCUSCHI, L. (s.d.). *Análise da Conversação2007* (6ª ed.). São Paulo: Ática.

MELLO, R., & CUNHA, C. (2003). Operacionalizando o Método da Grounded Theory nas pesquisas em estratégia: Técnicas e procedimentos de Análise com apoio do Software Atlas. *Ti –Encontro de Estudos em Estratégia 1. Anais*.

MINER, G., IV, J. E., NISBET, R., DELEN, D., & FAST, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications* (1ª ed.). OXFORD: ELSEVIER.

MINER, G., IV, J., HILL, T., NISBET, R., DELEN, D., & FAST, A. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* OXFORD - UK: ELSEVIER.

MORAIS, E. (2011). *Debatepapo: sequências conversacionais e visualização do co-texto para compreensão da conversação em bate-papo.* Dissertação de mestrado, Universidade Federal do Estado do Rio de Janeiro – UNIRIO, Rio de Janeiro.

MÜHLPFORDT, M., & WESSNER, M. (2005). Explicit referencing in chat supports collaborative learning. *Proceedings of Computer-Supported Cooperative Learning (CSCL)* , pp. 460–469.

NETTO, A. (2014). *Sugestão de associações entre mensagens de Bate Papo: Um Experimento com o Sistema Debatepapo V. 2.* Dissertação de Mestrado. UNIRIO, Rio de Janeiro.

NUNES, R. R. (2009). *Pergunta-sem-resposta: Sistema InterVIU para a pesquisa e o desenvolvimento de bate-papo para entrevista.* Dissertação de Mestrado, Departamento de Informática Aplicada, UNIRIO., Rio de Janeiro.

NUNES, R., UGULINO, W., GONCALVES, , J. C., & SANTORO, F. M. (2008). K2Chat: uma Ferramenta de Bate-Papo com Suporte ao Registro e Indexação das Sessões. *V Simpósio Brasileiro de Sistemas Colaborativos*.

OLIVA , J., SERRANO, J., CASTILLO, M., & IGLESIAS , Á. (2011). SyMSS: A syntax-based measure for short-text semantic similarity,. *Data & Knowledge Engineering, 70, Issue 4*, 390-405.

OUTLIER Calculator. (2015). Acesso em 25 de 07 de 2015, disponível em http://www.miniwebtool.com/outlier-calculator

PARK, S., JI, S., Ryu, D., & A, C. H. (2008). A new cognition-based chat system for avatar agents in virtual space. *Proceedings of The 7th ACM SIGGRAPH International*

Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '08)(ACM).

PEREZ, L. (2015). *Hiperônimos e hipônimos*. Acesso em 31 de Outubro de 2015, disponível em Brasil Escola: http://www.brasilescola.com/gramatica/hiponimos-hiperonimos.htm

PESSOA, E. (2002). *Entrevist@: uma ferramenta de bate-papo para entrevistas.* IM/UFRJ, Rio de Janeiro.

PIMENTEL, M. (2002). *HiperDiálogo: ferramenta de bate-papo para diminuir a perda de cotexto.* . Dissertação de Mestrado, NCE-IM-UFRJ, Rio de Janeiro.

PIMENTEL, M. (2012). *Estudo de caso em sistemas colaborativos*. Acesso em 26 de 08 de 2015, disponível em https://www.dropbox.com/sh/ftcq79y2aqq16u4/_sFxQwE8iO

PIMENTEL, M., & FUKS, H. (2009). Studying Response-Structure Confusion in VMT. In: G. Stahl, *Studying Virtual Math.*

PIMENTEL, M., & STRUC, M. (Novembro de 2012). Portal Tagarelas: Bate-papo para educação. *23º Simpósio Brasileiro de Informática na Educação*.

PIMENTEL, M., FUKS, H., & LUCENA, C. (2005). Mediated Chat Development Process: Avoiding Chat Confusion on Educational Debates. *Computer Supported Collaborative Learning*, pp. 499-503.

PIMENTEL, M., FUKS, H., & LUCENA, E. (03 de 2006). R-U-Typing-2-Me? Evolving a chat tool to increase understanding in learning activities. *International Journal of Computer-Supported Collaborative Learning, 1*, 117-142 .

PRIMO, A. (2007). *Interação Mediada por computador – comunicação, cibercultura e coginicação* (3ª ed.). Porto Alegre. RS: Sulina.

RABINER, L. (Fevereiro de 1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE , vol.77, no.2*, 257-286.

RAMOS, M. (2008). Linguagens Formais e Autômatos. Universidade Federal do Vale do São Francisco.

RECKER, J. (2013). *Scientific Research In Information Systems – A Beginner's Guide.* (3ª ed.). Berlin: Springer.

ROCHA, E. (2013). *Modelo de Participação em Bate-papo Educacional*. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO., Rio de Janeiro.

SHIMAKURA, S. E. (2015). *Interpretação do coeficiente de correlação*. Acesso em 19 de 08 de 2015, disponível em Universidade Federal do Paraná: http://leg.ufpr.br/~silvia/CE003/node74.html

SMITH, M., JJ, C., & BURKHALTER , B. (2000). Conversation Trees and Threaded Chats. *Microsoft Research*.

UNICAMP. (2013). *Plano prevê ampliação do Fies para ensino a distância*. Acesso em 18 de 12 de 2013, disponível em http://www.revistaensinosuperior.gr.unicamp.br/notas/plano-preve-ampliacao-do-fies-para-ensino-a-distancia

WAINER, J. (2015). Acesso em 26 de 08 de 2015, disponível em https://www.dropbox.com/sh/ftcq79y2aqq16u4/_sFxQwE8iO

Wikipédia. (2015). *Wikipédia*. Acesso em 25 de 07 de 2015, disponível em DIAGRAMA de caixa.: https://pt.wikipedia.org/wiki/Diagrama_de_caixa

YIN, R. (2009). *Case Study Research: Design and methods* (4ª ed.). (T. Oaks, Ed.) California - USA: Sage .

ZHAI, C. X. (2012). *Mining Text Data – An introduction to text data.* Londres – Reino Unido: Springer.

Apêndice I

Termo de consentimento livre e esclarecido

UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO)
Av. Pasteur, 458, Urca, CEP 22290-240 – Rio de Janeiro, RJ
Programa de Pós-Graduação em Informática (PPGI)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Convidamos você para participar da Pesquisa **UMA ABORDAGEM COMPUTACIONAL PARA CRIAR ASSOCIAÇÕES ENTRE MENSAGENS DE BATE PAPO**, sob a responsabilidade do pesquisador LUIZ EDUARDO XAVIER DE CASTRO PEREIRA, a qual pretende **entender as estratégias humanas para associar diferentes mensagens num ambiente de bate papo**.

Sua participação é voluntária e se dará por meio ***da participação e análise de um debate. Para o debate, usaremos um sistema de bate-papo. Para a análise do debate, usaremos um sistema de análise de associações de mensagens. O debate será sobre os pontos negativos e positivos do mestrado em termos educacionais, e busca de soluções.***

Os riscos decorrentes de sua participação na pesquisa estão associados à divulgação indevida de seu nome no histórico das mensagens e caso ocorra, poderemos retirar seu nome do histórico. Se você aceitar participar, estará **contribuindo para criação de dados para pesquisa de dissertação do pesquisador e futuras pesquisas na área de sistemas colaborativos e da educação à distância mediada por meio de sistemas de bate papo**.

Se depois de consentir em sua participação e desistir de continuar participando, tem o direito e a liberdade de **retirar seu consentimento em qualquer fase da pesquisa**, seja antes ou depois da coleta dos dados, independente do motivo e sem nenhum prejuízo a sua pessoa.

Você **não terá nenhuma despesa** e também não receberá nenhuma **remuneração**. Os resultados da pesquisa serão **analisados** e **publicados**, mas sua **identidade não será divulgada**, sendo guardada em sigilo. Para qualquer outra informação, você poderá entrar em contato com o pesquisador no endereço [redacted], pelo telefone (21) 98624-0892, pelo e-mail luiz.pereira@uniriotec.br, ou poderá entrar em contato com o programa de pós graduação em informática da UNIRIO.

Consentimento Pós–Informação

Eu,_____, fui informado sobre o que o pesquisador quer fazer e porque precisa da minha colaboração, e entendi a explicação. Por isso, eu concordo em participar do projeto, sabendo que não vou ganhar nada e que posso sair quando quiser. Este documento é emitido em duas vias que serão ambas assinadas por mim e pelo pesquisador, ficando uma via com cada um de nós.

Rio de Janeiro, 28 de novembro de 2014

_____ _____

Assinatura do Participante Assinatura do Pesquisador Responsável

Apêndice II

Implementação do Algoritmo

```
public class Core
{
	public static void main(String[] args)
			throws FileNotFoundException, IOException
	{
		Treinamento treinamento = new Treinamento();
		treinamento.Executar();

		List<Experimento> experimentos = new ArrayList<Experimento>();
		//experimentos.add(new TurmaSiscolab201402(treinamento));
		experimentos.add(new TurmaSiscolab201502(treinamento));


		for (Experimento experimento : experimentos)
		{
			experimento.Executar();
		}
	}
}
```

```
public class Treinamento extends  Comunicografo
{
	public Treinamento()
	{
		//configuração padrão dos dados de treinamento
		this.DiretorioLog                  = "src/recursos/logs";
		this.ArquivoCaracteristicas = "src/recursos/treinamento_aidchat/comunicografo.dat";
		this.ArquivoModelo                 = "";
	}

	@Override
	protected void ExecutarCore() throws FileNotFoundException, IOException
	{
		System.out.println("Iniciando a leitura dos dados de treinamento");
		//executa a criação das features no comunicógrafo básico e depois
		//cria o modelo a partir destas características
		this.ArquivoModelo = CreateModel.geraModelo(this.ArquivoCaracteristicas);

		System.out.println("Dados foram trienados e o modelo criado em " + this.ArquivoModelo);
	}
}
```

```
public abstract class Experimento extends Comunicografo
{
	public Experimento(Treinamento treinamento)
	{
		this.DiretorioLog                       = "src/recursos/siscolab2015.2/log";
		this.ArquivoCaracteristicas   = "src/recursos/siscolab2015.2/comunicografo.dat";
		this.ArquivoModelo                      = treinamento.ArquivoModelo;
	}

	@Override
	protected void ExecutarCore() throws FileNotFoundException, IOException
	{
		System.out.println("Iniciando o metodo de predicao");
		//arquivo de destino com o log atual

		String relacoes = "src/recursos/siscolab2015.2/passo2_predicao_2_features_predicao.txt";
		//passa o arquivo de características mas usa o arquivo treinamentoModel.txt
		//produzido na função de criar modelos a partir
		Predict.predictMessagesRelations(this.ArquivoCaracteristicas, this.ArquivoModelo, relacoes);

		System.out.println("Concluindo o metodo de predicao");
	}
}

public class TurmaSiscolab201502 extends Experimento
{
	public TurmaSiscolab201502(Treinamento treinamento) {
		super(treinamento);
		this.DiretorioLog                       = "src/recursos/siscolab2015.2/log";
		this.ArquivoCaracteristicas   = "src/recursos/siscolab2015.2/comunicografo.dat";
	}
}
```

```
public class TurmaSiscolab201402 extends Experimento
{
	public TurmaSiscolab201402(Treinamento treinamento) {
		super(treinamento);
		this.DiretorioLog                          = "src/recursos/siscolab2014.2/log";
		this.ArquivoCaracteristicas   = "src/recursos/siscolab2014.2/comunicografo.dat";
	}
}

public abstract class Comunicografo
{
	public String DiretorioLog;
	public String ArquivoCaracteristicas;
	public String ArquivoModelo;

	private void GerarCaracteristicas() throws FileNotFoundException, IOException
	{
		LeMensagensXML leMsgs = new LeMensagensXML();

		List<Discussao> discussoes = leMsgs.leituraXML(this.DiretorioLog);

	    File arquivo = new File(this.ArquivoCaracteristicas);
	    FileOutputStream fos = new FileOutputStream(arquivo);
	    Map<Integer, Integer> frequenciaDistancias = new HashMap<Integer, Integer>();

	    for (Discussao discussao : discussoes)
	    {
	      //remove stop words e converte para lowcase
	      List<Mensagem> mensagens = TratamentoTextual.executaTratamentosDiscussao(discussao.getMensagens());

	      //constrói as características dos dados de treinamento
	      //também aplica o algoritimo de ir nas 20 últimas mensagens
	      GeraArquivoFeatures.geraFeaturesDiscussaoTreinamento(mensagens, fos, frequenciaDistancias);

	    }
	    fos.close();
	}
```

```
	public void Executar() throws FileNotFoundException, IOException
	{
		try
		{
			//tanto para o experimento quando que para o treinamento, é preciso gerar
			//as características do log de bate papo
			this.GerarCaracteristicas();

			//no entanto, há uma variação entre treinamento e experimento,
			//implementado nas classes respectivas
			ExecutarCore();
		}
		catch (Exception e)
		{
			System.out.println("Erro geral" + e.getMessage());
		}
	}

	protected abstract void ExecutarCore() throws FileNotFoundException, IOException;
}
```

```
public class GeraArquivoFeatures
{

public static void geraFeaturesDiscussaoTreinamento(List<Mensagem> mensagens, FileOutputStream fos, Map<Integer,
Integer> frequenciaDistancias) {

            int numeroMaximoDistanciaAdjacencia=0;
            System.out.println("geraFeaturesDiscussaoTreinamento");
            try {


                  for (Mensagem mensagem : (List<Mensagem>) mensagens)
                  {
                        String[] palavras = mensagem.getTexto().split(" ");

                        //Percorre das 20 mensagens anteriores atÃ© a mensagem anterior a essa mensagem a qual ela
faz referencia
                        if(mensagem.getNumero() !=null && !mensagem.getNumero().equals(""))
                        {
                              Integer valor = Integer.valueOf(mensagem.getNumero())-20;
                              if(valor < 0)
                              {
                                    valor =0;
                              }
                              for(int i=Integer.valueOf(mensagem.getNumero())-1;i>=valor; i--)
                              {
                                    /*
                                     * A mensagem B Ã© a mensagem para qual estamos procurando
                                     * referencia A
                                     */
                                    Integer numeroMsgsEntreAeB;
                                    Integer numeroPalavrasEmA;
                                    Integer numeroPalavrasEmB;
                                    Integer numeroPalavrasEmAeB = 0;
                                    String primeiraPalavraA;
                                    String ultimaPalavraA;
                                    String primeiraPalavraB;
                                    String ultimaPalavraB;
                                    Integer nGrams3PresentesEmAEB = 0;
                                    Integer nGrams4PresentesEmAEB = 0;
```

```
Integer nGrams5PresentesEmAEB = 0;
Boolean mensagemATemNomeB = false;
Boolean mensagemBTemNomeA = false;


//Recupera a mensagem que queremos saber se eh a referencia
Boolean temMensagem = false;
try
{
      temMensagem = (mensagens.get(i) != null);
}
catch (Exception e)
{
      System.out.println("Chora " + e.getMessage());
}
finally
{

}


if(temMensagem)
{
Mensagem mensagemA = mensagens.get(i);
String[] palavrasMensagemA = mensagemA.getTexto()
            .split(" ");


//FEATURES

//numero de mensagens em A e B
numeroMsgsEntreAeB = Integer.valueOf(mensagem.getNumero())
            - Integer.valueOf(mensagemA.getNumero())-1;

//Numero de palavras na mensagem anterior
numeroPalavrasEmA = palavrasMensagemA.length;
//Numero de palavras na mensagem atual
numeroPalavrasEmB = palavras.length;

for (int palavraMsg = 0; palavraMsg < palavras.length - 1; palavraMsg++) {
```

```
                                        for (int palavraMsgAnterior = 0; palavraMsgAnterior <
palavrasMensagemA.length - 1; palavraMsgAnterior++) {
                                            if (palavras[palavraMsg]
                                                        .equals(palavrasMensagemA[palavraMsgAnterior])) {
                                                  //Numero de palavras iguais nas duas mensagens
                                                  numeroPalavrasEmAeB++;
                                            }
                                            if(palavras[palavraMsg].length() >= 3 &&
palavrasMensagemA[palavraMsgAnterior].length()>=3 &&

      (palavras[palavraMsg].substring(0,3).equals(palavrasMensagemA[palavraMsgAnterior].substring(0, 3))))
                                            {
                                                  nGrams3PresentesEmAEB++;
                                            }
                                            if(palavras[palavraMsg].length() >= 4 &&
palavrasMensagemA[palavraMsgAnterior].length()>=4 &&

      (palavras[palavraMsg].substring(0,4).equals(palavrasMensagemA[palavraMsgAnterior].substring(0, 4))))
                                            {
                                                  nGrams4PresentesEmAEB++;
                                            }
                                            if(palavras[palavraMsg].length() >= 5 &&
palavrasMensagemA[palavraMsgAnterior].length()>=5 &&

      (palavras[palavraMsg].substring(0,5).equals(palavrasMensagemA[palavraMsgAnterior].substring(0, 5))))
                                            {
                                                  nGrams5PresentesEmAEB++;
                                            }

                                        }
                                  }
                                  for (int palavraMsg = 0; palavraMsg < palavras.length - 1; palavraMsg++) {
                                        if (palavras[palavraMsg].equals(mensagemA.getUsuario())) {
                                            //Retorna se a mensagem B contem o nome do usuario que enviou a
mensagem A em seu texto
                                            mensagemBTemNomeA = true;
                                        }
                                  }
                                  for (int palavraMsgAnterior = 0; palavraMsgAnterior < palavrasMensagemA.length -
1; palavraMsgAnterior++) {
```

```
                                        if (palavrasMensagemA[palavraMsgAnterior]
                                                    .equals(mensagem.getUsuario())) {
                                              //Retorna se a mensagem A contem o nome do usuario que enviou a
mensagem B em seu texto
                                              mensagemATemNomeB = true;
                                        }
                                  }

                                  primeiraPalavraA = palavrasMensagemA[0].replaceAll(" ","").replaceAll("/n", "");
                                  ultimaPalavraA = palavrasMensagemA[palavrasMensagemA.length-1].replaceAll("
","").replaceAll("/n", "");
                                  primeiraPalavraB = palavras[0].replaceAll(" ","").replaceAll("/n", "");
                                  ultimaPalavraB = palavras[palavras.length-1].replaceAll(" ","").replaceAll("/n",
"");

                                  Boolean parAdjacencia = false;

                                  if(mensagens.get(i) != null)
                                  {
                                        if(mensagem.getReferenciaTreinamento() != null &&
!mensagem.getReferenciaTreinamento().equals(""))
                                        {
                                              if(Integer.valueOf(mensagem.getReferenciaTreinamento()).equals(i))
                                              {
                                                    parAdjacencia = true;
                                              }
                                        }

                                  if(parAdjacencia)
                                  {
                                        String texto = "parAdjacencia-"+parAdjacencia.toString()+  "
numeroMsgsEntreAeB="+numeroMsgsEntreAeB
                                         + " "+ "primeiraPalavraA="+primeiraPalavraA + " " +
"ultimaPalavraA="+ultimaPalavraA+" "
                                        + "primeiraPalavraB="+primeiraPalavraB + " " +
"ultimaPalavraB="+ultimaPalavraB+" "
                                        + "mensagemATemNomeB="+mensagemATemNomeB.toString() +"
"+"mensagemBTemNomeA="+ mensagemBTemNomeA.toString()+" "+
                                        "ehPar"+"\n";
                                        if(numeroMsgsEntreAeB>numeroMaximoDistanciaAdjacencia)
```

```
                                        {
                                              numeroMaximoDistanciaAdjacencia = numeroMsgsEntreAeB;
                                        }
                                        if(frequenciaDistancias.containsKey(numeroMsgsEntreAeB))
                                        {
                                              frequenciaDistancias.put(numeroMsgsEntreAeB,
frequenciaDistancias.get(numeroMsgsEntreAeB)+1);
                                        }else{
                                              frequenciaDistancias.put(numeroMsgsEntreAeB, 1);
                                        }
                                        fos.write(texto.getBytes());

                                  }else{
                                        String texto = "parAdjacencia-"+parAdjacencia.toString()+  "
numeroMsgsEntreAeB="+numeroMsgsEntreAeB
                                         + " "+ "primeiraPalavraA="+primeiraPalavraA + " " +
"ultimaPalavraA="+ultimaPalavraA+" "
                                        + "primeiraPalavraB="+primeiraPalavraB + " " +
"ultimaPalavraB="+ultimaPalavraB+" "
                                        + "mensagemATemNomeB="+mensagemATemNomeB.toString() +"
"+"mensagemBTemNomeA="+ mensagemBTemNomeA.toString()+" "+
                                        "NaoEhPar"+"\n";
                                        fos.write(texto.getBytes());
                                  }
                            }
                                  }
                            }
                      }
                }
            }
            catch (Exception e)
            {
                  e.printStackTrace();
            }
            System.out.println("NUMERO MAXIMO="+numeroMaximoDistanciaAdjacencia);
      }
}
```

```
public class TratamentoTextual
{

	private static String removeStopWords(String mensagem) throws IOException
	{
		Analyzer analyzer = new BrazilianAnalyzer(Version.LUCENE_42);
		TokenStream tokenStream = new StandardTokenizer(Version.LUCENE_42, new StringReader(mensagem));
		CharArraySet stopSet = BrazilianAnalyzer.getDefaultStopSet();
		stopSet.add("Ah");
		stopSet.add("alem");
		stopSet.add("além");
		stopSet.add("desse");
		stopSet.add("nesse");
		stopSet.add("ai");
		stopSet.add("vc");
		stopSet.add("voce");
		stopSet.add("você");
		stopSet.add("nisso");
		stopSet.add("onde");
		stopSet.add("no");
		stopSet.add("na");
		stopSet.add("da");

		tokenStream = new StopFilter(Version.LUCENE_42, tokenStream, stopSet);
		//CharTermAttribute cattr = tokenStream.getAttribute(CharTermAttribute.class);
		CharTermAttribute termAttr = tokenStream.getAttribute(CharTermAttribute.class);

		tokenStream.reset();
		StringBuilder sb = new StringBuilder();
		  while (tokenStream.incrementToken()) {
			if (sb.length() > 0) {
			    sb.append(" ");
			}
			sb.append(termAttr.toString());
		    }
		    return sb.toString();

	}
```

```
public static List<Mensagem> executaTratamentosDiscussao(List<Mensagem> mensagens) throws FileNotFoundException,
IOException
	{
		for(Mensagem mensagem:(List<Mensagem>) mensagens)
		{
			String texto = TratamentoTextual.removeStopWords(mensagem.getTexto());
			texto = TratamentoTextual.converteParaMinuscula(texto);

			mensagem.setTexto(texto);
		}
		return mensagens;

	}


	public static String converteParaMinuscula(String mensagem)
	{
		return mensagem.toLowerCase();

	}
```

```java
/**
 * Main class which calls the GIS procedure after building the EventStream
 * from the data.
 *
 * @author  Chieu Hai Leong and Jason Baldridge
 * @version $Revision: 1.7 $, $Date: 2008/11/06 20:00:34 $
 */
public class CreateModel {

    // some parameters if you want to play around with the smoothing option
    // for model training.  This can improve model accuracy, though training
    // will potentially take longer and use more memory.  Model size will also
    // be larger.  Initial testing indicates improvements for models built on
    // small data sets and few outcomes, but performance degradation for those
    // with large data sets and lots of outcomes.
    public static boolean USE_SMOOTHING = false;
    public static double SMOOTHING_OBSERVATION = 0.1;

    private static void usage() {
      System.err.println("java CreateModel [-real] dataFile");
      System.exit(1);
    }

    /**
     * Main method. Call as follows:
     * <p>
     * java CreateModel dataFile
     */
    public static String geraModelo(String dataFileName) {
      boolean real = false;
      String type = "maxent";

      String modelFileName = dataFileName.substring(0,dataFileName.lastIndexOf('.')) + "Model.txt";

      try {
        FileReader datafr = new FileReader(new File(dataFileName));
        EventStream es;
        if (!real) {
          es = new BasicEventStream(new PlainTextByLineDataStream(datafr));
```

```
      }
      else {
        es = new RealBasicEventStream(new PlainTextByLineDataStream(datafr));
      }
      GIS.SMOOTHING_OBSERVATION = SMOOTHING_OBSERVATION;
      AbstractModel model;
      if (type.equals("maxent")) {

        if (real) {
          model = GIS.trainModel(es,USE_SMOOTHING);
        }
        else {
          model = GIS.trainModel(100, new OnePassRealValueDataIndexer(es,0), USE_SMOOTHING);
        }
      }
      else if (type.equals("perceptron")){
        System.err.println("Perceptron training");
        model = new PerceptronTrainer().trainModel(10, new OnePassDataIndexer(es,0),0);
      }
      else {
        System.err.println("Unknown model type: "+type);
        model = null;
      }

      File outputFile = new File(modelFileName);
      GISModelWriter writer =  new SuffixSensitiveGISModelWriter(model, outputFile);
      writer.persist();
    } catch (Exception e) {
      System.out.print("Unable to create model due to exception: ");
      System.out.println(e);
      e.printStackTrace();
    }

    return modelFileName;
  }
}
```

```
/**
 * Test the model on some input.
 *
 * @author Jason Baldridge
 * @version $Revision: 1.4 $, $Date: 2008/11/06 20:00:34 $
 */
public class Predict {
	MaxentModel _model;
	ContextGenerator _cg = new BasicContextGenerator();

	public Predict(MaxentModel m) {
		_model = m;
	}

	private void eval(String predicates) {
		eval(predicates, false);
	}

	private void eval(String predicates, boolean real) {
		String[] contexts = predicates.split(" ");
		double[] ocs;
		if (!real) {
			ocs = _model.eval(contexts);
		} else {
			float[] values = RealValueFileEventStream.parseContexts(contexts);
			ocs = _model.eval(contexts, values);
		}
		System.out.println("For context: " + predicates + "\n"
				+ _model.getAllOutcomes(ocs) + "\n");

	}
```

```
/* http://opennlp.apache.org/documentation/apidocs/opennlp-maxent/opennlp/maxent/io/GISModelReader.html
 * Retrieve a model from disk. It assumes that models are saved in the following sequence:
 GIS (model type identifier)
 1. # of parameters (int)
 2. the correction constant (int)
 3. the correction constant parameter (double)
 4. # of outcomes (int)
 * list of outcome names (String)
 5. # of different types of outcome patterns (int)
 * list of (int int[])
 [# of predicates for which outcome pattern is true] [outcome pattern]
 6. # of predicates (int)
 * list of predicate names (String)
 If you are creating a reader for a format which won't work with this (perhaps a database or xml file),
 override this method and ignore the other methods provided in this abstract class.
 * */
/**
 * Main method. Call as follows:
 * <p>
 * java Predict dataFile (modelFile)
 */
public static void predictMessagesRelations(String dataFileName, String modelFileName, String relacoes)
{
      boolean real = false;

      // Predict predictor = null;
      MaxentModel m = null;
      try
      {
            m = new GenericModelReader(new File(modelFileName)).getModel();
            // predictor = new Predict(m);
            System.out.println("Tirando o código da bianca");

            File file = new File(dataFileName);
            FileReader reader = new FileReader(file);
            DataStream linhas = new PlainTextByLineDataStream(reader);


            File arquivoComunicografo = new File(relacoes);
            FileOutputStream fos = new FileOutputStream(arquivoComunicografo);
```

```
int countAcertos = 0;
int countErros = 0;
int countEhparClassificadoEhpar = 0;
int countNaoehParClassificadoPar = 0;
int countNaoEhparClassificadoNaoEhpar = 0;
int countEhParClassificadoNaoEhPar = 0;

while (linhas.hasNext())
{
	// inicia a leitura linha a linha
	String linha	= (String) linhas.nextToken();
	String best	= m.getBestOutcome(m.eval(linha.split(" ")));

	String[] avaliado = linha.split(" ");
	String parCorreto = "";

	if (avaliado[0].contains("false"))
	{
		if (best.equals("NaoEhPar"))
		{
			countAcertos++;
			countNaoEhparClassificadoNaoEhpar++;
		}
		else
		{
			countErros++;
			countNaoehParClassificadoPar++;
		}
	}
	else
	{
		if (best.equals("ehPar")) {
			countAcertos++;
			countEhparClassificadoEhpar++;
			parCorreto = "-correto";
		} else {
			countErros++;
			countEhParClassificadoNaoEhPar++;
		}
```

```
                        }

                        String resultado = best + parCorreto + "\n";
                        fos.write(resultado.getBytes());
                  }

                  System.out.println("Acertos=" + countAcertos);
                  System.out.println("Erros=" + countErros);
                  System.out.println("Porcentagem=" + (countAcertos * 100)
                              / (countAcertos + countErros));
                  System.out
                              .println("Acerto: Nao eh par classificado como nao Eh par= "
                                          + countNaoEhparClassificadoNaoEhpar);
                  System.out.println("Acerto: Eh par classificado como Eh par= "
                              + countEhparClassificadoEhpar);
                  System.out.println("Erro: Nao eh par classificado como Nao par= "
                              + countNaoehParClassificadoPar);
                  System.out.println("Erro: Par classificado como Não par= "
                              + countEhParClassificadoNaoEhPar);

                  fos.close();
            }
            catch (Exception e)
            {
                  System.out.println("Unable to read from specified file: "
                              + modelFileName);
                  System.out.println();
                  e.printStackTrace();
            }
      }
}
```