Análise estrutural de bate-papo na Web: um método para organizar o diálogo baseado em comunicografia e máxima entropia visando facilitar a extração de informações relevantes

Bianca Munaro Lima, Maria Luiza Machado Campos

Programa de Pós-Graduação em Ciência da Computação NCE/UFRJ
C.P. 68530, CEP 21941-590, Ilha do Fundão, Rio de Janeiro – Brasil

bianca.munaro@nce.ufrj.br, mluiza@nce.ufrj.br

Abstract. This paper describes an approach for analyzing the structure of dialogues produced in textual communication tools. The text generated in chat sessions has a non-linear format, because the messages are not associated with the preceding message. Therefore, it is interesting to analyze the structure of the dialogue building a graph with the associations between messages for later analysis and a better understanding. This paper presents an approach based on *comunicografia*, a theoretical methodology for automating the analysis of dialogues.

Resumo. Este artigo descreve uma abordagem para a análise da estrutura de diálogos produzidos em ferramentas de comunicação textual. O texto gerado em uma sessão de bate-papo segue um formato não-linear, pois as mensagens não estão associadas com a antecedente. Portanto, é interessante analisar a estrutura do diálogo de forma a evidenciar as associações entre as mensagens para sua análise posterior e um melhor entendimento. Este artigo apresenta uma abordagem baseada na comunicografia, metodologia teórica para a automatização da análise de mensagens.

1. Introdução

A Internet tem revolucionado o mundo da comunicação como nenhuma outra invenção foi capaz de fazer antes. Desde o seu inicio, o objetivo principal da Web sempre foi facilitar a comunicação, possibilitando a troca de informações através de uma interface simples e intuitiva.

A comunicação mediada por computador trouxe diversas vantagens para os diferentes grupos de usuários, dentre elas seu alcance, o grande número de facilidades de apoio para tratamento e acesso às informações trocadas, além de possibilidades de análise sobre essas interações.

No entanto, esse tipo de comunicação pode também apresentar dificuldades, principalmente quando envolve apenas comunicação textual. O problema é que muitas vezes os usuários podem não expressar suas idéias com clareza, ou mesmo as palavras podem sofrer interpretações errôneas, por terem significado ambíguo ou mesmo por estarem apresentadas ou consideradas em contexto distinto daquele intensionado pelo emissor, causando problemas na comunicação.

Geralmente, o texto gerado em uma sessão de bate-papo segue um formato não-

linear, pois nem sempre as mensagens estão associadas com a mensagem anterior. Isso acontece porque muitos assuntos são discutidos ao mesmo tempo e é comum ocorrerem referências a mensagens anteriores.

As características cognitivas do interlocutor, como memória, atenção e interesse e as características do grupo, como número de participantes e rapidez na reprodução de mensagens também podem influenciar na comunicação. Assim, de acordo com as dificuldades descritas, os usuários acabam se sentindo "perdidos", podendo ocorrer a chamada "perda de co-texto" [Pimentel et al, 2003]. A expressão "perda de co-texto", segundo Pimentel, foi criada para representar o momento em que o leitor não consegue estabelecer o encadeamento da conversação.

Quando um participante não consegue acompanhar a discussão, ele procura nas mensagens anteriores a referencia em que se perdeu. No caso em que o usuário não entende uma parte do diálogo, ele pode fazer uma pergunta ou uma pesquisa sobre o assunto. Essas interrupções podem prejudicar o fluxo da conversa dificultando o acompanhamento da a discussão.

Portanto, esse trabalho faz parte de uma abordagem mais abrangente que visa analisar as mensagens de um chat, descobrir os pontos em que ocorreram problemas de comunicação e recomendar informações interessantes que colaborem construtivamente com a discussão. Esse artigo foca apenas na proposta de um método para analisar e estruturar a comunicação em chats.

2. Análise do diálogo – Trabalhos Relacionados

A partir da necessidade de analisar e organizar automaticamente a grande quantidade de informações textuais que são produzidas ao longo das discussões mediadas pelas atuais ferramentas de comunicação textual da Internet, uma metodologia chamada comunicografia [Pimentel et al. 2002] foi proposta. A comunicografia tem o objetivo de automatizar, totalmente ou parcialmente, a análise sobre o conjunto de textos produzidos por ferramentas de comunicação textual. Consiste em basicamente duas etapas: a transformação da comunicação em grafo e a análise da representação.

A representação do bate-papo no formato de grafo é chamada de comunicógrafo. Este faz uso da teoria dos grafos para representar as mensagens e as relações entre elas. Essa estrutura torna a comunicação mais compreensível, facilitando a análise posterior. No comunicógrafo, as mensagens são descritas a partir de um conjunto de informações que podem ser geradas pela ferramenta como usuário, e data, ou recuperadas a partir da análise da discussão, como a referência a uma mensagem anterior ou o assunto.

A representação do bate-papo no formato de grafo permite que diversos dados sejam extraídos das mensagens. Essa representação também pode ser apresentada ao usuário para que ele possa recordar a discussão e entende-la melhor.

Assim, para a elaboração do comunicógrafo, é necessário analisar e descobrir associações entre as mensagens. Em [Pimentel et al. 2002], os detalhes técnicos para identificar as associações não são discutidos. Portanto, nesse trabalho será proposta uma abordagem para isso.

Existem também diversas abordagens que utilizam softwares em que o usuário pode indicar informações complementares à mensagem, que podem facilitar a análise posterior. No Threaded Chat [Smith et al. 1999], por exemplo, o usuário indica qual

mensagem quer responder, encadeando assim as mensagens durante a comunicação. O problema dessa abordagem é que o usuário perde tempo nesta atividade de escolha da mensagem a responder.

Em [Chiru et al. 2008], foi proposta uma metodologia em que referências implícitas são verificadas automaticamente. Nesse estudo foram desenvolvidos métodos para identificar as associações entre mensagens.

Em um dos métodos foi verificado que uma mensagem com poucas palavras, concordando ou discordando, contendo no máximo uma palavra após a retirada de *stopwords*[1], tem uma grande probabilidade de se referir à última mensagem anterior a ela. Isso acontece pois uma mensagem pequena pode ser digitada rapidamente.

Usuário 1: Eu acho que hoje vai chover

Usuário 2: Eu discordo

Para que sejam identificadas ações na discussão é utilizada uma lista de padrões que são constituídos de um conjunto de palavras e aonde a referência se encontraria na mensagem. O problema é que as listas de padrões são definidas manualmente. Dessa forma, palavras comuns a eles podem ser esquecidas, fazendo com que algumas referências não sejam recuperadas corretamente.

Em [Rebedea et al. 2009], utilizando uma ferramenta que permite adicionar referencias, foram definidos os casos em que referências implícitas ocorrem. Segundo ele, os participantes só sentiam necessidade de incluir referências explicitas quando as implícitas não eram óbvias. Um caso em que a referência é óbvia é quando aparece repetição de palavras ou frases nas mensagens.

Outro padrão, encontrado em referências implícitas, são os pares de adjacência [Jefferson et al. 1974], que são duplas de mensagens que se completam logicamente.

Segundo Holmer (2008), identificar as relações entre mensagens é o ponto de partida para a análise estrutural da comunicação.

3. Proposta de algoritmos para a estruturar o diálogo

Nossa proposta compreende a utilização da revisão da literatura apresentada. como base para implementação da estratégia do comunicógrafo. Para isso foram implementados algoritmos para estruturar as mensagens geradas em bate-papos.

O texto encontrado em mensagens trocadas a partir de softwares para a comunicação textual online geralmente apresenta erros ortográficos e abreviações. Isso acontece devido à velocidade de digitação que esse meio de comunicação exige. Devido a isso, para facilitar a análise posterior dessas mensagens, é necessário que seja feito o processamento inicial do texto. Foram aplicados: corretor ortográfico, remoção de *stopwords* e radicalização. Dessa maneira, após o tratamento das mensagens, palavras com erros ortográficos ou conjugações verbais diferentes são transformadas e palavras irrelevantes para a análise são removidas.

Para a construção do comunicógrafo podem ser aplicados diversos algoritmos para identificar referências entre mensagens. Para isso definiu-se um modelo bastante

[1] Stopwords são palavras que são filtradas antes da aplicação de algoritmos de processamento de linguagem natural. Elas são palavras comuns e pequenas como as, os, de, para, entre outras.

simples para representar a discussão (figura 2). Ele é utilizado para organizar as mensagens enviadas durante a conversa.

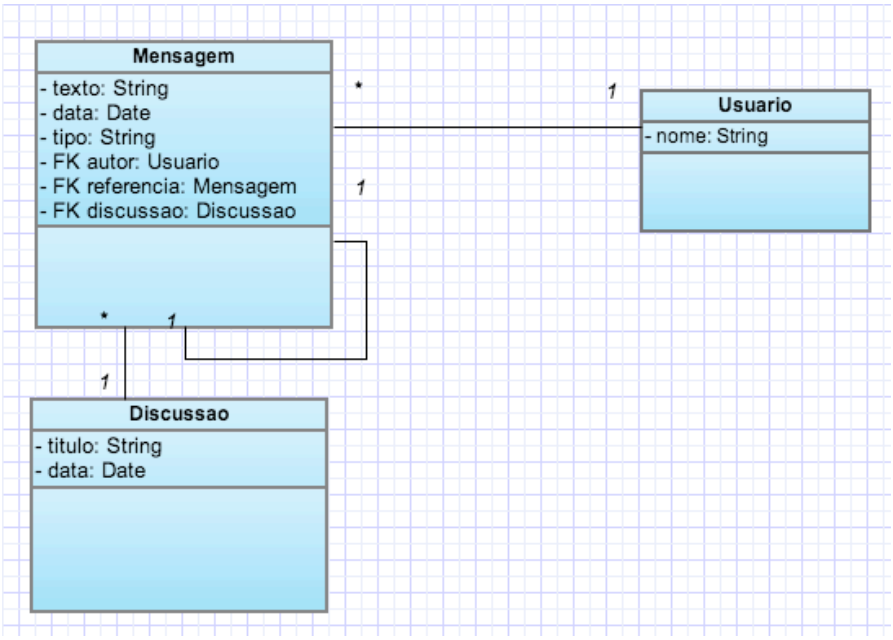

Figura 2: Modelo de dados para apoio à representação da discussão

Portanto, a partir da lista de mensagens enviadas, a ferramenta aplica os algoritmos de forma ordenada e ao identificar uma associação, salva a mensagem de referência no objeto Mensagem.

Os seguintes algoritmos foram desenvolvidos:

1. Mensagens pequenas concordando ou discordando

O algoritmo busca mensagens que têm apenas uma palavra, após a retirada de *stopwords*, e verifica se ela pertence a um padrão de concordância ou discordância. Caso pertença, é assinalada uma referência dessa mensagem para a anterior.

2. Repetição de palavras

Pares de adjacência são constituídos de duas mensagens que referenciam uma a outra e são enviadas por usuários diferentes. A sua identificação é muito importante pois permite identificar referências entre mensagens como os padrões: pergunta e resposta, cumprimento e cumprimento, entre outros.

Portanto, nosso problema é selecionar, dentre n possíveis mensagens {s1,…,sn}, a mensagem que maximiza a distribuição de probabilidade condicional [Shriberg et al, 2004]. Assim, é possível utilizar um modelo de máxima entropia para representar o problema. Para treinar o modelo, utilizamos características léxicas e estruturais. As características léxicas são informações sobre as palavras de cada mensagem. As características estruturais apresentam informações sobre o diálogo.

Abaixo estão descritas as caracteristicas utilizadas nesse trabalho. A mensagem B é a mensagem atual, para a qual estamos buscando a referência. A mensagem A é a mensagem anterior à mensagem B sendo testada.

Características estruturais:

- Número de mensagens entre A e B

Características léxicas:

- Número de palavras em A
- Número de substantivos em A
- Número de palavras em A que também existem em B
- Número de substantivos que existem em A e também existem em B
- Número de n-grams [2] presentes em A e B (n entre 2 e 4)
- Primeira e última palavra de A
- A contém o nome do usuário que enviou a mensagem B?
- B contém o nome do usuário que enviou a mensagem A?

5. Experimentação

Para o treinamento e teste dos algoritmos implementados foram utilizados 39 logs de bate-papos do gtalk e de fóruns e 17 audiências públicas coletadas no site e-democracia[3]. Esses diálogos foram transformados em XML de acordo com o esquema XSD desenvolvido. Posteriormente, os XML's foram anotados manualmente identificando os pares de adjacência.

Para treinar o modelo foi utilizada a API do Apache OpenNLP para o algoritmo de maxima entropia. A API descrita foi escolhida pela facilidade de uso e por utilizar técnicas para reduzir o tamanho dos modelos tornando mais rápida a sua leitura.

O objetivo do modelo de máxima entropia é integrar informações heterogeneas de diversas fontes para classificação. O algoritmo utiliza diversas características que devem ser descritas pelo experimentador. Elas são definidas de acordo com o conhecimento deste sobre os dados, representando as características mais importantes para a sua classificação.

Portanto, para o treinamento do modelo, foram utilizados 38 logs de bate-papos, representando 2/3 dos dados. O algoritmo desenvolvido percorre os logs, recupera os pares de adjacência anotados e as caracteristicas de cada par de mensagem. Essas características são gravadas em um arquivo de texto que será lido para criar o modelo de máxima entropia.

Finalmente, para testar os 18 logs de bate-papos restantes, para cada mensagem do log são percorridas as mensagens anteriores e o modelo criado no treinamento é utilizado para descobrir a probabilidade delas serem pares de adjacência. De acordo com os testes realizados, o algoritmo obteve uma acurácia de 63%.

Porém, essa performance ainda não é satisfatória. Devido a dificuldade de encontrar logs de bate-papos em português e anotar os pares de adjacência manualmente, o número de diálogos utilizados para treinamento e teste é escasso. Em [Shriberg et al, 2004] foram utilizadas 8135 mensagens para treinamento e teste enquanto nesse trabalho nós utilizamos 549 mensagens. Portanto, para obter um resultado melhor podemos aumentar a base de treinamento buscando e anotando mais diálogos de ferramentas textuais.

---

[2] Um n-gram é uma sequência contínua de n itens de uma sequência.

[3] e-democracia – Participação virtual, cidadania real Disponível em <http://edemocracia.camara.gov.br/> 04/04/2013

Outra medida para aumentar a eficácia do algoritmo é estudar a contribuição que cada característica tem para a recuperação de referências e adicionar novas características que tenham uma boa contribuição.

6. Conclusão

Nesse artigo foram mostradas abordagens para estruturar diálogos gerados em ferramentas de bate-papo. O algoritmo desenvolvido nesse trabalho ainda não alcançou os resultados esperados e necessita de melhorias para atender nossas necessidades. O objetivo do desenvolvimento de um algoritmo para estruturação automática das mensagens de um bate-papo é posteriormente extrair informações relevantes do diálogo para desenvolver uma ferramenta de comunicação online que melhore a qualidade de interação entre os usuários diminuindo problemas de comunicação.

Assim, como trabalhos futuros podemos citar a melhoria do algoritmo desenvolvido nesse artigo e o desenvolvimento de uma ferramenta de bate-papo online que analise o diálogo gerado ao longo da conversa e forneça informações interessantes para o usuário como dados sobre o assunto sendo discutido. Assim, é possível que dúvidas sejam solucionadas e usuários entendam melhor o assunto sendo discutido proporcionando uma conversa mais rápida e eficiente.

Referências

Galley, M.; McKeown, K.; Hirschberg, J.; Shriberg, E. (2004). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. Proc. 42nd Meeting of the ACL, pp. 669-676

Holmer, T., (2008) Discourse Structure Analysis of Chat Communication. In Language@Internet, Vol. 5.

Pimentel, M. G. E Sampaio, F. F. (2002) "Comunicografia", Revista Brasileira de Informática na Educação, v. 10, n. 1, pp. 53-59.

Pimentel, M. G., Fuks, H. E Lucena, C. J. P. (2003) "Co-text Loss in Textual Chat Tools", CONTEXT'03.

Rebedea, T.; Trausan-matu, S.; Chiru, C-G. (2008) "Extraction of Socio-semantic Data from Chat Conversations in Collaborative Learning Communities", EC-TEL, pp. 366-377.

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974) A Simplest Systematics for the Organisation of Turn-Taking for Conversation. Language 50(4i): pp. 696-735

Smith, M.; Cadiz, J. J.; Borkhalter, B. (1999) Conversation trees and threaded chats. SIGCHI Bulletin, Minneapolis, v. 31, n. 3, p. 21-23.

Trausan-Matu, S., Rebedea, T. (2009) Polyphonic Inter-Animation of Voices in VMT. In: Stahl, G. (ed.) Studying Virtual Math Teams, pp. 451-473.