

Detecção de quebras em diálogos humano-computador

Human-computer dialogue breakdown detection

Leonardo de Andrade  

Universidade de São Paulo
Escola de Artes Ciências e Humanidades

Ivandr  Paraboni  

Universidade de S o Paulo
Escola de Artes Ci ncias e Humanidades

Resumo

Com o crescimento constante no uso de tecnologias de relacionamento com o consumidor na Internet, os sistemas de *chatbot* se tornaram onipresentes no processamento de linguagem natural (PLN) e  reas relacionadas. Apesar dos avan os significativos nos  ltimos anos, no entanto, sistemas desse tipo nem sempre fornecem resultados plaus veis e consistentes, em muitos casos levando a uma quebra no di logo. Assim, h  grande interesse em investigar as circunst ncias nas quais erros deste tipo s o produzidos e, quando poss vel, aprimorar o projeto destes sistemas de modo a minimizar tais erros. Com base nestas observa es, neste trabalho abordamos a quest o da detec o autom tica de quebras em di logos humano-computador apresentando tr s modelos que levam em considera o o hist rico de di logo para decidir quando ele possui maior probabilidade de culminar em uma quebra. Os modelos propostos exploram uma variedade de m todos de PLN recentes, e s o avaliados tanto com base em um conjunto de dados de di logos reais em portugu s entre usu rios humanos e sistemas de *chatbot* desenvolvido especificamente para este fim, como tamb m utilizando *benchmarks* publicamente dispon veis para o idioma ingl s.

Palavras chave

classifica o textual, detec o de quebras em di logos, chatbots

Abstract

With the steady growth in the use of consumer relationship technologies on the Internet, chatbot systems have become ubiquitous in Natural Language processing (NLP) and related fields. Despite significant advances in recent years, however, systems of this kind do not always deliver plausible, consistent results, in many cases leading to a dialogue breakdown. As a result, there is a growing interest in how to improve systems of this kind so as to minimise errors. Based on these observations, this work addresses

the issue of automatic dialogue breakdown detection by presenting three models that take the dialogue history into account to decide when a conversation is likely to break. The models under consideration explore a range of recent NLP methods and are evaluated by using a purpose-built Portuguese dataset conveying real-world human-computer conversations, and also in publicly available benchmarks for the English language.

Keywords

text classification, dialogue breakdown detection, chatbots

1. Introdu o

Chatbots s o sistemas computacionais que interagem com usu rios humanos usando l ngua natural, atuando como uma interface para provedores de dados e servi os (Brandtzaeg & F lstad, 2017). Sistemas deste tipo podem fazer uso de m todos de Gera o de L ngua Natural para produ o de mensagens sem participa o de autores humanos e altamente customizadas, ou m todos de Recupera o de Informa o para selecionar mensagens a partir de uma base de textos predefinidos.   comum tamb m que estes sistemas fa am uso de avatares ou outros recursos com maior ou menor semelhan a com um ser humano (Chiar in & Chasaide, 2016), objetivando aumentar o n vel de engajamento dos usu rios.

Di logos com *chatbots* podem ser imperfeitos, e em certos casos o usu rio pode n o receber a resposta que estava procurando e ser for ado a tentar outras formula es at  que, eventualmente, obtenha a informa o desejada ou desista da intera o (Sandbank et al., 2018). Assim, define-se uma quebra em um di logo como sendo um ponto espec fico em que o usu rio humano n o tem mais condi es de prosseguir com o di logo (Martinovski & Traum, 2003).

Um exemplo de quebra deste tipo, adaptado a partir de exemplos do *cópus* a ser discutido na Seção 3, é apresentado a seguir.

Sistema

Olá, tudo bem? Como posso te ajudar?

Cliente

Olá, ontem paguei parcelado minha fatura em atraso, mas recebi sms que ainda está em atraso. Vc pode verificar?

Sistema

Não entendi sua pergunta! Por favor escreva com outras palavras

Cliente

Parcelamento da fatura

Sistema

Para parcelar sua fatura, faça...

Neste exemplo, observa-se que a partir de um erro de interpretação inicial o sistema acaba desviando a conversa para um tópico diferente do que era a intenção do cliente, o que pode culminar em uma quebra no diálogo mais adiante caso o cliente decida que não vale a pena continuar tentando. Antes que esse ponto seja alcançado, entretanto, o diálogo pode apresentar sintomas de desarranjo ou pequenos incidentes de insatisfação (neste caso, a própria falha de interpretação) que, de forma cumulativa, levam ao colapso final. Ao acompanhar esses incidentes menores e estudar as suas causas, podemos identificar oportunidades de melhoria na interação humano-computador e potencialmente aprimorar o sistema para que erros deste tipo não ocorram, ou ocorram com menor frequência.

Antecipar —e então tentar evitar— possíveis quebras em diálogos é uma tarefa importante para o aprimoramento de sistemas de *chatbot*, e uma linha de pesquisa ativa no Processamento de Língua Natural (PLN). O problema tem sido inclusive tema de desafios computacionais (ou ‘shared tasks’) da série *Dialogue Breakdown Detection Challenge* (DBDC) para os idiomas inglês e japonês (Higashinaka et al., 2017, 2019), e é destas competições que tomamos a própria definição do problema computacional a ser tratado, possivelmente pela primeira vez, em língua portuguesa. De acordo com esta definição, uma série de interações entre um *chatbot* e um usuário humano (como no exemplo acima) pode ser classificada de três formas: como uma situação que certamente levaria a uma quebra (B=*breakdown*), como uma situação que não levaria a uma quebra (NB=*no breakdown*), e ainda como um caso intermediário em que a sequência poderia ou não levar a uma quebra (PB=*possible breakdown*).

De modo geral, estudos existentes na área de detecção de quebra em diálogos levam em conta uma quantidade limitada de informação para decidir se há quebra ou não, muitas vezes considerando apenas um par pergunta-resposta de cada vez. Como alternativa a este tipo de abordagem, o presente trabalho apresenta três modelos de detecção de quebra em diálogos para o português que levam em conta o histórico (ou memória) da conversa, e que exploram diferentes métodos de sucesso recente em outras tarefas de PLN, a saber: o uso de *embeddings* estáticos (Mikolov et al., 2013; Pennington et al., 2014) e sensíveis ao contexto (Devlin et al., 2019), e o uso de modelos neurais baseados em *gate recurrent units* (GRUs). Para este fim, os modelos em questão fazem uso de um conjunto de diálogos reais produzidos por *chatbots* brasileiros, e quer foram coletados especificamente para este projeto. Além disso, como forma de demonstrar o poder de generalização dos modelos propostos, é conduzida também uma avaliação aos moldes das competições DBDC, utilizando para este fim *benchmarks* de diálogos em inglês publicamente disponíveis para este idioma.

O restante deste documento está organizado da seguinte forma. A Seção 2 discute trabalhos relacionados na área de detecção de quebras em diálogos. A Seção 3 descreve a construção do *cópus* de diálogos em português e os modelos de detecção de quebras propostos. A Seção 4 descreve o procedimento de avaliação destes modelos com base em diálogos em português e inglês, e a Seção 5 apresenta os resultados propriamente ditos. Finalmente, a Seção 7 resume as contribuições deste trabalho e oportunidades de melhorias futuras.

2. Trabalhos relacionados

No presente trabalho, a detecção de quebras em diálogos é vista uma tarefa de aprendizado de máquina supervisionado baseada em dados textuais (diálogos) anotados com pontos em que houve quebra nos moldes da série de desafios *Dialogue breakdown detection challenge*, ou DBDC (Higashinaka et al., 2016, 2017, 2019), conforme discutido na Seção anterior. Este problema se distingue, por exemplo, da detecção de quebras em diálogos em língua falada (Black & Eskenazi, 2009), da detecção de estados do diálogo (Williams et al., 2013), e também do caso de diálogos orientados a uma tarefa específica (Bear et al., 1992; Carpenter et al., 2001; Bulyko et al., 2005) por apresentar maior variedade de quebras possíveis (Higashinaka et al., 2016).

Uma parte considerável dos estudos de interesse para o presente estudo é assim organizada em torno da série de desafios DBDC e dos conjuntos de dados rotulados que estes eventos disponibilizam, denominados córpus DBDC3 e DBDC4¹. Uma visão geral dos desafios DBDC e dos córpus a eles associados é apresentada na Seção 2.1. Modelos computacionais para detecção de quebras em diálogos humano-computador, utilizando estes ou outros córpus, são revisados na Seção 2.2.

2.1. Os desafios DBDC e córpus

Os desafios DBDC3 (Higashinaka et al., 2017) e DBDC4 (Higashinaka et al., 2019) são competições (ou *shared tasks*) de sistemas de detecção automática de quebras em diálogos entre humanos e *chatbots*. Nestes eventos foram produzidos dois córpus de mesmo nome que se tornaram influentes na área, e que consistem de coleções de diálogos em inglês e japonês rotuladas com informação de quebras (B), não-quebras (NB) e possíveis quebras (PB). Com o possível intuito de facilitar a condução da tarefa da competição, os rótulos das categorias B, BP e PB ocorrem de forma balanceada em ambos os córpus, e portanto estes conjuntos de dados não necessariamente correspondem a um uso normal de sistemas deste tipo. No presente trabalho, apenas as porções em inglês serão discutida, respeitando-se a mesma divisão de treino e teste proposta nas competições de origem.

O córpus DBDC3 (Higashinaka et al., 2017) foi rotulado por um grupo de 30 anotadores em múltiplas rodadas de análise de concordância entre juízes. A porção em inglês deste córpus é composta de quatro coleções de diálogos de propósito geral entre *chatbots* e voluntários humanos e/ou recrutados por *crowd sourcing*, a saber: TKTK-100, de 100 sessões do conjunto do WOCHAT TickTock (Yu et al., 2016); IRIS-100, de 100 sessões do conjunto do WOCHAT IRIS (Banchs & Li, 2012); CIC-115, de 115 diálogos do *Conversational Intelligence Challenge*²; e YI-100, de 100 diálogos com um robô do Instituto de Física e Tecnologia de Moscou³. A Tabela 1 apresenta estatísticas descritivas do córpus DBDC3 inglês em cada um destes conjuntos.

O desafio DBDC3 contou com oito equipes participantes, sendo que seis delas trabalharam exclusivamente com a porção de dados em inglês. Dentre estes, tiveram mais destaques os estudos

de Iki & Saito (2017); Lopes (2017); Kato & Sakai (2017); Sugiyama (2019); Takayama et al. (2019), que são discutidos na Seção 2.2.

Para a edição seguinte do evento, denominada DBDC4 (Higashinaka et al., 2019), foi construído um novo córpus com características similares, ou seja, mantendo-se os mesmos dois idiomas e definições de classes, porém desta vez rotuladas por um time de apenas 15 anotadores. A porção em inglês do córpus DBDC4 consiste de diálogos produzidos pelo sistema IRIS (Banchs & Li, 2012), e diálogos produzidos por seis *chatbots* não especificados, denominados apenas como bot01..06, a partir do conjunto de dados ConvAI2⁴. A Tabela 2 apresenta estatísticas descritivas do subconjunto de desenvolvimento deste córpus, tal qual definido na competição.

O desafio DBDC4 contou com quatro equipes participantes, cujas abordagens e resultados são descritos por Sugiyama (2021); Shin et al. (2019); Hendriksen et al. (2021); Wang et al. (2019) e discutidos na Seção 2.2 a seguir. Todas equipes trabalharam com a porção de dados em inglês, e duas delas trabalharam também com o conjunto de dados em japonês (não considerado no presente trabalho).

2.2. Detecção automática de quebras em diálogos

A área de detecção automática de quebras em diálogos tem apresentado grande crescimento em anos recentes, possivelmente influenciado pela própria organização da série de desafios DBDC (Higashinaka et al., 2017, 2019). Alguns dos estudos deste tipo mais diretamente relacionados ao presente trabalho são sumarizados na Tabela 3, com informações sobre a língua-alvo dos diálogos considerados, o tipo de representação textual (e.g., *embeddings* de palavras, documentos ou sentenças, *part-of-speech* (POS) etc.), método de aprendizado de máquina (AM) adotado e, quando pertinente, a posição geral no ranque de sistemas participantes das competições DBDC3 e DBDC4 para as tarefas em inglês em primeira execução com base na medida F_1 relatada.

Os estudos aqui descritos foram identificados por meio de uma revisão exploratória da literatura, partindo-se dos próprios relatórios das competições DBDC, dos artigos publicados pelos seus participantes durante e após a participação no evento, bem como de suas próprias referências. Além disso, face ao grande número de submissões (cada participante podia submeter

¹A edição mais recente do evento, denominada DBDC5, ainda não disponibilizou o conjunto de dados utilizado.

²<https://convai.io/data/>

³<https://www.slideshare.net/sld7700/>

⁴<https://github.com/DeepPavlov/convai/tree/master/data>

	TKTK	IRIRS	CIC	YI
Di�logos	210	210	225	210
N�o quebra	38,6%	33,5%	29,0%	35,0%
Poss�vel quebra	28,2%	28,4%	33,1%	37,7%
Quebra	33,2%	38,1%	37,9%	27,3%

Tabela 1: Estat sticas descritivas do c rpus DBDC3 ingl s em Higashinaka et al. (2017).

	bot1	bot2	bot3	bot4	bot5	bot6	IRIS
Total de di�logos	39	38	42	41	2	6	43
N�o quebra	40,4%	40,8%	35,8%	39,9%	22,0%	16,4%	30,0%
Poss�vel quebra	29,4%	26,8%	29,5%	29,4%	37,0%	22,6%	30,4%
Quebra	30,2%	32,4%	34,7%	30,7%	41,0%	61,0%	39,6%

Tabela 2: Total de di logos e distribui o de frases do sistema por classe no c rpus DBDC4 ingl s (desenvolvimento) (Higashinaka et al., 2019).

at  tr s execu es), foram selecionados apenas os sistemas de maior destaque em cada competi o com base nos crit rios de avalia o considerados.

Com base neste levantamento, observa-se uma predomin ncia de estudos no idioma ingl s, uso de *embeddings* de palavras, e m todos de aprendizado neural como LSTM e BERT. Os estudos de Iki & Saito (2017); Lopes (2017); Kato & Sakai (2017); Sugiyama (2019); Takayama et al. (2019) descrevem sistemas participantes da competi o DBDC3 ou vers es aprimoradas destes, e os estudos de Sugiyama (2021); Wang et al. (2019); Shin et al. (2019); Hendriksen et al. (2021) s o relativos   competi o DBDC4. Posteriormente, os estudos de Almansor et al. (2021); Ng et al. (2020b) apenas reutilizaram estes (e outros) conjuntos de dados de forma independente. Detalhes adicionais s o discutidos a seguir.

2.2.1. Participantes da competi o DBDC3

O estudo de (Iki & Saito, 2017)   motivado pela observa o de que o hist rico de di logos humano-computador frequentemente inclui um grande n mero de palavras n o observadas durante o treinamento do modelo, o que pode impactar a qualidade da conversa. Como forma de contornar esta dificuldade,   proposto o uso de redes neurais *End-to-End* do tipo MemN2N (Sukhbaatar et al., 2015) em conjunto com representa es sentenciais baseadas em *embeddings* de caracteres, e uma rede do tipo CNN para o m dulo de aten o. O sistema final, denominado Pleco, obteve os melhores resultados de medida F_1 na competi o DBDC3 em ingl s considerando-se a tarefa de detec o da classe quebra (B). Na tarefa de detec o de poss vel quebras ou quebras (PB+B), entretanto, o sis-

tema ainda ficou (assim como todos os demais participantes) abaixo do *baseline* de classe majorit ria da competi o. Este sistema ser  utilizado como *baseline* tamb m em nossos experimentos relacionados ao c rpus DBDC3, descritos na Se o 3.

O trabalho de Lopes (2017) objetivou investigar a poss vel generaliza o do problema de detec o de quebras em di logos orientados a tarefas ou n o, comparando duas abordagens: uma baseada em um conjunto reduzido de atributos orientados a tarefas e classificadores SVM, e a outra (puramente textual) usando *embeddings* de senten a e RNNs. De modo geral, a abordagem n o orientada   tarefa apresenta melhor desempenho, obtendo a segunda melhor medida F_1 para a tarefa em ingl s da competi o DBDC3, e a melhor medida de precis o dentre os sistemas participantes.

O estudo de Sugiyama (2017), originalmente submetido   competi o DBDC3 apenas para a tarefa em japon s,   alargado em 2019 para contemplar tamb m a tarefa em ingl s. Neste caso, os modelos propostos objetivaram estimar o grau de adequa o da transi o de t picos a cada par pergunta-resposta do di logo, e obtiveram os melhores resultados da competi o neste idioma. De forma mais espec fica, foram computadas diversas caracter sticas textuais, como quantidade de palavras em comum entre pergunta e resposta, m tricas de similaridade variadas, tamanho das senten as em n mero de palavras e caracteres, quantidade de intera es, *embeddings* de senten as utilizando codifica o do tipo seq2sec, contagem de termos interrogativos, dist ncia em rela o a perguntas similares, contagens IDF e palavras de conte do abstrato. Como m todos de aprendizado, foi em-

Referência	Língua	Representação textual	Método	DBDC3	DBDC4
Iki & Saito (2017)	En	<i>character emb.</i>	CNN	#1	
Lopes (2017)	En	<i>word/document emb.</i>	LSTM, SVM	#2	
Sugiyama (2017)	Jp	<i>sentence emb.</i> , POS, TF-IDF	<i>Ensemble</i>		
Kato & Sakai (2017)	En	<i>word emb.</i> , TF-IDF	Similaridade		
Takayama et al. (2017)	Jp	<i>word emb.</i>	LSTM, CNN		
Sugiyama (2021)	En	<i>word emb.</i> , POS, TF-IDF	BERT		#1
Wang et al. (2019)	En	<i>word/sentence emb.</i>	<i>RF</i> , LSTM		#2
Shin et al. (2019)	En	<i>sentence emb.</i>	BiLSTM		
Hendriksen et al. (2021)	En	<i>word emb.</i>	LSTM		
Almansor et al. (2021)	En	TF-IDF, sentimento	<i>Ensemble</i>		
Ng et al. (2020b)	En	<i>word emb.</i>	BERT		

Tabela 3: Estudos recentes de detecção de quebra em diálogos humano-computador.

pregado um *ensemble* do tipo pilha de regressores (*Stack regressors* (van der Maaten & Hinton, 2008)) baseado em *Random Forest* (RF), *Extra-trees* (ETR), *K-nearest Neighbor* (KNN), *Gradient Boosting* (GBR) e *Support Vector* (SVR). O regressor ETR é o utilizado no nível superior da pilha para combinar as predições dos demais modelos.

O estudo de Kato & Sakai (2017) segue a abordagem de Sugiyama (2017) e também utiliza regressores ETR e outros métodos para estimar a média e a variância da distribuição de quebras, e então derivar as probabilidades de quebra a partir dessas estimativas. Para cálculo da similaridade entre *embeddings* de duas sentenças, melhores resultados foram observados utilizando-se a similaridade de cosseno entre todos os pares de termos. Esta abordagem seria posteriormente aprimorada e rerepresentada à competição DBDC4 com melhores resultados (Wang et al., 2019).

O estudo de Takayama et al. (2019) tem como foco a questão do viés de anotação de quebras em diálogos, e estende a submissão (Takayama et al., 2017) originalmente apresentada à competição DBDC3 para detecção de quebras em diálogos apenas em japonês. O estudo propõe uma abordagem para detecção de quebras que explora diferenças entre anotadores, na qual os dados de treinamento são agrupados de acordo com a distribuição de anotações, e então utilizados para treinar detectores específicos para cada agrupamento. A classificação é realizada com uso de um modelo de *embeddings* de palavras e redes do tipo LSTM e CNN combinadas em uma arquitetura do tipo *Ensemble* para a predição final de quebras.

2.2.2. Participantes da competição DBDC4

O estudo de Sugiyama (2021) apresenta uma abordagem que combina atributos tradicionais de

diálogo propostos em estudos prévios (Sugiyama, 2017, 2019) e outras, e modelo de língua pré-treinado BERT (Devlin et al., 2019). A proposta apresentou o melhor resultado de medida F_1 global (considerando possíveis quebras e quebras, ou PB+B) e a melhor acurácia da competição DBDC4 para o inglês, e os melhores resultados globais de classificação para o japonês. Este sistema, denominado NTTCS19, será utilizado como *baseline* também em nossos experimentos relacionados ao córpus DBDC4, descritos na Seção 3.

O estudo por Wang et al. (2019) estende a abordagem de Kato & Sakai (2017), originalmente apresentada na competição DBDC3, para a edição DBDC4 com diversas melhorias, incluindo a substituição do regressor ETR por *Random Forest*, e a predição direta das probabilidades dos rótulos em vez de estimar sua média e variância. A proposta utiliza um modelo de LSTM adaptado de Lopes (2017) com uso de uma CNN adicional para extração de características. Dentre várias arquiteturas consideradas, uma solução baseada em um *ensemble* de árvore de decisão e múltiplos modelos do tipo LSTM apresentou o segundo melhor resultado de medida F_1 em primeira execução na competição DBDC4 em inglês.

O estudo de Shin et al. (2019) utiliza redes bidirecionais LSTM (BiLSTM) com mecanismo de atenção global e *embeddings* BERT para detecção de quebras no córpus DBDC4 inglês, além de um mecanismo de atenção local para lidar com casos de quebra raros. Os melhores resultados são observados na detecção de quebras próximas ao fim do diálogo, sugerindo que a disponibilidade de mais informação contextual facilita a tarefa. Apesar do uso de métodos mais sofisticados do que os de vários outros participantes, entretanto, o sistema não alcançou resultados competitivos.

O estudo de Hendriksen et al. (2021) compara uma gama de modelos LSTM (*vanilla*, empilhado e bidirecional) e tipos de *embeddings* (Word2Vec e GloVe de diferentes origens) para detec o de quebras no c rpus DBDC4 ingl s. Os melhores resultados foram observados na configura o que usa a LSTM do tipo *vanilla* com *embeddings* GloVe *Common Crawl*, mas ainda assim inferiores aos de outros sistemas participantes.

2.2.3. Outras abordagens

Posteriormente  s competi es DBDC3/4, dois estudos relacionados s o ainda dignos de nota. O estudo de Almansor et al. (2021) utiliza m todos de an lise de sentimentos para detectar mudan as indicativas de quebra ou poss vel quebra em di logos em sistemas de atendimento ao consumidor. O modelo proposto utiliza um l xico afetivo e contagens TF-IDF para classificar o sentimento associado a cada intera o como sendo positivo, neutro ou negativo, utilizando um *ensemble* de classificadores do tipo *Multinomial Naive Bayes*, *Bernoulli Naive Bayes*, regress o log stica e SVM. Resultados observados no c rpus DBDC3 s o superiores aos do *baseline* CRF da competi o para o caso de quebra individual, mas ainda inferior no caso de soma das quebras e poss veis quebras.

Finalmente, o estudo de Ng et al. (2020b) investiga o uso de m todos de aprendizagem semi-supervisionada para aprimorar a detec o de quebras em di logos, incluindo pr -treinamento cont nuo em um conjunto de dados da rede social Reddit e um m todo de aumento de dados baseado em m ltiplas dobras (Ng et al., 2020a). O conjunto de dados aumentado   utilizado em um modelo de classifica o composto de um m dulo BERT e um classificador *Multilayer Perceptron* (MLP). O modelo proposto obteve os melhores resultados na recente competi o DBDC5⁵, sendo 12% superior aos sistemas de *baseline* e outros participantes. Os resultados para o c rpus DBDC4 n o s o entretanto diretamente compar veis com os de outros sistemas porque a m trica de avalia o utilizada foi a medida F_1 da classe majorit ria, e n o das classes quebra e quebra + poss vel quebra originalmente adotadas por Higashinaka et al. (2019).

3. Materiais e m todos

O presente trabalho consiste da cria o de um novo conjunto de dados em portugu s brasileiro

⁵<http://workshop.colips.org/wochat/@iwsds2020/shared.html>

contendo di logos humano-computador rotulados com informa o de quebras, e da proposta e avalia o de tr s novos modelos de detec o de quebras em di logos humano-computador que levam em conta o hist rico (ou mem ria) da conversa. Estes dois itens s o descritos individualmente a seguir, e o c digo desenvolvido para este fim encontra-se dispon vel para re so⁶.

3.1. Constru o do c rpus DBDBR portugu s

Conforme discutido na Se o 2.2, existem conjuntos de dados publicamente dispon veis para estudo de problemas de detec o de quebras em di logos nos idiomas ingl s e japon s. No caso do idioma portugu s, entretanto, n o foram identificados recursos semelhantes. Al m disso, observa-se que a anota o de grandes massas de dados deste tipo representa um custo consider vel, tipicamente envolvendo um grande n mero de ju zes e problemas de concord ncia. Com base nestas observa es, optou-se assim por efetuar a constru o de um c rpus de di logos reais em portugu s entre humanos e *chatbots*, aqui denominado DBDBR, contendo quebras sinalizadas pelos pr prios usu rios e contornando assim a necessidade de anota o manual por terceiros.

O c rpus DBDBR foi constru do a partir de dados cedidos por uma empresa brasileira que comercializa um sistema de *chatbot* de atendimento para seus clientes, e com a qual o primeiro autor desta pesquisa mant m v nculo profissional. Por meio deste v nculo, foi obtida permiss o de uso de parte dos dados gerados pelo sistema.

Al m do conjunto de mensagens propriamente dito, cada di logo pode incluir informa es de *feedback* do usu rio, que tem a op o de sinalizar sua insatisfa o com uma resposta usando conceitos como ‘N o foi isso que eu perguntei’ e ‘Resposta incorreta’. No presente trabalho, respostas associadas a este tipo de *feedback* negativo s o tomadas (ainda que de forma aproximada) como pontos de quebra no di logo, o que contorna a necessidade de uma anota o manual de alto custo. Diferentemente dos c rpus DBDC3/4 descritos na Se o 2.1, entretanto,   importante observar que o presente m todo s  oferece a distin o bin ria entre quebra e n o quebra, ou seja, n o existe a classe intermedi ria (poss vel quebra).

O c rpus contempla dados em tr s dom nios de di logo que apresentaram o maior volume de intera es em um per odo de dois meses: uma provedora de TV por assinatura, um banco e uma corretora. Di logos na  rea de TV por assina-

⁶<https://github.com/landrady/DialogBreakdown>

tura incluem dúvidas sobre instalação de equipamentos, agendamento de serviço técnico, contratação de pacotes e outros; diálogos da área bancária incluem dúvidas sobre empréstimos, prazos de cartões, limites, emissão de boleto e outros; e diálogos na área de corretora incluem dúvidas sobre investimentos, cancelamentos de operações, juros, uso de cartão de crédito etc. Em cada domínio, foram extraídos aproximadamente 10.000 diálogos de forma aleatória.

A Tabela 4 apresenta estatísticas descritivas de cada domínio do cópuz coletado.

3.2. Modelos propostos

Seguindo o trabalho de Higashinaka et al. (2017, 2019), no presente trabalho a detecção de quebras nos cópuz DBDC em inglês será definida como um problema de classificação ternária (quebra, possível quebra e não-quebra). Para o caso do cópuz DBDBR em português, entretanto, a tarefa será definida como um problema de classificação binária dado que o cópuz não possui rótulos de ‘possível quebra’.

Em ambos os casos, a detecção de quebras será investigada considerando-se três estratégias que levam em conta um histórico mais amplo da conversa. Estas estratégias, denominadas RegW2V, RegBERT e GruGloVe, foram escolhidas por nos permitir explorar métodos alternativos de aprendizado (em especial, do tipo regressão e baseados em *Gate Recurrent Units*), e representações textuais de *embeddings* estáticos e sensíveis ao contexto, conforme detalhado a seguir. Exceto quando indicado, a definição dos valores ótimos para os hiper-parâmetros de cada modelo foi realizada por meio de um procedimento de *grid search* a ser detalhado na Seção 4.

O modelo RegW2V objetiva representar uma estratégia de classificação textual padrão baseada em *embeddings* estáticos aplicada à detecção de quebras em diálogos. Para este fim, o modelo utiliza um método de aprendizado do tipo *Gradient Boosting* (Friedman, 2001) e *embeddings* do tipo Word2Vec (Mikolov et al., 2013). O modelo recebe como entrada a concatenação de dois tipos de informação: (i) dois vetores representando o par de sentenças usuário-sistema, e (ii) dois vetores representando a memória do diálogo, que é o conjunto de perguntas ou respostas das cinco últimas interações usuário-sistema. As sentenças usuário-sistema (i) são representadas como vetores de contagens TF-IDF reduzidas com uso de *Principal Component Analysis* (PCA), e a memória (ii) é representada por um vetor de *embeddings* médios de 300 dimensões do

tipo Skip-gram pré-treinados em português, obtidos de Hartmann et al. (2017), e inglês⁷.

O modelo RegBERT objetiva constituir uma solução mais sofisticada para o problema na qual os *embeddings* estáticos são substituídos por *embeddings* sensíveis ao contexto. RegBERT é em grande parte semelhante a RegW2V, porém utilizando representações textuais do tipo BERT (Devlin et al., 2019) de 257 dimensões para o português, obtidos por Souza et al. (2020), e de 77 dimensões para o inglês⁸. Estes parâmetros, de grande impacto no tempo de treinamento de modelos BERT, foram escolhidos com base no tamanho médio das sentenças em cada cópuz.

Finalmente, o modelo GruGloVe objetiva representar uma solução de tratamento mais tradicional para a noção de histórico da conversa, baseada na classificação de sequências implementada com uso de uma rede neural baseada em *Gate Recurrent Units* (GRUs) e *embeddings* estáticos do tipo GloVe (Pennington et al., 2014). A rede recebe como entrada dois tipos de informação: (i) sequências de *embeddings* de sentenças do usuário e do sistema, em ambos os casos compostos da média dos *embeddings* de 150 palavras, e (ii) características não textuais adicionais representando o tamanho médio das sentenças do usuário e do sistema e o identificador do diálogo. Estas informações são fornecidas à primeira camada da rede em blocos de 10 interações cada. A seguir, os blocos de sentenças do usuário e do sistema são combinados em dois vetores médios representando os dois participantes do diálogo (i.e., humano e computador), que são então fornecidos a três camadas recorrentes do tipo GRU concatenadas ao conjunto de características não textuais, e a duas camadas densas do tipo ReLU e Softmax, respectivamente. Esta arquitetura é ilustrada na Figura 1.

Dado que as classes são desbalanceadas, as probabilidades obtidas pelos modelos de regressão (no intervalo entre zero e um) são convertidas em classes nominais ordenadas (não quebra, possível quebra e quebra) com a definição de pontos de corte ajustados para cada intervalo. Mais especificamente, foi construído um modelo do tipo floresta aleatória para cada tarefa de classificação usando os parâmetros fixos de profundidade máxima 4, com até 3 folhas, pesos de classes balanceados e estado aleatório zero. A partir da árvore de cada tarefa, foram identificados os intervalos de probabilidade que correspondem a cada classe nominal.

⁷<https://code.google.com/archive/p/word2vec/>

⁸<https://huggingface.co/bert-base-cased>

M�tricas	TV	Banco	Corretora
Quantidade de di�logos	9.990	9.988	9.973
Usu�rios �nicos	9.936	9.813	9.080
M�dia de palavras do consumidor	4,64	3,74	9,7
M�dia de palavras do <i>chatbot</i>	21,12	18,60	18,49
M�dia de intera��es por di�logo	7,32	7,65	6,32
Tamanho do vocabul�rio do consumidor	11.088	9.728	15.902
Tamanho do vocabul�rio do <i>chatbot</i>	30.367	18.556	23.973
Quantidade de quebras	7.932	6.131	7.044
Quantidade de n�o quebras	140.126	146.832	121.968

Tabela 4: Estat sticas descritivas do c rpus DBDBR portugu s.

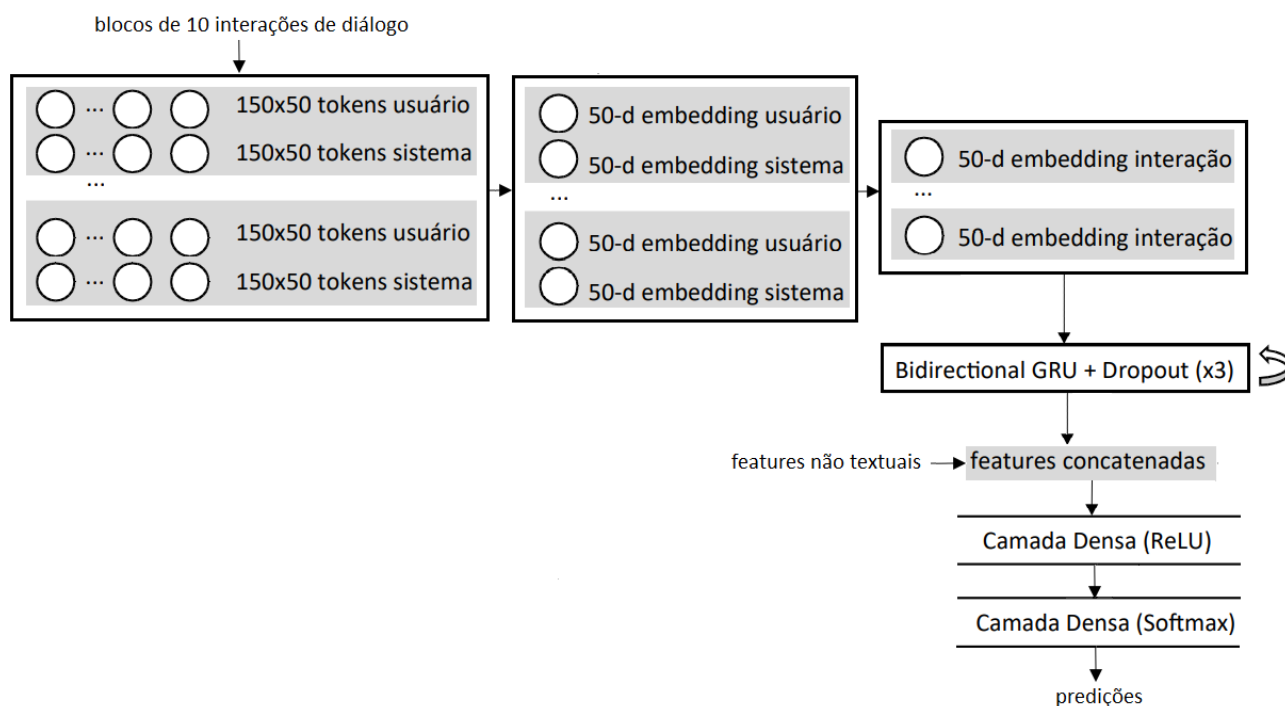


Figura 1: Arquitetura do modelo GruGloVe.

Um exemplo desta representa  o   apresentado na Figura 2, ilustrando o caso da  rvore de decis o gerada para o modelo RegW2V com base nos dados do c rpus DBDC3. Com base na probabilidade $X[0]$ do modelo de regress o subjacente,   exibido o n mero de inst ncias das classes (*value*) n o quebra, poss vel quebra e quebra, respectivamente, e o  ndice gini associado a cada n  do da estrutura. O mesmo procedimento foi utilizado para a obten  o dos r tulos de classe nas demais tarefas de classifica  o aqui discutidas.

4. Avalia  o

oram conduzidos experimentos para avaliar seu desempenho na tarefa de detec  o de quebras nos di logos no c rpus DBDBR em portugu s (cf. Se  o 3.1) e nos c rpus DBDC3 (Higashinaka et al., 2017) e DBDC4 (Higashinaka et al.,

2019) do ingl s. O objetivo do experimento foi assim o de identificar a melhor estrat gia computacional para cada cen rio de avalia  o, e ilustrar a aplica  o destes modelos a di logos em portugu s.

Para os c rpus DBDC3/4, foi utilizada a mesma divis o de treino/teste seguida nas respectivas competi  es de modo a permitir uma compara  o direta com sistemas existentes, e para o c rpus DBDBR utilizou-se uma divis o aleat ria   propor  o 70/30. Em todos os casos, a por  o de teste foi reservada para avalia  o final dos modelos treinados.

Ao contr rio dos conjuntos de dados em ingl s, observa-se que o c rpus portugu s (DBDBR)   fortemente desbalanceado, com um n mero de n o-quebras v rias vezes superior ao n mero de quebras. Isso ocorre porque as avalia  es destes di logos s o feitas pelos pr prios usu rios do

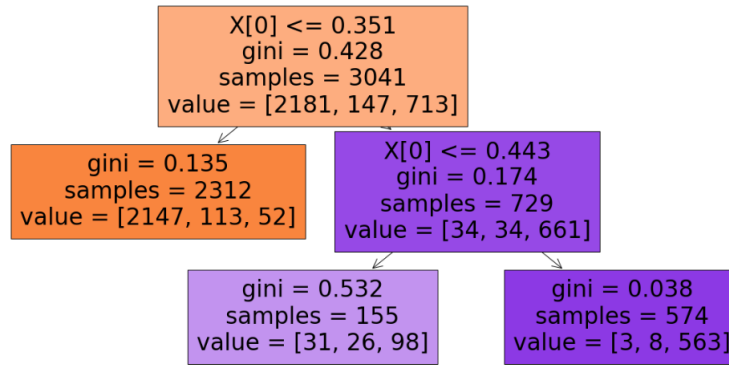


Figura 2: Árvore de decisão para o modelo RegW2V para o córpus DBDC3

serviço que, na maior parte das vezes, não fornece nenhuma resposta, o que configura o rótulo ‘sem quebra’. Como forma de reduzir este desbalanceamento — permitindo assim a observação de diferenças mais expressivas entre os modelos avaliados— os experimentos a serem conduzidos utilizam apenas um subconjunto de cerca de 50% das instâncias da classe ‘não quebra’, selecionadas aleatoriamente a partir do córpus original. A Tabela 5 apresenta o número de instâncias para cada classe, conjunto e córpus.

Córpus	Conjunto	NB	PB	B
DBDBR	treino	134.108	-	14.774
	teste	57.476	-	6.333
DBDC3	treino	1.414	1.834	1.087
	teste	846	479	764
DBDC4	treino	1087	443	766
	teste	1239	446	481

Tabela 5: Instâncias de treino e teste para classes não quebra (NB), possível quebra (PB) e quebra (B).

No caso do modelo GruGloVe, a otimização foi feita considerando-se a variação de pesos validados pela função do erro médio quadrático (MSE). Para os modelos de regressão RegW2V e RegBERT, foi utilizado um otimizador de hiperparâmetros aleatórios com os atributos tamanho de janela (1 a 15), valor mínimo de DF para TF-IDF (DFmin, de 1 a 6), valor máximo de DF (DFmax, de 0,5 a 1), quantidade de estimadores (est, de 1 a 150), taxa de aprendizagem (lr, de 0,001 a 0,1) e profundidade máxima (P, de 3 a 6). Os valores destes hiperparâmetros que obtiveram resultado ótimo de medida F_1 em cada conjunto de treinamento são sumarizados na Tabela 6.

Para fins de avaliação, os modelos propostos são comparados ao *baseline Conditional Random Fields* (CRF) de Higashinaka et al. (2017), e também com o sistema de melhor resultado em

primeira execução em cada competição. No caso do conjunto DBDC3, isso corresponde ao sistema Pleco descrito por Iki & Saito (2017), e no caso do conjunto DBDC4 corresponde ao sistema NTTCS19 de Sugiyama (2021). Para o conjunto DBDBR, apenas o *baseline* CRF foi considerado. Os resultados dos modelos propostos foram computados utilizando-se a ferramenta de avaliação oficial de cada evento, e os resultados dos demais sistemas foram extraídos dos respectivos relatórios de Higashinaka et al. (2017, 2019).

Na avaliação propriamente dita foram consideradas as métricas de medida F_1 da classe ‘quebra’ $F_1(B)$ e, com exceção do córpus DBDBR (que não possui possíveis quebras), também a medida F_1 da soma das classes ‘possível quebra’ e ‘quebra’ $F_1(PB + B)$ propostas por Higashinaka et al. (2017). Todos os resultados são referentes à porção inédita de teste de cada córpus, à qual nenhum dos modelos teve acesso durante o treinamento.

5. Resultados

A seguir são apresentados os resultados individuais obtidos para cada córpus, seguidos da avaliação da sua significância estatística e de uma breve análise de erros.

5.1. Resultados principais

A Tabela 7 sumariza os resultados para a classe ‘quebra’ (B) no córpus DBDBR português (Seção 3.1), com o melhor resultado em destaque.

Observa-se que o modelo RegW2V é superior às alternativas consideradas, incluindo o próprio *baseline* CRF das competições DBDC3/4.

A Tabela 8 sumariza os resultados para as classes ‘quebra’ (B) e ‘possível quebra’ com ‘quebra’ (PB+B) no córpus DBDC3 inglês (Higashinaka et al., 2017). O melhor resultado de cada métrica é destacado.

C�rpus	Modelo	janela	DFmin	DFmax	est	lr	P
DBDC3	RegW2V	5	4	0,92	15	0,001	4
	RegBERT	7	5	0,97	142	0,01	3
DBDC4	RegW2V	4	4	0,76	50	0,001	3
	RegBERT	4	4	0,76	50	0,001	3
DBDBR	RegW2V	6	2	0,91	100	0,1	5
	RegBERT	6	2	0,97	92	0,03	5

Tabela 6: Hiper-par metros  timos para RegW2V e RegBERT em cada c rpus.

Modelo	$F_1(B)$
Baseline CRF (Higashinaka et al., 2017)	0,53
RegW2V	0,57
RegBERT	0,56
GruGloVe	0,23

Tabela 7: Resultados de medida F_1 do c rpus DBDBR portugu s para a classe ‘quebra’.

Modelo	$F_1(B)$	$F_1(PB+B)$
Baseline CRF (Higashinaka et al., 2017)	0,35	0,76
Pleco (Iki & Saito, 2017)	0,36	0,87
RegW2V	0,46	0,85
RegBERT	0,46	0,86
GruGloVe	0,56	0,84

Tabela 8: Resultados de medida F_1 do c rpus DBDC3 ingl s para a classe ‘quebra’ (B) e ‘poss vel quebra’ com ‘quebra’ (PB+B).

Nestes resultados observa-se que, para a m trica $F_1(B)$, o modelo GruGloVe supera as alternativas com ampla vantagem. No caso da m trica $F_1(PB+B)$, por outro lado, nenhum dos modelos propostos supera o sistema Pleco da competi o DBDC3, ainda que a margem seja pequena (especialmente em rela o ao modelo RegBERT).

Finalmente, a Tabela 9 sumariza os resultados para as classes ‘quebra’ (B) e ‘poss vel quebra’ com ‘quebra’ (PB+B) no c rpus DBDC4 ingl s (Higashinaka et al., 2019). O melhor resultado de cada m trica   destacado.

Modelo	$F_1(B)$	$F_1(PB+B)$
Baseline CRF (Higashinaka et al., 2017)	0,34	0,58
NTTCS19 (Sugiyama, 2021)	0,46	0,77
RegW2V	0,42	0,75
RegBERT	0,39	0,68
GruGloVe	0,41	0,78

Tabela 9: Resultados de medida F_1 do c rpus DBDC4 ingl s para a classe ‘quebra’ (B) e ‘poss vel quebra’ com ‘quebra’ (PB+B).

No caso da m trica $F_1(B)$, observa-se que os modelos propostos n o atingem o resultado obtido pelo sistema NTTCS19 da competi o DBDC4. J  no caso da m trica $F_1(PB+B)$, o modelo GruGloVe apresenta uma pequena vantagem sobre os demais.

Para an lise de signific ncia estat stica, os tr s modelos propostos (RegW2V, RegBERT e GruGloVe) foram comparados ao *baseline* CRF⁹ utilizando-se o m todo de *bootstrap* em Efron & Tibshirani (1994). De forma mais espec fica, para cada sistema um dos sistemas propostos (RegW2V, RegBERT e GruGloVe), e tamb m para o *baseline* CRF, foram extra das 100 amostras aleat rias de cada conjunto de predi o com uma taxa de amostragem de 95%, e ent o foi calculada a medida F_1 m dia do sistema considerando a classe PB+B no caso da tarefa em ingl s, ou apenas a classe B no caso da tarefa em portugu s. Finalmente, os resultados de cada um dos sistemas propostos foram comparados aos resultados obtidos pelo *baseline* por meio de um teste-*t*. A Tabela 10 sumariza os testes realizados, na qual todas diferen as em rela o ao *baseline* CRF s o significativas para $p < 0,0001$.

Com base nestes resultados, constatou-se que os modelos propostos s o significativamente superiores ao *baseline* CRF em todos os cen rios, com exce o do modelo GruGloVe para o c rpus DBDBR, em que foi observado um efeito significativo no sentido oposto.

5.2. An lise de erros

Como forma de identificar poss veis problemas de classifica o e oportunidades de melhoria futura, foi realizada tamb m uma breve an lise de erros frequentes dos modelos RegW2V e GruGloVe sobre os c rpus em ingl s DBDC3 e DBDC4, j  que estes apresentavam maior variedade do que a proporcionada pela rotula o bin ria do c rpus

⁹A an lise n o inclui os sistemas Pleco (Iki & Saito, 2017) e NTTCS19 (Sugiyama, 2021) porque os resultados detalhados de suas predi o n o est o dispon veis, e porque n o s o aplic veis ao c rpus DBDBR.

Córpus	CRF	RegW2V		RegBERT		GruGloVe	
	F_1	F_1	teste t	F_1	teste t	F_1	teste t
DBDBR	0,725	0,743	138	0,734	76	0,402	2077
DBDC3	0,243	0,361	452	0,327	233	0,262	115
DBDC4	0,205	0,238	485	0,269	539	0,251	513

Tabela 10: Medida F_1 e estatísticas do teste t comparando o baseline CRF a cada um dos modelos propostos. Todas as diferenças em relação ao baseline são significativas para $p < 0,0001$.

em português. Para este fim, foram selecionados aleatoriamente 50 diálogos de cada córpus, totalizando 1034 interações humano-computador. Estas interações foram analisadas de forma empírica pelo primeiro autor deste estudo, que identificou quatro categorias de erros mais frequentes, aqui denominadas ‘Erro de continuidade’, ‘Erro de anotação majoritária’, ‘Erro de saudação + pergunta’ e ‘Erro de quebras consecutivas’. A proporção de erros identificados em cada uma destas categorias é apresentada na Tabela 11, e detalhes adicionais são discutidos a seguir.

Erros de continuidade, exclusivos do modelo RegW2V, ocorrem quando o usuário continua o assunto de uma interação anterior porém o modelo identifica a não-quebra como sendo uma possível quebra, ou seja, ‘esquecendo’ o histórico do diálogo. Este tipo de problema foi melhor contornado com a classificação de sequências do modelo GruGloVe.

Erros de anotação majoritária representam os casos de maior ambiguidade na anotação de quebras presentes nos córpus DBDC3/4. Dado que os modelos consideram (assim como nas respectivas competições) o rótulo da classe como sendo aquele que tenha o maior número de anotações (ou votos) da equipe de juízes, observa-se que os modelos propostos tendem a classificar como possível quebra os casos em que a distribuição dos votos é mais balanceada, ou seja, quando não há uma tendência forte para quebra ou para não quebra.

Erros do tipo ‘Saudação + Pergunta’ são referentes ao uso combinado de uma saudação do usuário e de uma solicitação na mesma sentença, como em ‘Olá, então quem você está visitando?’. Solicitações deste tipo são frequentemente respondidas pelo *chatbot* considerando-se apenas a saudação, e produzindo respostas como em ‘Olá’. Todos estes casos constituem quebras de diálogo genuínas, mas tendem a ser classificadas apenas como possível quebra pelos modelos avaliados.

Finalmente, os erros do tipo ‘Quebras consecutivas’ são referentes ao efeito cumulativo de uma sequência de falhas no diálogo. Em casos deste tipo, os modelos propostos tendem a clas-

sificar incorretamente a sequência em sua totalidade mesmo quando parte das respostas era na verdade apropriada.

6. Discussão

Os experimentos realizados apresentam grande variação de resultados, o que era de certa forma esperado dada a variedade de conjuntos de dados, idiomas, definições de classe e métricas de avaliação. A seguir apresentamos de forma resumida algumas considerações a esse respeito.

Em primeiro lugar, observa-se que nas três tarefas abordadas, os melhores resultados foram obtidos por um dos sistemas propostos (RegW2V, RegBERT ou GruGloVe), ou por um sistema com resultados similares (i.e., sem diferença estatística significativa) em relação a estes. De forma mais específica, RegW2V obteve o melhor resultado para o córpus DBDBR, RegBERT ficou um ponto de medida F_1 abaixo do melhor modelo (Iki & Saito, 2017) para o córpus DBDC3, e GruGloVe obteve o melhor resultado para o córpus DBDC4.

Em segundo lugar, é interessante observar o papel de destaque dos modelos baseados em *transformers* do tipo BERT nestes experimentos. Não houve diferença significativa entre o melhor modelo de cada tarefa e RegBERT no caso dos córpus DBDBR e DBDC3, e somente na tarefa do córpus DBDC4 este modelo apresenta vantagem real em relação às alternativas avaliadas. Ainda assim, cabe observar que o sistema NTTCS19 (Sugiyama, 2021), vencedor da competição DBDC4, é também baseado em um modelo de língua pré-treinado do tipo BERT.

Finalmente, embora as tarefas para o português e inglês não sejam verdadeiramente comparáveis (já que utilizam córpus diferentes e modelam tarefas de classificação diferentes), é interessante observar que, considerando-se os valores médios de medida F_1 obtidos, a tarefa em português parece ser mais complexa do que suas contrapartidas em inglês. Enquanto o melhor resultado de medida F_1 para o córpus DBDBR português foi 0,57, para as tarefas DBDC3 e

Categoria de erro	RegW2V		GruGloVe	
	# de erros	% de erros	# de erros	% de erros
Erro de continuidade	23	3,8%	0	0,0%
Erro de anotação majoritária	161	26,9%	145	25,1%
Erro de saudação + pergunta	50	8,4%	61	10,5%
Erro de quebras consecutivas	99	16,6%	135	23,4%
Outros	265	44,3%	237	41,0%
Total de erros	598		578	

Tabela 11: Número (#) e percentual de erros cometidos pelos modelos RegW2V e GruGloVe na classificação de quebras nos  rpus DBDC3/4, por categoria de erro.

DBDC4 em ingl s obteve-se F_1 m ximo de 0,87 e 0,78, respectivamente, o que   de certa forma inesperado considerando-se que a tarefa em portugu s era bin ria, e portanto mais simples do ponto de vista computacional. Uma poss vel explica o para esta discrep ncia pode estar ligada ao tipo de fen meno representado pelos r tulos de cada  rpus. Como os r tulos dos  rpus DBDC3/4 foram obtidos por consenso de grandes equipes de anotadores,   poss vel que estes r tulos representem uma classe mais restrita de problemas de quebra em di logo, e que um alto grau de consist ncia na anota o facilite a tarefa de classifica o autom tica. No caso do  rpus DBDBR, por outro lado, o uso de r tulos derivados das indica es fornecidas por usu rios, e talvez o pr prio uso de dados de di logos reais, contempla uma gama possivelmente muito maior de motiva es para a quebra no di logo, e com alto grau de subjetividade. Embora esta complexidade adicional em certo sentido torne o problema computacional mais realista,   poss vel tamb m que isso explique o menor desempenho de todos os modelos empregados na tarefa em portugu s.

7. Considera es finais

Este trabalho apresentou uma investiga o de m todos de detec o autom tica de quebras em di logos humano-computador em portugu s e ingl s levando em conta o hist rico (ou mem ria) da conversa para decidir se a ocorr ncia de uma quebra   ou n o prov vel. Para este fim, fora propostos modelos que fazem uso de regress o e GRU bidirecional, e utilizando *embeddings* de palavra est ticos e contextuais. Al m disso, foi constru do um novo  rpus em portugu s composto de di logos reais produzidos por *chatbots* brasileiros que  , at  onde temos conhecimento, um recurso in dito na  rea para este idioma.

Os resultados obtidos variam conforme a classe e o conjunto de dados considerado, n o havendo uma solu o  tima  nica para todos os

cen rios de avalia o. Ainda assim, os resultados dos modelos propostos s o de modo geral pr ximos ou superiores aos dos sistemas de *baseline* considerados, incluindo os melhores sistemas participantes das competi es DBDC3/4.

O estudo realizado deixa uma s rie de oportunidades de melhorias e trabalhos futuros. Em especial, destacamos que um trabalho mais extenso de otimiza o do modelo GRU pode levar a resultados superiores aos atuais, assim como combina es de *embeddings* contextuais BERT com outros m todos de classifica o al m da regress o log stica da proposta atual. Outras possibilidades incluem, por exemplo, o estudo de quebras de refer ncias pronominais¹⁰ e o uso de conhecimento autoral como caracter sticas de personalidade¹¹ do usu rio em aux lio   tarefa de detec o de quebras em di logos.

No que diz respeito ao  rpus em portugu s utilizado, observamos que o presente trabalho concentrou-se apenas nas quebras identificadas automaticamente por terem sido sinalizadas pelos usu rios do sistema.   bastante prov vel, entretanto, que estes di logos contenham muitas outras quebras n o sinalizadas, e que seria igualmente importante conhecer e tratar computacionalmente. Um trabalho de anota o desta natureza, aos moldes do desenvolvido nas competi es DBDC para o ingl s e japon s,   tamb m deixado como trabalho futuro, assim como a pr pria tarefa de cria o de uma vers o anonimizada dos dados, a ser disponibilizada para futuras pesquisas na  rea.

Agradecimentos

O segundo autor contou com apoio da Universidade de S o Paulo.

¹⁰Paraboni (1997); Paraboni & de Lima (1998).

¹¹Silva & Paraboni (2018a,b); dos Santos et al. (2017).


Referências

- Almansor, Ebtesam Hussain, Farookh Kha-deer Hussain & Omar Khadeer Hussain. 2021. Supervised ensemble sentiment-based framework to measure chatbot quality of services. *Computing* 103. 491–507. doi 10.1007/s00607-020-00863-0.
- Banchs, Rafael E. & Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. Em *ACL 2012 System Demonstrations*, 37–42.
- Bear, John, John Dowding, & Elizabeth Shriberg. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. Em *30th Annual Meeting of the Association for Computational Linguistics*, 56–63. doi 10.3115/981967.981975.
- Black, Alan & Maxine Eskenazi. 2009. The spoken dialogue challenge. Em *10th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL)*, 337–340.
- Brandtzaeg, Petter Bae & Asbjørn Følstad. 2017. Why people use chatbots. Em *International Conference on Internet Science (INSCI)*, 377–392.
- Bulyko, Ivan, Katrin Kirchhoff, Mari Ostendorf & J. Goldberg. 2005. Error-correction detection and response generation in a spoken dialogue system. *Speech Communication* 45(3). 271–288.
- Carpenter, Paul, Chun Jin, Daniel Wilson, Rong Zhang, Dand Bohus & Alexander I. Rudnicky. 2001. Is this conversation on track? Em *EUROSPEECH 2001 Scandinavia; 7th European Conference on Speech Communication and Technology and 2nd INTERSPEECH Event*, 2121–2124.
- Chiaráin, Neasa Ní & Ailbhe Ní Chasaide. 2016. Chatbot technology with synthetic voices in the acquisition of an endangered language: motivation, development and evaluation of a platform for irish. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 3429–3435.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.
- Efron, Bradley & Robert Tibshirani. 1994. *An introduction to the bootstrap*. CRC Press.
- Friedman, Jerome. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5). 1189–1232. doi 10.1214/aos/1013203451.
- Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Rodrigues & Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.
- Hendriksen, Mariya, Artuur Leeuwenberg & Marie-Francine Moens. 2021. LSTM for dialogue breakdown detection: Exploration of different model types and word embeddings. Em *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, 443–453. doi 10.1007/978-981-15-9323-9_41.
- Higashinaka, Ryuichiro, Luis Fernando D’Haro, Bayan Abu Shawar, Rafael E. Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi & Joao Sedoc. 2019. Overview of dialogue breakdown detection challenge 4. Em *Dialog System Technology Challenge*, em linha.
- Higashinaka, Ryuichiro, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi & Nobuhiro Kaji. 2017. Overview of dialogue breakdown detection challenge 3. Em *Dialog System Technology Challenge*, em linha.
- Higashinaka, Ryuichiro, Kotaro Funakoshi, Yuka Kobayashi & Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 3146–3150.
- Iki, Taichi & Atsushi Saito. 2017. End-to-end character-level dialogue breakdown detection with external memory models. Em *Dialog System Technology Challenges Workshop*, em linha.
- Kato, Sosuke & Tetsuya Sakai. 2017. RSL17BD at DBDC3: Computing utterance similarities based on term frequency and word embedding vectors. Em *Dialog System Technology Challenges Workshop*, em linha.
- Lopes, José. 2017. How generic can dialogue breakdown detection be? the KTH entry to DBDC3. Em *Dialog System Technology Challenges Workshop*, em linha.
- van der Maaten, Laurens & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(86). 2579–2605.

- Martinovski, Bilyana & David R. Traum. 2003. Breakdown in human-machine interaction: the error is the clue. Em *ISCA tutorial and research workshop on Error handling in dialogue systems*, 11–16.
- Mikolov, Tomas, Scott Wen-tau & Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.
- Ng, Nathan, Kyunghyun Cho & Marzyeh Ghassemi. 2020a. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1268–1283. doi 10.18653/v1/2020.emnlp-main.97.
- Ng, Nathan, Marzyeh Ghassemi, Narendran Thangarajan, Jiacheng Pan & Qi Guo. 2020b. Improving dialogue breakdown detection with semi-supervised learning. Em *34th Conference on Neural Information Processing (NeurIPS)*, em linha.
- Paraboni, Ivandré. 1997. *Uma arquitetura para a resolução de referências pronominais possessivas no processamento de textos em língua portuguesa*. Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul. Tese de Mestrado.
- Paraboni, Ivandré & Vera Lucia Strube de Lima. 1998. Possessive pronominal anaphor resolution in Portuguese written texts. Em *17th international conference on Computational linguistics-Volume 2*, 1010–1014.
- Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. doi 10.3115/v1/D14-1162.
- Sandbank, Tommy, Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, John Richards & David Piorkowski. 2018. Detecting egregious conversations between customers and virtual agents. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 1802–1811. doi 10.18653/v1/N18-1163.
- dos Santos, Vitor Garcia, Ivandré Paraboni & Bárbara Barbosa Claudino Silva. 2017. Big five personality recognition from multiple text genres. Em *Text, Speech and Dialogue (TSD)*, 29–37. doi 10.1007/978-3-319-64206-2_4.
- Shin, JongHo, Alireza Dirafzoon & Aviral Anshu. 2019. Context-enriched attentive memory network with global and local encoding for dialogue breakdown detection. Em *Workshop on Chatbots and Conversational Agent Technologies*, em linha.
- Silva, Bárbara Barbosa Claudino & Ivandré Paraboni. 2018a. Learning personality traits from Facebook text. *IEEE Latin America Transactions* 16(4). 1256–1262. doi 10.1109/TLA.2018.8362165.
- Silva, Bárbara Barbosa Claudino & Ivandré Paraboni. 2018b. Personality recognition from Facebook text. Em *13th International Conference on the Computational Processing of Portuguese (PROPOR)*, 107–114. doi 10.1007/978-3-319-99722-3_11.
- Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 403–417. doi 10.1007/978-3-030-61377-8_28.
- Sugiyama, Hiroaki. 2017. Dialogue breakdown detection based on estimating appropriateness of topic transition. Em *Dialog System Technology Challenges Workshop*, em linha.
- Sugiyama, Hiroaki. 2019. Empirical feature analysis for dialogue breakdown detection. *Computer Speech & Language* 54. 140–150. doi 10.1016/j.csl.2018.09.007.
- Sugiyama, Hiroaki. 2021. Dialogue breakdown detection using BERT with traditional dialogue features. Em *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, Springer. doi 10.1007/978-981-15-9323-9_39.
- Sukhbaatar, Sainbayar, Arthur Szlam, Jason Weston & Rob Fergus. 2015. End-to-end memory networks. Em *Advances in Neural Information Processing Systems (NIPS)*, em linha.
- Takayama, Junya, Eriko Nomoto & Yuki Arase. 2017. Dialogue breakdown detection considering annotation biases. Em *Dialog System Technology Challenges Workshop*, em linha.
- Takayama, Junya, Eriko Nomoto & Yuki Arase. 2019. Dialogue breakdown detection robust to variations in annotators and dialogue systems. *Computer Speech & Language* 54. 31–43. doi 10.1016/j.csl.2018.08.007.
- Wang, Chih-Hao, Sosuke Kato & Tetsuya Sakai. 2019. RSL19BD at DBDC4: Ensemble of decision tree-based and LSTM-based models. Em

4th *Dialogue Breakdown Detection Challenge*, em linha.

Williams, Jason, Antoine Raux, Deepak Ramachandran & Alan W. Black. 2013. The dialog state tracking challenge. Em *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 404–413.

Yu, Zhou, Ziyu Xu, Alan W. Black & Alexander Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. Em *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 404–412.  [10.18653/v1/W16-3649](https://doi.org/10.18653/v1/W16-3649).