DAB: NOVA METODOLOGIA E FERRAMENTA DE ANOTAÇÃO DE DIÁLOGOS PARA PROCESSAMENTO DE LINGUAGEM NATURAL

Palavras-Chave: Anotação, Dados, Inteligência Artificial

Autores/as:

Andreis Gustavo Malta Purim, IC, UNICAMP

Henrique Theodor Schutz Foerste, IC, UNICAMP

Rafael Roque de Souza, IC, UNICAMP

Prof.(ª) Dr.(ª) Júlio Cesar dos Reis (orientador(a)), IC, UNICAMP

INTRODUÇÃO:

A popularização e o crescente acesso Modelos de Linguagem Grande (MLL) – parte da área de Processamento de Linguagem Natural (*NLP*) - como o ChatGPT, levam muitas empresas a adotarem *Chatbots* cada vez mais inteligentes para acelerar o atendimento virual ao cliente. No entanto, para que os modelos de Inteligência Artificial aprendam diversos contextos e tópicos conversacionais, é necessário treina-los com enormes quantidades de dados – que por sua vez passam por um exaustivo e caro processo de coleta e anotação (Coppola 2021).

Dois tipos de anotações são necessárias dentro de uma mensagem: Entidades (uma informação específica e identificável dentro de um texto ou sentença, como nomes de pessoas, organizações, locais, datas, valores numéricos e outras entidades nomeadas) e Inteções (intenção se refere ao objetivo ou propósito que um usuário deseja alcançar ao expressar uma frase ou mensagem). A figura 1 apresenta um exemplo de uma mensagem anotada de dialogo:

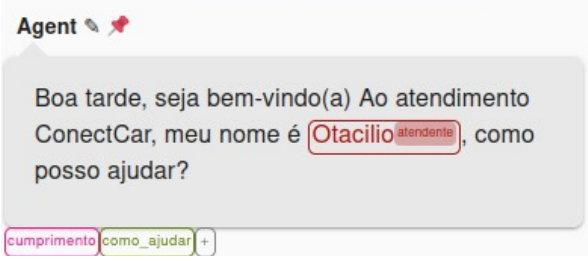

Figura 1 - Exemplo de uma mensagem anotada na ferramenta Assis, desenvolvida como parte deste estudo, com entidades (Nome do Atendente) e Intenções: Cumprimentar e perguntar se pode ajudar.

Para suprir a demanda de anotação de milhares de dialogos, existem diversas Softwares de Criação e Anotação de Dados – no entanto, elas são em grande parte pouco acessíveis à usuários sem conhecimento técnico em computação. Outro problema é a falta de metodologia unificada para anotação –

uma vez que cada ferramenta e cada equipe pode adotar diferentes filosofias no processo de anotação, resultando em sistemas sem integração automática.

Este estudo apresenta uma metodologia unificada para criação e anotação de diálogos (*DAB: Dialog Annotation Blueprint*) baseado em diversos estudos de casos executados, servindo como base para ferramentas de anotação de dados. Além disso, durante o estudo, essa metodologia foi base para a criação de dois softwares disponíveis gratuitamente para criação e anotação de dados.

DESENVOLVIMENTO:

A filosofia da metodologia unificada de anotação é englobar o maior número de ferramentas, processos e etapas entre a coleta e anotação de dados de forma interoperável. Isto é, diversos softwares de anotação podem se interligar por meio de chamadas *API (application programming interface).*

A base escolhida para o formato dos dados é o dataset MultiWoZ (Budzianowski et al, 2018), devido à sua popularidade como *Corpus* para Modelos de Linguagem. A metodologia é dividida em 6 macroetapas, conforme apresentado na figura 2:

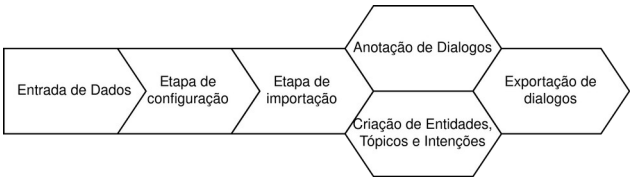

Figura 2 – Macroetapas de anotação de dados de dialogo na Metodologia DAB

Cada macroetapa se relaciona de forma independente, permitindo à diferentes softwares de atenderem diferentes especificações conforme necessário. A figura 3 mostra as diferentes etapas necessárias para a entrada de dados, criado de forma a ser compatível com a metodologia MCCD (Sanches et al, 2022) de aquisição de dados:

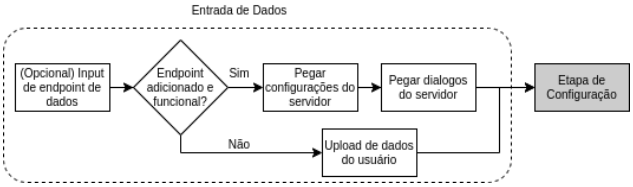

Figura 3 – Etapa da entrada de dados. Em ambos os casos (upload de dados e/ou endpoint configurado com outra ferramenta), o formato dos dados é na formatação MultiWoz.

A figura 4 apresenta as etapas de Configuração de *APIs* no Software, necessária para a interoperabilidade de softwares utilizando esta metodologia de anotação.

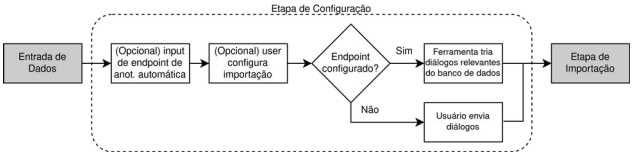

Figura 4 – Etapa de configuração de APIs para interoperabilidade de ferramentas

A etapa de importação de ontologias, uma das inovações propostas pela metodologia *DAB*, está representada na Figura 5. A importação de ontologias é um dos principais passos verificados no ganho de tempo com anotação de grandes datasets.

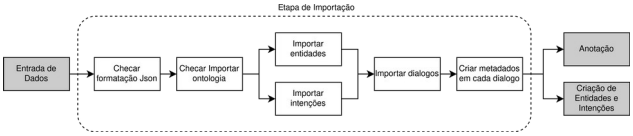

Figura 5 – Etapa de importação de ontologias e diálogos

As três macroetapas iniciais são algumas das principais inovações da Metodologia de Anotação proposta, pois até então não há na literatura um padrão proposto de configuração pré-anotação.

Os dois passos seguintes se dão em paralelo. A figura 6 apresenta a etapa da criação de entidades e intenções na forma de um laço de anotação até todos os dialogos estarem anotados:

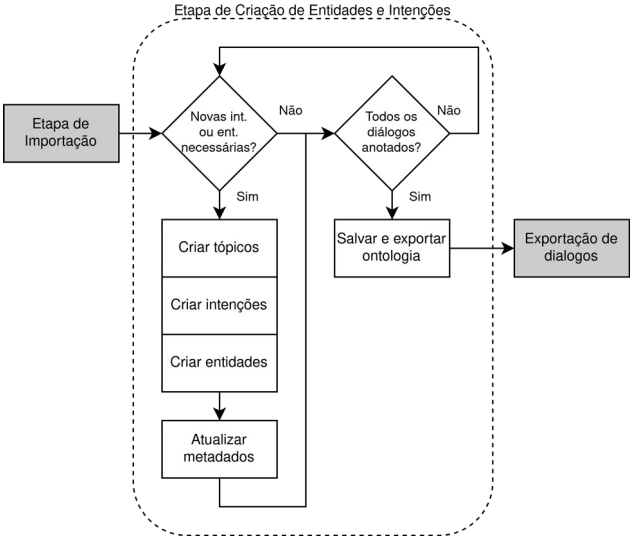

Figura 6 – Laço de criação de Entidades e Intenções.

Enquanto a figura 7 mostra o laço principal de anotação de diálogos. É neste laço que se encaixam os atuais softwares disponíveis. A metodologia propõe a divisão de softwares em dois módulos opcionais: um módulo de anotação por ferramenta, e um módulo de anotação assistida por Inteligência Artificial.

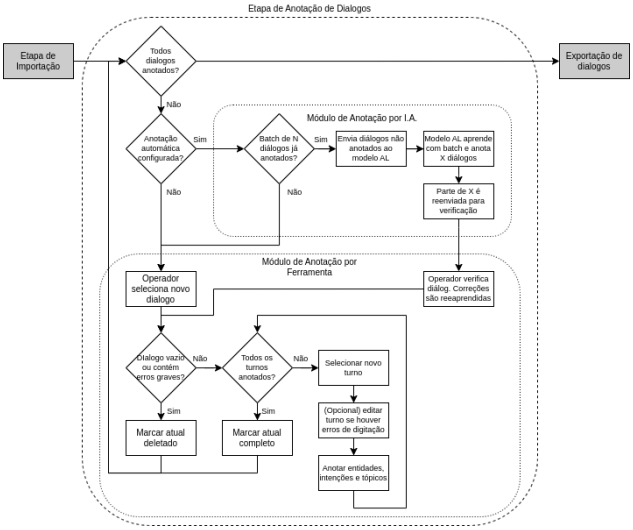

Figura 5 – Laço de anotação, com os dois módulos de software separados.

Finalmente, a anotação termina com a reestruturação dos dados no formato MultiWoZ, além da exportação da ontologia criada durante a anotação, comforme demonstrado na figura 8.

Figura 8 – Laço de criação de Entidades e Intenções.

APLICAÇÕES E DISCUSSÃO:

Conforme explicado anteriormente, a metodologia de anotação de diálogos é um proposta inovadora em esquematizar e propor as etapas de anotação de diálogos para chatbots de genérica. Para testar o ganho de valor dado, dois softwares foram desenvolvidas com base na metodologia:

- *Mechanical Dialog Creation (MDC):* criado por Matheus F. Sanches. Uma ferramenta online de criação de diálogos entre um atendente automático e um usuário, dado um número de tasks. Esta ferramenta aplica as macroetapas de criação de dados e importação. Ela é capaz de ser interligada com a ferramenta *Assis* de anotação, enviando automaticamente os dados gerados para anotação.
- *Assis Tool*: criado por Andreis Purim e Henrique Foerste e apresentado no Simpósio Brasileiro de Sistemas de Informação 2023, é uma plataforma online

de anotação de diálogos, que aplica todas as macroetapas da metodologia. O módulo de anotação por ferramenta é em uma *Single-Page Application*, enquanto o módulo de anotação por I.A. se dá na forma de um servidor capaz de anotar diálogos por de um modelo BERT com Active Learning. Após um número pre-determinado de dialogos anotados, o modelo poderá anotar os dialogos restantes. Cada anotação automática corrigida pelo usuário é reutilizada para o treinamento do modelo.

Figura 9 – Interface da ferramenta Assis durante a anotação de um diálogo. Até o atual momento, Assis é a única ferramenta capaz de ser utilizada em um smartphone.

Ambas as ferramentas foram utilizadas em casos de estudos com empresas interessadas na criação de Chatbots para o atendimento aos seus clientes. Foram anotados mais de 96 diálogos com uma média de 7 mensagens cada, totalizando 800 mensagens anotadas – com a equipe de anotação sendo usuários das empresas e voluntários da universidade.

Uma bateria final de testes foi realizada com 26 participantes para a avaliação de tempo na anotação entre o uso da ferramenta e a anotação manual. A tabela 1 apresenta o tempo médio de anotação dos 26 usuários anotando 4 pacotes de diferentes diálogos.

	Assis (min)	Manual (min)
Pacote A1	13:12	26:09
Pacote A2	21:29	34:05
Pacote B1	19:11	39:32
Pacote B2	31:26	45:16

Tabela 1 – Média do tempo de anotação dos 4 pacotes de diálogos entre os 26 voluntários.

Uma pesquisa de satisfação com 37 usuários demonstrou que 33 preferiam a ferramenta Assis além das outras ferramentas apresentadas. O ganho significativo de tempo nas anotações (quase metade no pacote A1 e B1), demonstra o valor na metodologia de anotação proposta.

Outros testes, melhorias e integrações das ferramentas estão sendo propostas, além de constante uso em anotação de outros dados. Ambas as ferramentas estão disponíveis online para uso.

CONCLUSÕES:

Modelos conversacionais – principalmente *chatbots* - necessitam de enormes volumes de dados anotados para que atinjam um nível de fluência adequado, e apesar da constante expanção e popularização da área, ainda há poucas ferramentas disponíveis para facilitar a anotação. Outro problema enfrentado é são as inúmeras formas de anotar – que resultam em dados e softwares incompatíveis uns com os outros.

A *Dialog Annotation Blueprint (DAB)* é uma proposta de metodologia unificada de anotação que atende todos os passos de criação e importação dos dados até sua

exportação para o treinamento de modelos de linguagem – que propõe um formato único e consistente (MultiWoZ) de dados anotados em macroetapas separadas, que podem ser atendidas por diferentes ferramentas de software se comunicando via *APIs*. Duas ferramentas foram criadas com esta metodologia, uma para criação de diálogos e a outra para anotação – ambas sendo validadas com satisfação de usuários em casos de estudo com dados reais.

BIBLIOGRAFIA

Riccardo Coppola and Luca Ardito. 2021. **Quality Assessment Methods for Textual Conversational Interfaces: A Multivocal Literature Review.** Information 12, 11 (2021), 437

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. **MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling.** arXiv preprint arXiv:1810.00278 (2018)

Matheus F. Sanches, Jader M. C. de S, Allan M. de Souza, Diego A. Silva, Rafael R. de Souza, Julio C. Dos Reis and Leandro A. Villas. **MCCD: Generating Human Natural Language Conversational Datasets from Online Forums.** International Conference on Enterprise Information Systems (ICEIS 2022).

Henrique Theodor Schutz Foerste, Andreis Gustavo Malta Purim, Rafael Roque Souza, and Julio Cesar Dos Reis. 2023. **Assis: Online Semi-Automatic Dialog Annotation Tool.** In Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI '23). Association for Computing Machinery, New York, NY, USA, 37–44.