OIEAnnotator

Uma ferramenta para construção e anotação de corpora para Extração de Informação Aberta

Igor Tironi
tironiigor@gmail.com
FORMAS Research Group – Institute of Computing –
Federal University of Bahia
Salvador, Bahia

Daniela Barreiro Claro
dclaro@ufba.br
FORMAS Research Group – Institute of Computing –
Federal University of Bahia
Salvador, Bahia

Resumo

A disponibilidade de corpora anotados é uma importante tarefa de *Open Information Extraction (Open IE)*. Porém, essa é uma tarefa difícil pois demanda trabalho manual de anotadores. Essa tarefa se torna ainda mais complicada no contexto da língua portuguesa, dada a sua complexidade e a falta de uma estrutura prévia para tarefas de anotação nesta língua. Ferramentas que possam agilizar esse processo tem um grande valor para a construção de conhecimento nesta área. Esse trabalho propôs uma ferramenta capaz de auxiliar no processo de construção de *corpora* anotados, através da anotação e identificação de novas triplas relacionais nas sentenças. Para validação, foi definido um grupo de especialistas, composto por três especialistas na tarefa, e um grupo de controle, composto por indivíduos sem conhecimento no processo para teste de usabilidade da ferramenta. A ferramenta foi utilizada para anotação de um *corpus* em português, mas não foi identificado nenhum impedimento para a utilização desta para outras linguas.

Keywords: Open IE, Corpora, Português, Anotação, Ferramenta

1 Introdução

Pesquisadores e cientistas enfrentam um grande desafio quando confrontados com a grande quantidade de informação que, embora disponível, se apresenta de forma não estruturada. A *Open Information Extraction (Open IE)*, área de estudo que possibilita estruturar dados de meios não estruturados, é responsável pela identificação e representação de informações contidas nesse tipo de dados em um conjunto de triplas representando relações [3].

O objetivo de um sistema de *Open IE* é que as triplas extraídas sejam uma sequência de palavras com alto poder discriminatório e potencial informativo [5]. Apesar da possibilidade da criação de regras para extração dessas informações com recursos humanos, utilizar estes recursos para fazer a anotação de *corpus* que possam ser utilizados por algoritmos de Aprendizado de Máquina (AM) supervisionados é mais vantajoso [5].

Apesar da área de Open IE ter tido avanços na última década, a maioria teve foco na língua inglesa, com um número pequeno de estudos para a língua portuguesa nos últimos cinco anos. Conforme apontado por Bender [1], esse fato introduz um viés nos modelos quando considerados para a língua portuguesa. Assim, se justifica a necessidade da criação de recursos que possibilitem a validação dos modelos importados, bem como o desenvolvimento de ferramentas e técnicas focadas na língua portuguesa [6].

Desta forma, como a criação de *corpora* anotados requer um trabalho manual de anotadores capacitados, ferramentas que facilitem a criação de *corpora* anotados podem contribuir para o desenvolvimento de sistemas de *Open IE*. Esse processo de anotação pode se tornar ainda mais complexo quando múltiplos anotadores precisam validar as mesmas informações extraídas, uma vez que a detecção e resolução de divergências é trabalhosa sem o auxílio de uma ferramenta específica.

Esse trabalho propõe, então, uma ferramenta *web* capaz de auxiliar a construção de *corpora* anotados, através da anotação e identificação de relações e argumentos para construção de *corpora* anotados. A ferramenta foi testada para anotação de um *corpus* em português, mas a arquitetura permite a inclusão de outros *corpora* para anotação na língua portuguesa.

2 Método

Esse estudo foca na construção de uma ferramenta, tomando como referência o processo e ferramentas utilizados previamente pelo grupo de pesquisa FORMAS da UFBA para construção de *corpora* para *Open IE*, e foi desenvolvida em cinco etapas conforme Figura 1.

Na primeira etapa foi realizada uma análise detalhada da ferramenta utilizada pelo grupo, buscando entender a semântica das relações e funcionamento do sistema. Durante esse

In: XXI Workshop de Ferramentas e Aplicações (WFA 2022), Curitiba, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2022.

ISSN 2596-1683

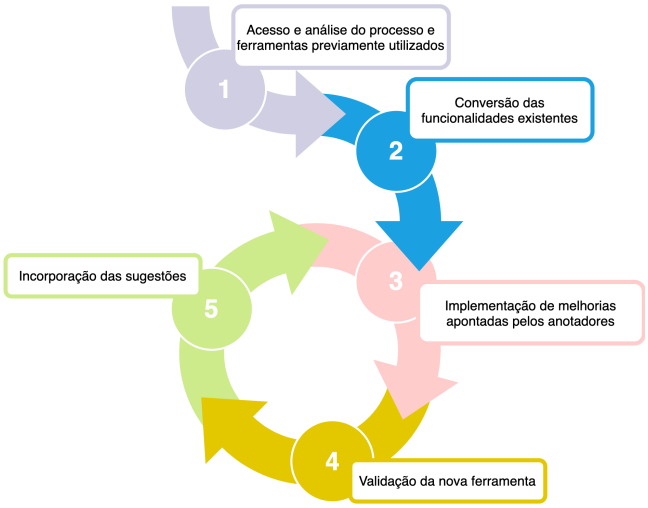

Figura 1. Fluxo da Metodologia [Fonte: Autor]

processo foram elencadas algumas melhorias que poderiam ser feitas na ferramenta para facilitar o processo de anotação.

Na segunda etapa, uma nova ferramenta foi construída, tomando como referência a ferramenta anterior. A nova ferramenta foi feita em uma estrutura cliente-servidor, utilizando a framework Laravel [8] como estrutura para servidor e o framework React [10] como estrutura para o cliente.

Na terceira etapa, os pontos de melhorias já identificados pelo grupo de pesquisa durante a utilização da ferramenta inicial e também da ferramenta proposta durante o desenvolvimento foram implementados. Além disso, interfaces foram incluídas para melhor visualização das relações textuais nas sentenças do *corpora*.

Na quarta etapa, a ferramenta foi utilizada pelos anotadores do grupo, experientes e não experientes. Após a utilização, foi feita uma coleta de feedback sobre a ferramenta. O formulário de feedback consistiu em 14 perguntas utilizando a Escala de Likert, bem como um campo aberto para sugestões.

Por fim, na quinta etapa, as sugestões da etapa quatro foram levadas em consideração e algumas delas também foram incorporadas na ferramenta. As etapas 3, 4 e 5 foram repetidas até que a ferramenta estivesse atendendo ao processo de anotação do grupo de anotadores.

3 A ferramenta OIEAnnotator

A ferramenta está disponível na web, o que facilita sua utilização pelos anotadores, e permite a inclusão de múltiplos *corpora*, que podem ser utilizados em diferentes processos de anotações por grupos diferentes. A licença do software é gratuita para uso acadêmico e pode ser solicitada para o grupo FORMAS.

Na ferramenta, o *corpus* é composto por um conjunto de sentenças, que, por sua vez, possui Classe Gramatical (POS Tags), Trechos de Texto (Chunks) e Árvore de Dependência (Dependency Tree), que são utilizados no processo de anotação e possuem interface para visualização na ferramenta.

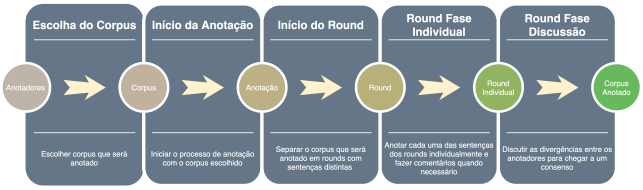

Figura 2. Processo de Anotação [Fonte: Autor]

Uma vez que o *corpus* esteja carregado na ferramenta, é possível incluir triplas relacionais (extrações) vinculando-as às sentenças, podendo estas serem criadas manualmente pelos usuários ou gerados por um sistema de *Open IE*. Esse processo de inclusão de extrações pode ser realizado diretamente ao *corpus*, ou durante a etapa de anotação que será descrita a seguir.

O processo de anotação consiste na validação e comentários das extrações do *corpus* pelos anotadores alocados, permitindo também a inclusão de novas extrações manualmente criadas durante essa etapa. A ferramenta permite a divisão desse processo de anotação em *rounds* que são responsáveis por um conjunto disjunto das sentenças do *corpus* sendo anotado.

Cada *round* consiste em duas etapas: individual e discussão. Durante a etapa individual cada anotador deve fazer a validação das extrações das sentenças deste *round* sem interação com outros anotadores, adicionando comentários para explicar suas decisões ou elencar dúvidas. Essa etapa permite que os anotadores analizem as extrações sem a influência dos outros. Na segunda etapa todos os anotadores têm acesso às validações e comentários dos outros anotadores, o que permite que eles façam as discussões necessárias sobre as validações divergentes para chegar a uma conclusão final sobre a validade da extração ou aceitação da divergência de opiniões.

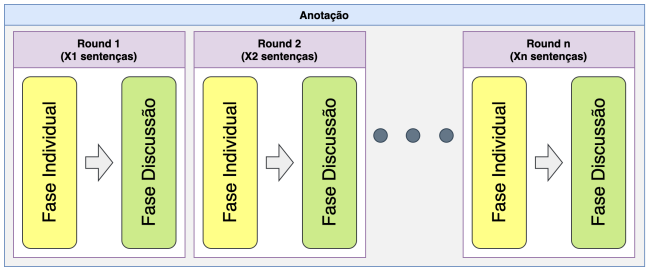

Figura 3. Anotação e *Rounds* [Fonte: Autor]

Essas validações das extrações podem ser exportadas então no formato de um *corpus* anotado para que seja utilizado em outras tarefas de treinamento.

3.1 Requisitos da Ferramenta

Requisitos funcionais são afirmações sobre o que um sistema deve prover, como deve reagir com determinadas entradas e como deve se comportar em determinados cenários [Sommerville, 2007 apud [2]].

3.1.1 Requisitos Funcionais.

- **RF1 - Autenticação dos Anotadores.** A ferramenta deve fornecer um sistema de autenticação para acesso individual aos anotadores
- **RF2 - Cadastro de *Corpora*** A ferramenta deve permitir que múltiplos *corpora* sejam cadastrados, cada um composto por um conjunto de sentenças com *POS Tags*, *Chunks* e *Dependency Tree*
- **RF3 - Processo de Anotação.** A ferramenta deve permitir a criação de um processo de anotação de um *corpus* vinculada a múltiplos anotadores. Múltiplos processos de Anotação devem ser permitidos para um mesmo *corpus* sem que haja interação entre eles.
- **RF4 - Inclusão de Extrações.** A ferramenta deve permitir a inclusão de extrações para sentenças de um *corpus*. Essa extração pode estar vinculada diretamente ao *corpus* ou apenas à anotação.
- **RF5 - Validação das Extrações.** A ferramenta deve permitir a validação das extrações no nível da anotação de forma individual para cada anotador alocado no processo.
- **RF6 - Processo de *Rounds* na Anotação.** A ferramenta deve permitir a criação de *rounds*. Esses *rounds* devem ser responsáveis por um conjunto disjunto das sentenças do *corpus* sendo anotado.
- **RF7 - Comentários nas Extrações do *Round* na Anotação.** A ferramenta deve permitir a criação de comentários nas extrações vinculados ao *round*.
- **RF8 - Fases do *Round*.** A ferramenta deve permitir a divisão de um *round* em duas fases, uma individual e uma de discussão. Na primeira etapa os anotadores não devem ter acesso às validações e aos comentários dos outros anotadores e na segunda etapa todas as informações devem estar visíveis para todos.
- **RF9 - Extração do *Corpus* Anotado.** A ferramenta deve permitir a extração do *Corpus* anotado.
- **RF10 - Visualização dos componentes da Sentença.** A ferramenta deve permitir a visualização dos componentes da sentença (*POS Tags*, *Chunks* e *Dependency Tree*) de forma visual para os anotadores.
- **RF11 - Identificação de Divergências.** A ferramenta deve identificar e alertar sobre divergências das validações entre os anotadores.

3.1.2 Requisitos Não Funcionais.

- **RNF1 - Fácil Usabilidade.** A ferramenta deve estar disponível na *web* e permitir que os anotadores compartilhem o processo de anotação do *corpus*.
- **RNF2 - Disponibilidade de Interface para Aplicação.** A ferramenta deve disponibilizar uma Interface para Aplicação (API) para integrações futuras.
- **RNF3 - Segurança.** Por ser uma plataforma web, é necessário que a aplicação seja segura para manter a privacidade e consistência dos *corpora* sendo anotados.
- **RNF4 - Manutenibilidade.** Por se tratar de uma ferramenta que será mantida pelo grupo de pesquisa, é ideal que sejam mantidas boas práticas, permitindo que futuros membros possam dar suporte a aplicação.

4 Avaliação

Para validação da ferramenta, foi realizado um questionário com três experts, escolhidos por serem especialistas na tarefa e/ou futuros usuários, bem como algumas pessoas sem familiaridade com a ferramenta para teste de usabilidade.

O formulário foi elaborado com 14 perguntas utilizando a Escala de Likert e um campo aberto para sugestões para a ferramenta, sendo aplicado após a utilização da ferramenta para anotação de sentenças de um *corpus* em português seguindo a metodologia proposta. Essas perguntas visavam avaliar o entendimento sobre o processo, o desempenho da ferramenta e o impacto das funcionalidades. Por questões de espaço, a explicação de cada pergunta foi suprimida.

Discussão da Avaliação. . A avaliação da ferramenta foi feita por 9 indivíduos, sendo desses 3 experts na tarefa de anotação, compondo o Grupo de Experts (GE), e 6 utilizados como Grupo de Controle (GC). Os indivíduos foram considerados como GE se respondessem o valor máximo na pergunta "Qual o seu nível de conhecimento sobre uma tarefa de anotação?". Como podemos ver na Tabela 1, a média das avaliações para o GE foi maior do que a do GC, o que era esperado por ser um processo complexo que o GC não tinha experiência. Além disso, levando em consideração as respostas para a pergunta aberta e para a pergunta sobre a distinção de *Corpus*, Anotação e *Round*, mostra-se necessário uma melhoria na forma como o processo de anotação é explicado dentro da ferramenta para auxiliar na capacitação de novos anotadores.

	Grupo de Controle			Grupo Experts			Total		
	Mínimo	Máximo	Média	Mínimo	Máximo	Média	Mínimo	Máximo	Média
A disponibilidade da ferramenta via Web facilita o processo de anotação e avaliação das sentenças e triplas.	4	5	5	3	5	4	3	5	4.56
Quão fácil é o entendimento da interface da ferramenta?	2	5	4	4	5	5	2	5	4.11
Está clara a distinção entre Corpus, Anotação e Round na ferramenta?	1	5	4	3	5	4	1	5	4.00
Quão fácil foi a identificação do corpus a ser trabalhado e das sentenças anotadas?	2	5	4	4	5	4	2	5	4.00
O que você achou da usabilidade da ferramenta?	2	5	4	4	5	5	2	5	4.11
O que você achou da navegabilidade da ferramenta?	3	5	4	4	5	5	3	5	4.44
Você compreendeu o que são as diferentes fases em cada Round?	2	5	4	4	5	5	2	5	4.22
Quanto a ferramenta lhe auxiliou no processo de anotação?	1	5	4	4	5	5	1	5	4.11
O progresso do Round auxilia na tarefa de Anotação?	3	5	5	5	5	5	3	5	4.67
A visualização da Árvore de Dependência facilita a tarefa de anotação?	3	5	4	5	5	5	3	5	4.56
A visualização dos Chunks facilita a tarefa de anotação?	4	5	5	5	5	5	4	5	4.78
A possibilidade de comentários dentro da ferramenta facilita o processo de discussão nos Rounds?	4	5	5	5	5	5	4	5	4.78
Quão satisfeito você está com a ferramenta?	2	5	4	5	5	5	2	5	4.33
Qual o seu nível de conhecimento sobre uma tarefa de anotação?	1	3	3	5	5	5	1	5	3.33

Tabela 1. Tabela Agregada de Avaliação

5 Trabalhos Relacionados

Algumas outras ferramentas já foram propostas com o intuito de facilitar o processo de criação de conteúdo para o processo de Open IE.

A OPEN IE - Annotation tool [9] foi desenvolvida em PHP, com uma interface web e uma ideia similar de criação de extrações. Apesar da ferramenta possibilitar que os anotadores incluam novas extrações e validem as extrações existentes, o módulo para discussão das anotações não foi desenvolvido, forçando assim que os anotadores recorressem a métodos manuais para a fase de discussão do *Round*. Além disso, a ferramenta não possui visualização gráfica dos componentes da sentença, o que dificulta o processo de anotação.

A Doccano [7] é uma ferramenta de propósito geral para tarefas de anotação de texto. A estrutura de reconhecimento de entidades nomeadas da ferramenta não seria ideal para visualização das relações. Além disso, a ferramenta não possui interface para visualização dos componentes da sentença, o que dificulta o processo de avaliação e criação de novas extrações. Por fim, a falta de uma interface de discussão faria com que os anotadores precisassem fazer um processo manual para identificação e resolução de conflitos, o que não é ideal para o contexto.

A Prodigy [4] também é uma ferramenta de propósito geral para tarefas de anotação de texto. O módulo da ferramenta de dependências e relações pode ser utilizado para a construção de *corpus*, mas o fato de não ter uma interface para discussão torna o trabalho mais complicado para os anotadores. Além disso, a falta de visualização dos componentes da sentença também dificultaria a tarefa.

Com o intuito de compará-las os seguintes critérios foram utilizados: (i) Disponibilidade na Web; (ii) Adequado ao processo; (iii) Visualização gráfica dos componentes da sentença; (iv) Compartilhamento entre anotadores; (v) Discussão das divergências; (vi) Facilidade de expansão da ferramenta; (vii) Open Source. Por questões de espaço, as explicações sobre cada critério foi suprimida.

	Ferramentas			
Atributos	Ferramenta Proposta	OPEN IE - Annotation tool [Prates e Emanuel 2021]	Doccano [Nakayama et al. 2018]	Prodigy [Explosion 2017]
Disponibilidade na Web	Sim	Sim	Sim	Sim
Adequado ao processo (1-5)	5	4	3	3
Visualização gráfica dos componentes da sentença	Sim	Não	Não	Não
Compartilhamento entre anotadores	Sim	Sim	Sim	Sim
Discussão das divergências	Sim	Não	Não	Não
Facilidade de expansão da ferramenta (1-5)	4	2	5	1
Open Source	Não	Não	Sim	Não
Linguagem	PHP e Javascript	PHP	Python	Javascript

Tabela 2. Quadro Comparativo das Ferramentas

Desta forma, podemos ver que existem ferramentas que conseguem entregar experiências para anotação de *corpora* anotados, mas a entrega de todas as funcionalidades necessárias para o processo de anotação proposto em uma única ferramenta é um diferencial da ferramenta proposta. Além disso, criar uma comunidade *Open Source* ativa em volta da ferramenta pode ser bom para a criação de novos módulos e melhoria dos existentes, o que está relacionado com a facilidade de expansão da ferramenta.

6 Conclusão e Trabalhos Futuros

Essa ferramenta sistematiza o processo de anotação da tarefa de extração de informação, facilitando a execução do processo, gerando ganho de tempo, aumentando efetividade e evitando falhas. A disponibilidade na Web permite que a mesma seja executada sem a instalação de novos softwares e sendo mais amigável para anotadores com menos familiaridade.

A possibilidade da visualização dos dados em uma interface dentro da ferramenta também facilita no processo de validação da extração, pois assim os anotadores tem um contexto melhor sobre a sentença sendo anotada.

Por fim, estes resultados podem ser replicados por outros grupos, com a tarefa de anotação em outros *corpora* e em línguas diferentes do português.

Referências

[1] Emily Bender. 2019. English isn't generic for language, despite what NLP papers might lead you to believe. In *Symposium and Data Science and Statistics* (Bellevue WA). http://faculty.washington.edu/ebender/papers/Bender-SDSS-2019.pdf [Online; accessed 15-may-2020].

[2] Bruno Cabral. 2014. SPLICE: A Flexible SPL Lifecycle Management Tool. (2014). Universidade Federal da Bahia. Graduação sob orientação de Eduardo Santana de Almeida.

[3] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Commun. ACM* 51, 12 (2008), 68–74.

[4] Explosion. 2017. Prodigy: Radically efficient machine teaching. An annotation tool powered by active learning. https://prodi.gy/ Disponível em https://prodi.gy/.

[5] Cláudia Freitas, Milena Uzeda-Garrão, and Claudia Oliveira. 2005. A anotação de um corpus para o aprendizado supervisionado de um modelo de SN. *XXV Congresso da Sociedade Brasileira de Computação* (01 2005). https://www.researchgate.net/publication/242091301_A_anotacao_de_um_corpus_para_o_aprendizado_supervisionado_de_um_modelo_de_SN

[6] Rafael Glauber, Leandro Souza de Oliveira, Cleiton Fernando Lima Sena, Daniela Barreiro Claro, and Marlo Souza. 2018. Challenges of an Annotation Task for Open Information Extraction in Portuguese. In *Computational Processing of the Portuguese Language*, Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold (Eds.). Springer International Publishing, Cham, 66–76.

[7] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human. https://github.com/doccano/doccano Software available from https://github.com/doccano/doccano.

[8] Taylor Otwell. 2011. Laravel: The PHP Framework for Web Artisans. https://laravel.com/ Software available from https://github.com/laravel/laravel.

[9] Arley Prates and Luis Emanuel. 2021. OPEN IE - Annotation tool. https://github.com/arleyprates/openie-annotation-tool

[10] Jordan Walke. 2013. React: A JavaScript library for building user interfaces. https://reactjs.org/ Software available from https://github.com/facebook/react.