

METODOLOGIAS E FERRAMENTAS PARA ANOTAÇÃO DE NARRATIVAS CLÍNICAS

Gabriel Herman Bernardim Andrade^{1,2}, Lucas Emanuel Silva e Oliveira² e Cláudia Maria Cabral Moro²

¹ Engenharia de Computação, Escola Politécnica, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brasil

² Programa de Pós-Graduação em Tecnologia em Saúde, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brasil

Resumo: A utilização de técnicas de Processamento de Linguagem Natural (PLN) em textos clínicos é amplamente dependente de grandes quantidades de dados textuais anotados, denominados corpus ou padrão ouro. Sendo essenciais para a modelagem da linguagem durante a fase de treinamento de diversos algoritmos de PLN. Porém, para a criação de um padrão ouro é necessário um extenso e custoso trabalho manual de anotação, que demanda um grande esforço de especialistas. **Objetivo:** Realizar uma revisão da literatura, visando o estudo de metodologias e ferramentas utilizadas em procedimentos de anotação de textos. **Método:** Levantamento em bases científicas referentes à elaboração de corpus morfológicos, sintáticos e morfossintáticos foi realizado, analisando 32 estudos de anotação e mais 12 ferramentas. **Resultados:** Foram levantados os principais aspectos nos processos de anotação, bem como realizada uma avaliação dentre critérios pré-definidos de cada das ferramentas de suporte encontradas.

Palavras-chave: Processamento de Linguagem Natural; Anotação de Textos; Ferramentas de Anotação; Corpus padrão ouro.

Abstract: The use of natural language processing techniques (NLP) in clinical texts is dependent on large amounts of annotated text data, called corpus or gold standard. Are essential for the modelling language during the training phase of NLP algorithms. However, for the creation of a gold standard is required extensive and costly manual annotation task, that demands a great deal of experts. **Objectives:** To review the literature to identify methodologies and tools applied to text annotation. **Methods:** Scientifics databases search regarding the development of morphological, syntactic and morphosyntactic corpus was performed by analyzing 32 annotation studies and 12 tools. **Results:** Main aspects of the annotation process description, as well as an assessment from pre-defined criteria for each one of the annotation tools identified.

Keywords: Natural Language Processing; Text Annotation; Annotation Tools; Gold Standard Corpus.

Introdução

As técnicas de Processamento de Linguagem Natural (NLP) necessitam de uma grande quantidade de informação anotada de forma para apresentarem um bom desempenho. Tal informação é de extrema importância na fase de aprendizado de padrões e modelagem de linguagem. A anotação desses dados, conhecidos como corpus ou padrões ouro, depende de extensiva análise humana e anotação manual, processo que é extremamente demorado e propenso a erros, pois requer um grande esforço intelectual de especialistas no domínio linguístico e/ou específico. A anotação de um grande corpus frequentemente depende de um time de anotadores e uma série de orientações para solucionar divergências ou inconsistências¹⁻³.

Entretanto, a tarefa de criação de corpus vem sendo auxiliada por ferramentas computacionais desde a década de 1990, como no caso da anotação do Penn TreeBank Corpus¹. Dentre tais ferramentas, pode-se citar simples processadores de texto com funcionalidade de auto completar, até interfaces gráficas que permitem que anotadores selecionem e etiquetem *tokens* ou montem estruturas em sentenças⁴⁻⁷, gerenciamento da coleção de textos ou análise automática de não concordância entre anotadores. Outro aspecto que minimiza o trabalho a ser realizado é a semi-automação do processo, onde o texto pode ser pré-etiquetado ou estruturas podem ser pré-formadas por algoritmos já treinados. Isso permite que os anotadores tenham o trabalho de apenas corrigir os erros da fase automática, o que se mostra muito mais produtivo, além de reduzir taxa de erro dos anotadores^{1,8}.

Essas ferramentas variam bastante no quesito de suas funcionalidades, permitindo o desenvolvimento de corpus anotados sobre diferentes características. Certas ferramentas permitem a anotação de diversas camadas de informação (como morfológica, sintática, semântica e relacionamento de conceitos)⁶, algumas focam apenas em propriedades linguísticas do texto⁵ e outras incorporam conceitos do domínio semântico específico do texto⁷.

Com o conhecimento de múltiplas metodologias e ferramentas com diferentes funcionalidades é possível obter uma melhor avaliação de qual abordagem é mais apropriada para o processo de anotação, visando as características que se deseja obter no corpus final. Uma escolha apropriada permite não só um produto final melhor, mas também um processo de desenvolvimento menos custoso, pois o trabalho manual dos anotadores pode ser reduzido e sua produtividade aumentada. O conhecimento do que se encontra disponível na literatura permite também que seja evitado o desperdício de tempo no desenvolvimento de uma ferramenta completa de anotação em casos desnecessários.

Corpora desenvolvidos a partir de textos jornalísticos podem ser facilmente encontrados na literatura. Especificamente para a língua Portuguesa, existem trabalhos como o desenvolvido por Bick⁹. Entretanto, corpora anotados sobre textos do contexto médico são praticamente inexistentes e, segundo os resultados obtidos por Pakhomov³ e Peters⁸, técnicas de NLP aplicadas a diversos tipos de textos médicos podem ter seu desempenho elevado quando treinados sobre um padrão ouro específico ao domínio. Desta forma, este trabalho foi realizado buscando estudar as metodologias e identificar as ferramentas utilizadas, bem como os processos que rodeiam a tarefa de anotação e compilação de textos em um padrão ouro morfológico e sintático, visando a definição de uma metodologia para a construção de um corpus de textos médicos em português brasileiro. Foram analisadas 32 diferentes abordagens de anotação em diferentes idiomas e domínios de estudo incluindo a área de saúde. Além disso, foi possível avaliar 12 diferentes ferramentas de anotação. São apontados também os pontos relevantes de cada uma das metodologias e a composição das mesmas que resultou na técnica escolhida para construção do corpus para o grupo de pesquisa.

Métodos

Este trabalho foi elaborado a partir de uma revisão de literatura através das bases de dados *IEEE-Xplore*, *ACM*, *ScienceDirect* e *Google Scholar*, além de pesquisas realizadas diretamente na ferramenta de busca do Google por publicações referentes à elaboração de corpus morfológicos, sintático ou morfossintáticos, excluindo anotações voltadas apenas a informações semânticas ou relações entre entidades textuais ou conceitos. Não foi feita distinção do domínio de estudo dos textos ou o propósito de uso. Foram aceitas publicações referentes a metodologias e procedimentos para anotação de textos, bem como publicações referentes a ferramentas de suporte a anotação.

As palavras-chave utilizadas para a realização da busca foram “corpus (corpora) development”, “corpus (corpora) building”, “corpus annotation”, “annotated corpus” e “text annotation tool”. Buscou-se por publicações sem realizar nenhuma limitação referente a data de publicação.

Somando-se todas as bases de dados, foram obtidas 58 diferentes publicações (realizando a exclusão de repetições entre as diferentes bases), selecionadas através da análise dos títulos e/ou resumo,

de forma que fossem pertinentes ao tema buscado. A partir da leitura destes, foram excluídas 11 publicações por se tratarem de linguagens com estruturas muito distantes do Português Brasileiro (tais como Árabe, Chinês ou Japonês) ou por não terem foco na metodologia ou ferramenta utilizadas, mas sim na forma de aquisição dos dados anotados ou em comparações de desempenho entre diferentes padrões ouro. Foram também excluídas 3 publicações por se tratarem de outras revisões de literatura. Dessa forma, foram analisadas 44 publicações que preencheram os critérios inicialmente propostos.

Dentre os estudos revisados, foram encontradas 12 ferramentas disponíveis para uso. Tais ferramentas foram instaladas e testadas, visando identificar as suas funcionalidades, e características que facilitem e/ou reduzam o trabalho manual dos anotadores. A partir da leitura foram identificados como critérios importantes a possibilidade de pré-anotar os textos automaticamente, uma interface de anotação simples, intuitiva e que permita anotação rápida, o suporte à anotação na de textos da língua portuguesa e a existência de estatísticas referentes a concordância entre diferentes anotadores. Verificou-se também a licença da ferramenta, a possibilidade da alteração do código para inserir funcionalidades (*OpenSource*) e a existência de suporte ou documentação.

Por fim, selecionou-se a ferramenta que mais se enquadrou dentro dos quesitos citados para ser utilizada como base no processo de construção de um corpus morfossintático de textos médicos em Português Brasileiro.

Resultados e Discussão

O objetivo deste estudo foi apresentar e discutir os achados da literatura referentes à metodologia de anotação de textos para a construção de padrões ouro para treinamento de algoritmos de Recuperação de Informação em Processamento de Linguagem Natural, ressaltando os pontos que se mostram importantes para o andamento do processo. Porém, julgou-se também válido realizar a análise e avaliação de diferentes ferramentas de apoio à anotação dos textos encontradas nos estudos avaliados, visto que são parte importante dentro do processo de construção.

Metodologias de Anotação – A maioria dos autores não entram em detalhes sobre o processo utilizado para a anotação em si, focando principalmente nos resultados obtidos após a construção do corpus, bem como a avaliação de desempenho do mesmo para treinamento de etiquetadores distintos. Esta tendência fez com que a obtenção das informações desejadas por este estudo fosse prejudicada. Dessa forma, dentro dos estudos avaliados foram excluídos todos os que não detalhavam a metodologia de anotação ou ferramenta utilizada, reduzindo para apenas 12 trabalhos de construção efetivamente avaliados, como pode ser observado na Quadro 1.

Uma grande exceção é o projeto de desenvolvimento do Penn Treebank¹, que forneceu a metodologia detalhada de estudo, avaliação e construção do corpus. Este foi um dos principais estudos de anotação de textos em inglês, sendo considerado como referência e que influenciou diversos estudos posteriores, mesmo em outros idiomas. Seu desenvolvimento utilizou um processo de anotação semiautomática, onde os textos foram pré-etiquetados por um algoritmo estocástico (o qual foi previamente treinado utilizando o Brown Corpus). Dessa forma, os anotadores tiveram o trabalho de revisar e corrigir o que foi etiquetado incorretamente na fase automática. Segundo experimentos realizados durante o processo de anotação, percebeu-se que o processo de anotação manual levava duas vezes mais tempo que a simples correção. Além disso, o nível de discordância entre as etiquetas de diferentes anotadores foi reduzido drasticamente com a inserção do passo automático.

O uso do processo de pré-anotação automático foi utilizado em todos os outros estudos avaliados, com exceção dos de Dinakaramani¹⁰ e Alexin¹¹, onde não haviam informações pré-existentes para a realização do treinamento de um *tagger*. A abordagem de Nguyen¹², pelo mesmo motivo, utilizou um processo totalmente manual para as primeiras sentenças, passando a treinar um etiquetador e etiquetar as próximas sentenças com os dados anotados.

Apesar de se mostrar como uma boa alternativa para reduzir tanto o tempo de anotação quanto aumentar a precisão da informação anotada, mesmo quando o etiquetador utilizado não é tão preciso, o uso de uma técnica de pré-anotação deve levar em conta a propensão que um anotador tem em seguir a escolha do etiquetador automático.

Quadro 1 - Análise das metodologias de anotação

Estudo	Idioma	Tipo de Textos	Tipo de Anotação	Processo Semiautomático	Tamanho do Corpus	Múltiplos Anotadores	Ferramenta Utilizada	Interface de Anotação
Marcus (1993) ¹	Inglês	Jornalístico	Morfológica e Sintática	Sim	4,5 milhões de palavras	Sim, 4 anotadores	PARTS, Fiddich	Interface em Lisp
Pala (1998) ¹³	Tcheco	<i>Não informado</i>	Morfológica	Sim	1 milhão de palavras	<i>Não informado</i>	LEMMA	CQP
Brants (2002) ¹⁴	Alemão	Jornalístico	Morfológica e Sintática	Sim	35 mil sentenças	Sim, 2 anotadores	TnT Tagger, LFG Parsing	Annotate
Alexin (2003) ¹¹	Húngaro	Diversos	Morfológica	Não	1 milhão de palavras	<i>Não informado</i>	HuMor parser	Ferramenta própria
Galicia-Haro (2003) ²	Espanhol	Internet	Morfológica	Somente automática	1592 MB em texto	Não	MACO	Nenhuma
Pakhomov (2006) ³	Inglês	Narrativas Clínicas	Morfológica	Sim	100 mil palavras	Sim, 3 anotadores	MaxEnt Tagger	Editor de XML
Huseth (2007) ¹⁵	Norueguês	Prontuário Eletrônico	Morfológica e Sintática	Sim	74 mil palavras	Não, 1 anotador	Tagger próprio	Ferramenta gráfica
Areta (2007) ¹⁶	Basco	Livros Científicos	Morfológica e Sintática	Sim	1,6 milhões de palavras	Não	<i>Não informado</i>	Corpusgile, Eulia
Nguyen (2009) ¹²	Vietnamita	Jornalístico	Morfológica e Sintática	Início manual, depois semiautomático	210 mil palavras	Sim, no mínimo 2 anotadores	vnTokenizer, JvNTagger, MaxEnt	SynAF
Dandapat (2009) ¹⁷	Bangla e Hindi	<i>Não informado</i>	Morfológica	Sim	<i>Não informado</i>	Sim	Tagger Próprio	Ferramenta gráfica
Wazsczuk (2010) ¹⁸	Polonês	Jornalístico	Sintática	Sim	85 mil sentenças	Sim, no mínimo 2 anotadores	Spejd	TrEd
Peters (2010) ⁸	Português	Narrativas Clínicas	Morfológica	Sim	123 mil palavras	Não	OpenNLP	MALT
Dinakaramani (2014) ¹⁰	Indonésio	Jornalístico	Morfológica	Não	260 mil palavras	Sim, 2 anotadores	Nenhuma	<i>Não informado</i>

Outra característica que se julgou muito interessante para um processo de anotação foi a utilização de um aprendizado interativo e treinamento incremental do etiquetador utilizado, permitindo que maior quantidade de informação seja anotada com qualidade, reduzindo o esforço humano. Este processo, conhecido como *Active Learning*, é uma abordagem onde um etiquetador é treinado com um modelo baseado em uma pequena quantidade informação de padrão ouro e na medida em que mais informações são fornecidas por um “oráculo” (no caso, um anotador), tal modelo é atualizado e o etiquetador retreinado. Isso permite com que o desempenho do etiquetador aumente gradativamente

a medida em que mais informações de treinamento são providas a ele¹⁹we are constrained by a fixed budget. A fully annotated corpus is required, but we can afford to label only a subset. We train a Maximum Entropy Markov Model tagger from a labeled subset and automatically tag the remainder. This paper addresses the question of where to focus our manual tagging efforts in order to deliver an annotation of highest quality. In this context, we find that active learning is always helpful. We focus on Query by Uncertainty (QBU). Esta técnica foi utilizada no estudo de Huseth¹⁵, onde cada texto anotado foi adicionado ao conjunto de treinamento do etiquetador (baseado em modelos ocultos de Markov), fazendo com que as probabilidades fossem recalculadas e o próximo texto fosse etiquetado em seguida. Sua avaliação mostra que a precisão do etiquetador automático, subiu de menos de 80% para 95% do início ao final da anotação. Além disso, os anotadores reportaram que o processo de anotação automática melhorou e que ficaram mais inclinados a confiar nas etiquetas sugeridas. O estudo de Peters⁸ também utilizou um processo similar de aprendizado ativo durante a anotação dos textos.

A presença de múltiplos anotadores é outra característica presente em quase todos os estudos avaliados, de forma a reduzir erros de anotação e evitar que as tendências de um anotador sejam transferidas para as informações anotadas. Porém, tal metodologia introduz a necessidade de uma avaliação de concordância entre os anotadores e uma forma de solucionar possíveis divergências.

Pode-se citar como exemplo a metodologia de Pakhomov³, onde foram realizadas discussões sobre os dados e os possíveis casos de dificuldade e, periodicamente, “sessões de consenso” para analisar as divergências encontradas. Nestas seções além de estarem presentes os anotadores, havia o auxílio de um pesquisador de NLP. No estudo de Nguyen¹², cada sentença anotada foi revisada por no mínimo dois anotadores, sendo que um deles é responsável por realizar a revisão do algoritmo automático e o outro revisa o trabalho realizado pelo primeiro.

O treinamento dos anotadores também se mostra como um fator importante para redução do tempo de anotação, até mesmo no caso de anotadores do domínio linguístico. No estudo de Marcus¹, foi realizado um treinamento de 15 horas de correção e 6 horas de etiquetagem. Percebeu-se que a curva de aprendizado para etiquetagem morfológica leva menos de um mês com a velocidade de anotação chegando a 3.000 palavras por hora. Já no caso da anotação sintática, devido a maior complexidade da análise a ser realizada, a anotação após seis semanas de trabalho foi de 375 a 475 palavras por hora. O uso de uma árvore sintática simplificada permitiu um aumento de 100 a 200 palavras por hora sobre o que havia sido obtido.

Ferramentas de Anotação – O uso de ferramentas de suporte à anotação busca facilitar o processo e reduzir o tempo e esforço necessário para concluir a tarefa. Geralmente possuindo interfaces gráficas, essas ferramentas possibilitam que o anotador realize seu trabalho interagindo com o computador de forma amigável.

Dentre os estudos avaliados, foram encontradas 12 ferramentas de suporte à anotação que foram avaliadas e testadas, sendo estas: Annotate (<http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/annotate.html>), Argo²⁰, Atomic²¹, Brat⁷, CLaRK²², GATE²³, Knowtator⁶, Lexical Annotation Workbench (<http://ufal.mff.cuni.cz/~hana/law.html>), MMAX2⁵, UAM Corpus Tool²⁴, WebAnno⁴ e Wordfreak²⁵. As ferramentas variam em questão de objetivo, sendo que algumas são voltadas para a notação manual, outras para adjudicação (revisão de textos já anotados), porém a maioria suporta a realização de ambas tarefas.

Não foi possível realizar a instalação das ferramentas Annotate e Lexical Annotation Workbench, pois não foi possível encontrar seus arquivos executáveis ou instaladores disponíveis na *internet*.

Como se observa no Quadro 2, a avaliação das ferramentas foi realizada com base nas características levantadas através do estudo das metodologias de anotação, buscando comparar as funcionalidades que aceleram o processo de anotação, como possibilidade de pré-anotação automática e aprendizado ativo.

Quadro 2 - Análise das ferramentas de suporte à anotação.

Ferramenta	Processo de Anotação Morfológica Manual	Processo de Anotação Sintática Manual	Visualização sintática	Etiquetagem Automática	Aprendizado Ativo	Importa Texto Etiquetado	Estatística de concordância	Etiqueta personalizada	Suporte ao Português	Formato de Saída	OpenSource	Tipo / linguagem
Annotate	<i>Não testado</i>	<i>Não testado</i>	<i>Não testado</i>	Sim	Não	<i>Não testado</i>	Não	Não	Não	<i>Não testado</i>	Não	Desktop (C)
Argo ²⁰	Seleciona-se o texto referente à palavra, clica-se em “Create” e seleciona-se ou digita-se a etiqueta. Texto fica marcado com a cor da etiqueta.	Igual ao morfológico, porém seleciona-se a sentença desejada.	Não	Sim	Não	Não	Não	Não	Sim	<i>Não testado</i>	Não	Aplicativo Web
Atomic ²¹	Arrasta-se a etiqueta até a posição desejada no texto, então atribui-se o valor da etiqueta digitando.	Arrasta-se o componente da estrutura até a posição e ajusta-se os limites, então atribui-se o valor da etiqueta digitando.	Não	Não	Não	Sim	Não	Sim	Sim	Diversos	Sim	Desktop (Java)
Brat ⁷	Duplo clique sobre a palavra, seleção da etiqueta dentro de poucas possibilidades. Etiqueta aparece sobre a palavra	Selecionar a sentença e clicar duas vezes, seleção de etiqueta dentro de poucas possibilidades. Permite relacionar etiquetas. Linhas que indicam a região da etiqueta aparecem.	Linhas de relação	Sim	Não	Sim	Não	Sim	Sim	ANN e TXT	Sim	Aplicativo Web
CLaRK ²²	<i>Não testado</i>	<i>Não testado</i>	Em árvore	Não	Não	Sim	Não	Sim	Sim	XML	Não	Desktop (Java)
GATE ²³	Seleciona-se o texto referente à palavra, a janela de anotação se abre e então digita-se o tipo de anotação e seleciona-se a etiqueta.	Igual ao morfológico, porém seleciona-se a sentença desejada.	Não	Sim	Não	Sim	Não	Sim	Sim	Diversos	Sim	Desktop (Java)
Knowtator ⁶	Seleciona-se o texto referente à palavra e então clica-se na etiqueta correspondente. A cor da etiqueta é exibida na palavra	Igual ao morfológico, porém seleciona-se a sentença desejada.	Não	Não	Não	Sim	Sim	Sim	Sim	XML	Sim	Plug-in (Protegé) (Java)
Lexical Annotation Workbench	<i>Não testado</i>	<i>Não testado</i>	<i>Não testado</i>	Não	Não	Não	Não	Sim	Não	PDT e TNT	Sim	Desktop (Java)

MMAX2 ⁵	Seleciona-se o texto referente à palavra para criar um item marcável. Clicar no item gerado abre a janela para selecionar a etiqueta em uma lista. Etiqueta não é mostrada junto ao texto.	Da mesma forma que o morfológico, porém seleciona-se a sentença. Etiqueta não é mostrada junto ao texto.	Não	Não	Não	Não	Sim	Sim	Sim	XML	Sim	Desktop (Java)
UAM Corpus Tool ²⁴	Clicar sobre a palavra e selecionar etiqueta em uma lista	Selecionar sentença e escolher etiqueta.	Não	Sim	Não	Não	Sim	Sim	Sim	XML e PTB	Não	Desktop (Java)
WebAnno ⁴	Clicar sobre a palavra e selecionar etiqueta em uma lista pré-definida. A etiqueta aparece sobre a palavra.	Selecionar sentença e escolher etiqueta. Relações podem ser inseridas arrastando sobre outra. A estrutura é mostrada através de linhas com a etiqueta.	Linhas de relação	Sim	Sim	Sim	Sim	Sim	Sim	Diversos	Sim	Aplicativo Web (Java)
Wordfreak ²⁵	Texto é exibido em formato de árvore, clica-se na palavra e em uma tabela seleciona-se a etiqueta. Etiqueta é exibida na coluna ao lado da palavra.	Clica em um ramo para alterar sua etiqueta. Modifica a estrutura por botões de seta que controlam os níveis na árvore. Etiqueta é exibida na coluna ao lado do nó da árvore.	Em árvore	Sim	Sim	Não	Não	Não	Sim	XML	Sim	Desktop (Java)

A arquitetura de funcionamento dos sistemas estudados basicamente se estendeu por aplicativos do tipo *Desktop* ou *Web*, sejam estes funcionando através da nuvem (como o caso do Argo) ou através de servidores que podem ser locais. Os aplicativos do tipo *Web* apresentam a vantagem de armazenarem os dados de forma centralizada em um servidor e permitir o acesso de diversos anotadores ao mesmo tempo e de diferentes locais para a realização do processo de anotação. Vale também ressaltar que a ferramenta WebAnno é a única que apresenta um sistema de gerenciamento de projetos, através do qual um gerente de projeto pode controlar o trabalho dos anotadores, visualizar seus resultados e acompanhar o andamento do projeto.

Foi analisado também o processo de anotação manual para inserção ou correção da anotação. Visto que os anotadores geralmente precisam de horas de treinamento antes de iniciar a anotação efetivamente, o uso de uma interface simples e intuitiva permite que seu aprendizado seja mais fácil e rápido. Além disso, uma interface que possibilite a etiquetagem em poucos passos é preferível, por reduzir o tempo de etiquetagem pôr termo e evitar o cansaço prematuro do anotador com a execução de tarefas repetitivas. A forma de visualização das etiquetas também é importante, principalmente para a estrutura sintática. Uma interface simples que permita uma visualização clara e concisa do que foi anotado diretamente em conjunto com o texto, sem a necessidade de abrir janelas ou caixas de diálogo adicionais também facilita o processo de anotação. No caso de uma anotação sintática, a estrutura pode se tornar complexa pelos diferentes níveis existentes, porém visualizações no formato de árvore ou através de ligações entre as classes é uma boa alternativa.

Para a construção do corpus anotado proposto pelo grupo de pesquisa, motivo pelo qual este estudo foi realizado, fez-se a análise do suporte das ferramentas à anotação de textos na língua portuguesa,

bem como a utilização de um conjunto de etiquetas personalizado, de forma permitir o uso do conjunto fornecido pelo OpenNLP (<http://opennlp.apache.org>) e CoGrOO (<http://cogroo.sourceforge.net>), os quais já vem sendo utilizados há 9 anos em nossos estudos.

Percebe-se também que um grande número de ferramentas tem seu código aberto, o que permite que suas falhas ou falta de alguma funcionalidade seja compensada pela integração com diferentes sistemas.

Uma característica interessante que não foi encontrada em nenhuma das ferramentas estudadas, foi a sugestão de etiquetas para a anotação baseada nos resultados estatísticos do etiquetador. No estudo de Huseth¹⁵, a ferramenta era capaz de exibir a lista de etiquetas ordenadamente de acordo com a probabilidade calculada. Esta funcionalidade, entretanto, pode não ser tão útil em casos que haja a utilização de aprendizado ativo, pois a tendência é que o etiquetador cometa cada vez menos erros, não justificando o esforço para a implementação de tal funcionalidade.

Com a avaliação dos estudos de anotação, percebeu-se que diversas ferramentas foram desenvolvidas internamente aos grupos de pesquisa responsáveis e não foram liberadas à comunidade científica para uso. Além disso, a avaliação de algumas ferramentas foi difícil ou prejudicada pela documentação pobre ou até inexistente, como foi o caso das ferramentas Brat, CLaRK, MMAX2, UAM Corpus Tool e Wordfreak.

Percebeu-se que muitas ferramentas apresentam limitações quanto à uma visualização gráfica da estrutura semântica anotada. Muitas não possuíam tal função, enquanto outras se limitavam a estruturas simplificadas ou apenas parte da árvore sintática da sentença. Além disso, poucas ferramentas apresentam a avaliação de estatísticas do processo e anotação.

Outra grande limitação encontrada em grande parte das ferramentas é a falta da funcionalidade de anotação automática por um algoritmo etiquetador, forçando com que todo o trabalho realizado seja puramente manual. Além disso, nenhuma das ferramentas avaliadas foi capaz de realizar a geração automática de estruturas sintáticas das sentenças.

Com a análise realizada, a ferramenta WebAnno⁴ se mostrou como a mais adequada para a utilização no processo de anotação futuramente proposto. A ferramenta, além de possuir extensa documentação, foi claramente a que apresentou a interface mais amigável e simplificada, porém funcional, fazendo com que a curva de aprendizado para sua utilização seja rápida e suave. Esta é uma característica muito importante, visto que, em muitas situações, os anotadores não pertencem a área de tecnologia e a dificuldade de familiarização com o *software* pode impactar seriamente no desempenho do processo de anotação. Além disso, esta foi a única das ferramentas testadas que apresentou todos os aspectos julgados importantes para o processo de anotação (como apresentado no Quadro 2). Em especial, ressalta-se o suporte para a pré-anotação dos textos da língua portuguesa utilizando um conjunto de etiquetas personalizado. Isto permite que, em conjunto com a importação de textos já anotados, todos os dados que possuímos podem ser utilizados para agilizar o processo de anotação, além de permitir a produção de um corpus anotado de maior qualidade.

Entretanto, é importante ressaltar que as ferramentas são diversas em funcionalidades e diferem substancialmente em muitos aspectos, como sua arquitetura básica, suporte à automação, formatos de visualização dos textos e características que podem ser anotadas. Dessa forma, é difícil definir uma única ferramenta que possa ser aplicada em qualquer situação. Além disso, caso os requerimentos de anotação de um projeto não sejam muito específicos, é possível encontrar uma ferramenta que seja capaz de fornecer as funcionalidades necessárias.

Conclusão

Neste trabalho foram avaliadas 12 metodologias de diferentes projetos de compilação e anotação de textos morfológica e/ou sintaticamente, bem como analisadas e avaliadas outras 12 ferramentas de

suporte à anotação. Este estudo serve como guia para futuros projetos de construção de corpus anotados, de forma a conhecer diferentes estudos com abordagens distintas para a solução de um mesmo problema, podendo assim, ser identificadas as características pertinentes para um dado projeto.

Quanto às ferramentas de anotação, percebe-se a existência de uma extensa gama de ferramentas pertinentes para anotação morfossintática. Entretanto, todas possuem seus pontos fortes e fracos, sendo necessária a avaliação de qual pode ser utilizada para cada caso. Nenhuma das ferramentas é capaz de satisfazer todas as necessidades, porém soluções capazes de resolver diversos problemas existem. Além disso, muitas ferramentas são disponibilizadas com seu código fonte aberto, o que pode contribuir para a integração com outros sistemas e complementação de suas deficiências.

Referências

- [1] Marcus MP, Marcinkiewicz MA, Santorini B. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics* 19.2 (1993): 313-330.
- [2] Galicia-Haro SN. Using electronic texts for an annotated corpus building. *Proc Mex Int Conf Comput Sci.* 2003;2003-Janua:26-32.
- [3] Pakhomov SV, Coden A, Chute CG. Developing a corpus of clinical notes manually annotated for part-of-speech. *Int J Med Inform.* 2006;75(6):418-429.
- [4] Yimam SM, Gurevych I, Eckart de Castilho R, Biemann C. WebAnno: A flexible, web-based and visually supported system for distributed annotations. *Proc 51st Annu Meet Assoc Comput Linguist Syst Demonstr.* 2013;1(1):1-6.
- [5] Müller C, Strube M. Multi-level annotation of linguistic data with MMAX2. *Corpus Technol Lang Pedagog New Resour new tools, new methods* 3. 2006:197-214.
- [6] Ogren, PV. Knowtator : A Protégé plug-in for annotated corpus construction. *Proc 2006 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol companion Vol Demonstr.* 2006;(June):273-275.
- [7] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. *Proc Demonstr 13th Conf Eur Chapter Assoc Comput Linguist.* 2012;(1):102-107.
- [8] Peters AC, Oleynik M, Pacheco EJ, Moro CMC, Schulz S, Nohama P. Elaboração de um Corpus Médico baseado em Narrativas Clínicas contidas em Sumários de Alta Hospitalar. *An do XII Congr Bras Informática em Saúde.* 2010;(1).
- [9] Afonso S, Bick E. Floresta Sintá(c)tica: um “treebank” para o português. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas de Gran Canaria, Espanha; 2002:1698-1703. Disponível em <http://www.linguateca.pt/documentos/AfonsoetalAPL2001.pdf>. Acesso em 25 mai. 2016.
- [10] Dinakaramani A, Rashed F, Luthfi A, Manurung R. Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. *Proc Int Conf Asian Lang Process 2014, IALP 2014.* 2014:66-69.
- [11] Alexin Z, Gyimóthy T, Hatvani C, et al. Manually annotated Hungarian corpus. *Proc tenth Conf Eur chapter Assoc Comput Linguist - EACL '03.* 2003;2:53.
- [12] Nguyen PT, Vu XL, Nguyen TMH, Nguyen VH, Le HP. Building a Large Syntactically-Annotated Corpus of Vietnamese. *Proc Third Linguist Annot Work.* 2009;(August):182-185.
- [13] Pala K, Rychlý P, Smrž P. Corpus Annotation in Inflectional Languages: Czech. In: *Proceedings of the Ninth International Workshop on Database and Expert Systems Applications.* 1998:149-153.
- [14] Brants S, Dipper S, Hansen S, Leizius W, Smith G. The TIGER treebank. *Proc Work treebanks Linguist Theor.* 2002:24-41. Disponível em <http://www.coli.uni-saarland.de/publikationen/softcopies/Brants:2002:TT.pdf>. Acesso em 25 mai. 2016.

- [15] Huseth O, Røst TB. Developing an annotated corpus of patient histories from the primary care health record. Proc - 2007 IEEE Int Conf Bioinformatics and Biomed Work BIBMW. 2007:165-173.
- [16] Areta N, Gurrutxaga A, Leturia I, et al. ZT Corpus: Annotation and tools for Basque corpora. Corpus Linguist 2007. 2007;Birmingham:1-19.
- [17] Dandapat S, Biswas P, Choudhury M, Bali K. Complex Linguistic Annotation – No Easy Way Out! A Case from Bangla and Hindi POS Labeling Tasks. Proc 3rd Linguist Annot Work LAW '09. 2009;(August):10-18.
- [18] Waszczuk J, Głowińska K, Savary A, Przepiórkowski A. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. Proc Int Multiconference Comput Sci Inf Technol. 2010:531-539.
- [19] Ringger E, McClanahan P, Haertel R, et al. Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation. Proc Linguist Annot Work. 2007;(June):101-108.
- [20] Rak R, Rowley A, Black W, Ananiadou S. Argo: an integrative, interactive, text mining-based workbench supporting curation. Database: The Journal of Biological Databases and Curation. 2012;2012:bas010.
- [21] Druskat S, Bierkandt L, Gast V, Rzymiski C, Zipser F. Atomic: An Open-Source Software Platform for Multi-Level Corpus Annotation. Proc 12th Konf zur Verarbeitung natürlicher Spr. 2014:228-234..
- [22] Simov K, Peev Z, Kouylekov M, Simov A, Dimitrov M. CLaRK - an XML-based System for Corpora Development. Proc Corpus Linguist 2001 Conf. 2001:558-560.
- [23] Cunningham H, Maynard D, Bontcheva K. Text Processing with GATE (Version 6). Gateway Press CA; 2011.
- [24] O'Donnell M. The UAM CorpusTool: Software for Corpus Annotation and Exploration. Appl Linguist Now Underst Lang Mind. 2008;00(April):1433-1447.
- [25] Morton T, Lacivita J. WordFreak: an open tool for linguistic annotation. Proc 2003 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Demonstr Vol 4. 2003;4:17-18.

Contato

Gabriel Herman Bernardim Andrade
Estudante de Engenharia de Computação,
Escola Politécnica, Pontifícia Universidade
Católica do Paraná. Curitiba, Paraná, Brasil.
Telefone: +55 41 8778-6743
E-mail: gabrielhbandrade@outlook.com.br
Endereço: R. Imaculada Conceição, 1155 –
Rebouças, Curitiba/PR