A REVISÃO DA ETIQUETAGEM MORFOSSINTÁTICA DE UM CORPUS DE TEXTOS ACADÊMICO DA UNILAB.

Raquel Furtado Mesquita [1], Tiago Martins da Cunha [2]

RESUMO

Os estudos sobre o Processamento de Linguagem Natural no Brasil têm recentemente despertado o interesse de acadêmicos e entusiastas. No entanto, a escassez de recursos limita o desenvolvimento tecnológico que faz uso desses recursos. Dessa forma, realizamos a coleta de textos acadêmicos produzidos nas disciplinas de Leitura e Produção Textual I e II para compor um corpus de textos acadêmicos da UNILAB. Esse corpus foi construído para futuros acessos de pesquisa. Esse corpus foi enriquecido com informações morfossintáticas, i.e. foram atribuídas a classe gramatical para cada palavra do corpus. Esse processo apesar de automático apenas acerta 96% das palavras, ainda requerendo uma revisão humana para chegar aos 100% e tornar-se um corpus *Gold-Standard*. Nessa pesquisa visamos realizar a revisão em um padrão Ouro (Gold-Standard) , mas a revisão foi incompleta.Foi realizada a revisão de todas as etiquetas envolvendo substantivos e verbos, ou seja, mais de 70% do corpus já está revisado em seu padrão-ouro. As atividades desse projeto agem diretamente sobre o projeto realizados em 2016 e 2017, que realizou a criação e adaptação do programa de etiquetagem automática, criação de um corpus inicial de textos acadêmicos produzidos na UNILAB e anotação inicial desse corpus. Este corpus ainda requer uma expansão do quanto ao seu tamanho e sua qualidade. Um corpus etiquetado em um padrão Ouro pode ser utilizado com base para a etiquetagem automática de novos textos. Este corpus pode, também, ser utilizado em uma gama de pesquisa linguísticas com o intuito da investigação dos aspectos morfossintáticos, sintáticos e lexicais a cerca do discurso contido no corpus.

Palavras-chave:

Linguística Computacional. Linguística de Corpus. Etiquetagem.

[1] UNILAB, ILL, Discente, e-mail: raquelfurtadom@gmail.com
[2] UNILAB, ILL, Docente, e-mail: tiagotmc@unilab.edu.br

INTRODUÇÃO

O Processamento de Linguagem Natural (PLN) tem sido responsável pelo desenvolvimento tecnológico que podemos observar nas ultimas duas décadas. O motivo para esse desenvolvimento é o casamento entre a criação de ferramentas acessíveis e a disponibilização de recursos. Os recursos que nos referimos são os linguísticos e as ferramentas citadas são aquelas responsáveis pelo processamento desses recursos linguísticos.

Há uma gama de etapas e abordagens diferentes a serem escolhidas durante qualquer estudo de processamento de linguagem natural. Dessa forma, buscamos com esse projeto disponibilizar recursos e ferramentas para auxiliar nos estudos dessa área a Linguística Computacional (LC). O Brasil, muito recentemente, tem se preocupado em gerar recursos e ferramentas para a LC. Essa preocupação em países do primeiro mundo data os anos 50. Em países do primeiro mundo, os estudos de LC confundem-se com a surgimento de estudos nas áreas da Inteligência Artificial, Tradução Automática e a própria Ciência da Computação.

Os recursos no Brasil na área de LC ainda não são representativos. Estes recursos são bancos de dados linguísticos, como: dicionários, gramáticas, corpora e etc. Assim, visamos, como produto dessa pesquisa, disponibilizar um corpus anotado morfossintáticamente para a comunidade acadêmica.
O corpus será composto pelo discurso português escrito de alunos estrangeiros da UNILAB e será anotado morfossintáticamente com um conjunto de etiquetas ainda a ser definido. O processo de etiquetagem morfossintática exige alguns recursos como etiquetadores e toquenizadores. Essas ferramentas serão desenvolvidas ao longo dessa pesquisa.

O processo de etiquetagem de um corpus exige um árduo e criterioso trabalho de revisão. Esta tarefa será realizada na etapa final desta pesquisa por voluntários. Os voluntários serão devidamente selecionados e treinados.
Uma vez que o corpus esteja devidamente etiquetado e revisado, ele já poderá ser utilizado como parâmetro na etiquetagem de novos corpora ou em processos de estudos linguísticos ou outras etapas de etiquetagem, i.e. anotação sintática e/ou semântica.

METODOLOGIA

A criação de ferramentas computacionais para o PLN requer os recursos básicos de qualquer pesquisa computacional, i.e. um computador. A utilização de sistemas operacionais da família UNIX possibilitam a utilização de uma gama de recursos de programação e softwares gratuitamente disponibilizados na rede mundial de computadores (internet).
Nativa da sistema operacional UNIX, Python é uma linguagem de programação gratuita que apresenta uma vasta biblioteca e recursos disponíveis para o PLN. O Natural Language Tool Kit (NLTK) é uma biblioteca de módulos de Python especialmente desenvolvida para o PLN.
O uso dos módulos do NLTK possibilitam a criação de diversas ferramentas de PLN, inclusive a ferramenta de etiquetagem morfossintática visada nesta pesquisa. Além da ferramenta, é necessário a utilização recursos para a modelação dos padrões a serem utilizados na etiquetagem. Esses padrões serão utilizados dos diferente corpora apresentados na seção anterior.

Após a escolha do modelo e conjunto de etiquetas morfossintáticas ideais para compor o nosso corpus, eles serão atribuídos ao nosso corpus e submetido a revisão. O processo de revisão será realizado através de recursos gratuitos de editores de texto e para a resolução conflitos utilizaremos softwares gratuitos de controle de versão como SVN e Tortoise.
Os recursos de controle de versão são úteis para a visualização de discrepâncias no processo de revisão e auxiliam os revisores a concordarem a resposta. Caso seja necessário, um programa para o pareamento de etiquetas entre revisores será construído para auxiliar a logística de revisão.

A logística do processo de etiquetagem morfossintática que estamos inclinados a seguir consiste na escolha de um modelo, atribuição de etiquetas e revisão. Os modelos são extraídos de um corpus previamente anotado morfossintáticamente, como comentamos anteriormente. A atribuição de etiquetas parte de um levantamento estatístico. Este levantamento estatístico mesmo bem calibrado ao contexto e gênero textual ainda está sujeito a uma margem de erro. O status atual de bons etiquetadores já academicamente testados para o português (ALENCAR, 2013) é de cerca de 95% de acerto. Os restantes 5% devem ser revisados e corrigidos manualmente para o corpus tornar-se uma versão ideal, em padrão Ouro (gold-standard).
O processo de construção da ferramenta de etiquetagem e atribuição a um corpus pode ser considerado relativamente rápido em relação ao processo de revisão. Diferentes corpora como Macmorpho e Tycho Brahe, citados anteriormente, apesar de considerarem-se em sua versão gold-standard ainda encontram-se em revisão constante, mesmo quase 10 anos após a sua disponibilização.
Nossa pesquisa visa a etiquetagem de um corpus significativo que possa equiparar-se aos disponíveis à comunidade acadêmica. Dessa forma visamos um corpus em padrão ouro de 500k de toquens no primeiro ano.

RESULTADOS E DISCUSSÃO

No desenvolvimento de nossa pesquisa realizamos a coleta e organização de 450k toquens e a revisão destes em um corpus anotado em padrão Ouro a ser disponibilizado como modelo para novas etiquetagens morfossintáticas.
Buscamos, também, ao longo dessa pesquisa, procuramos cativar o interesse de alunos, colaboradores e entusiastas para o desenvolvimento de pesquisas no âmbito da Linguística Computacional. Dessa forma, foram promovidos minicursos e oficinas preparatórios para o ingresso de voluntários no processo de revisão do corpus.

Ao darmos início a pesquisa em agosto de 2016, realizamos o inicio das coletas de textos e construção da ferramenta de etiquetagem. No ano seguinte da pesquisa foram realizadas as etapas de organização do corpus e revisão dos 450k toquens em textos coletados.

Apesar da revisão não ter sido concluída satisfatoriamente, é possível afirmar que todos os substantivos e verbos do corpus já foram revisados. Isso equivale em 70% de revisão concluída. Mesmo ainda não sendo um corpus em Padrão-ouro, este corpus já pode ser utilizado para pesquisas paralelas que visam investigar aspectos sintáticos e de gênero em relação às produções acadêmicas de estudantes da UNILAB.

CONCLUSÕES

Percebemos durante o período dessa pesquisa a necessidade da construção de um corpus de textos acadêmicos que sirva de referencial pedagógico ou de pesquisa para os próprios professores da UNILAB. Apesar do corpus não me mostrar ainda como uma referência para essa amostragem, faltam poucas etapas para essa ferramenta poder auxiliar outras pesquisas e investigações pedagógicas sobre as produções dos estudantes da Unilab.

AGRADECIMENTOS

Agradeço a Pro-Reitoria de Pesquisa pela oportunidade de desenvolvimento dessa pesquisa e ao CNPq pelo provimento de bolsas de pesquisa para duas estudantes, Rita e Lara Carolina, que puderam participar e colaborar com essa pesquisa.

REFERÊNCIAS

Aluísio, Sandra Maria, and Gladis Maria de Barcellos Almeida. "O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística." *Calidoscópio*4.3 (2006): 156-178.

Sardinha, Tony Berber. *Lingüística de corpus*. Editora Manole Ltda, 2004.

Severo, Cristine Gorski. "Política (s) linguística (s) e questões de poder." *ALFA: Revista de Linguística*57.2 (2013).

SHEPHERD, T. O estatuto da Linguística de Corpus: metodologia ou área da Linguística. Matraga, Rio de Janeiro, v. 16, n. 24, p. 150–172., 2009.