Corretor Gramatical Para o Emacs

Trabalho de Conclusão de Curso

Thiago Maciel batista

Orientador: Prof. Dr. Marcelo Finger
Instituto de Matemática e Estatística
Universidade de São Paulo

16 de novembro de 2010

Roteiro

Antes de começarmos...

Não confundir

Corretor gramatical não é corretor ortográfico

Principal objetivo do trabalho

Acoplar um corretor gramatical ao emacs

Usar o corretor gramatical CoGrOO

Mas como?

Como o CoGrOO consegue corrigir um texto?

Usando processamento de linguagem natural

Processamento de Linguagem Natural(PLN)

Definição

Área da inteligência artificial que se concentra no desenvolvimento de algoritmos para manipular linguagens naturais.

Serve para que?

O computador poderá interpretar e gerar textos em linguagem humana.

PLN Estatístico

Definição

- Utiliza modelos prababilísticos para processar a linguagem.

Conceitos e técnicas

- Token

Conceitos e técnicas

- Token
- N-grama

Conceitos e técnicas

- Token
- N-grama
- Modelos n-grama

Conceitos e técnicas

- Token
- N-grama
- Modelos n-grama
- O principio da máxima entropia

Conceitos e técnicas

- Token
- N-grama
- Modelos n-grama
- O principio da máxima entropia
- Corpus

Conceitos e técnicas

- Token
- N-grama
- Modelos n-grama
- O principio da máxima entropia
- Corpus
- Aprendizado de máquina

CoGrOO

A versão usada neste trabalho foi a 3.0.5

Descrição

O que é o CoGrOO?

- Corretor gramatical para a língua portuguesa do Brasil
- Feito para o editor OpenOffice.org
- Atualmente usa a linguagem java
- Utiliza PLN estatístico para processar os textos

Descrição

O que é o CoGrOO?

- Corretor gramatical para a língua portuguesa do Brasil
- Feito para o editor OpenOffice.org
- Atualmente usa a linguagem java
- Utiliza PLN estatístico para processar os textos

Descrição

O que é o CoGrOO?

- Corretor gramatical para a língua portuguesa do Brasil
- Feito para o editor OpenOffice.org
- Atualmente usa a linguagem java
- Utiliza PLN estatístico para processar os textos

Descrição

O que é o CoGrOO?

- Corretor gramatical para a língua portuguesa do Brasil
- Feito para o editor OpenOffice.org
- Atualmente usa a linguagem java
- Utiliza PLN estatístico para processar os textos

Tipos de erros que são detectados

São detectados erros relacionados a...

- Colocação pronominal
- Concordância nominal
- Concordância entre sujeito e verbo
- Concordância verbal
- Uso de crase
- Regência nominal
- Regência verbal
- Erros comuns da língua portuguesa escrita

Ferramentas auxiliares

Dependências

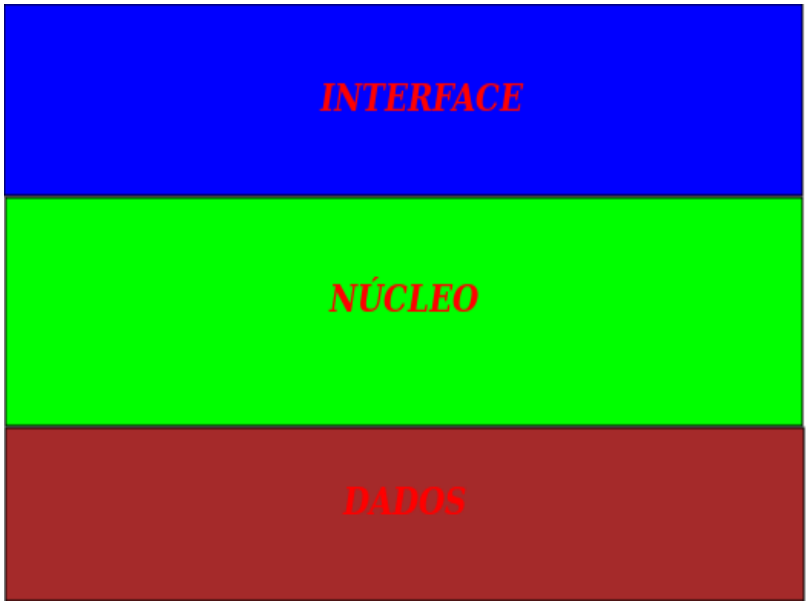
- OpenNLP - Framework para auxiliar no desenvolvimento de projetos de PLN
- MAXENT - Pacote que utiliza o conceito do modelo da máxima entropia para fazer aprendizado de máquina

Ferramentas auxiliares

Dependências

- OpenNLP - Framework para auxiliar no desenvolvimento de projetos de PLN
- MAXENT - Pacote que utiliza o conceito do modelo da máxima entropia para fazer aprendizado de máquina

Estrutura Geral



FAZ A INTERAÇÃO COM O APLICATIVO QUE IRÁ USAR O COGROO

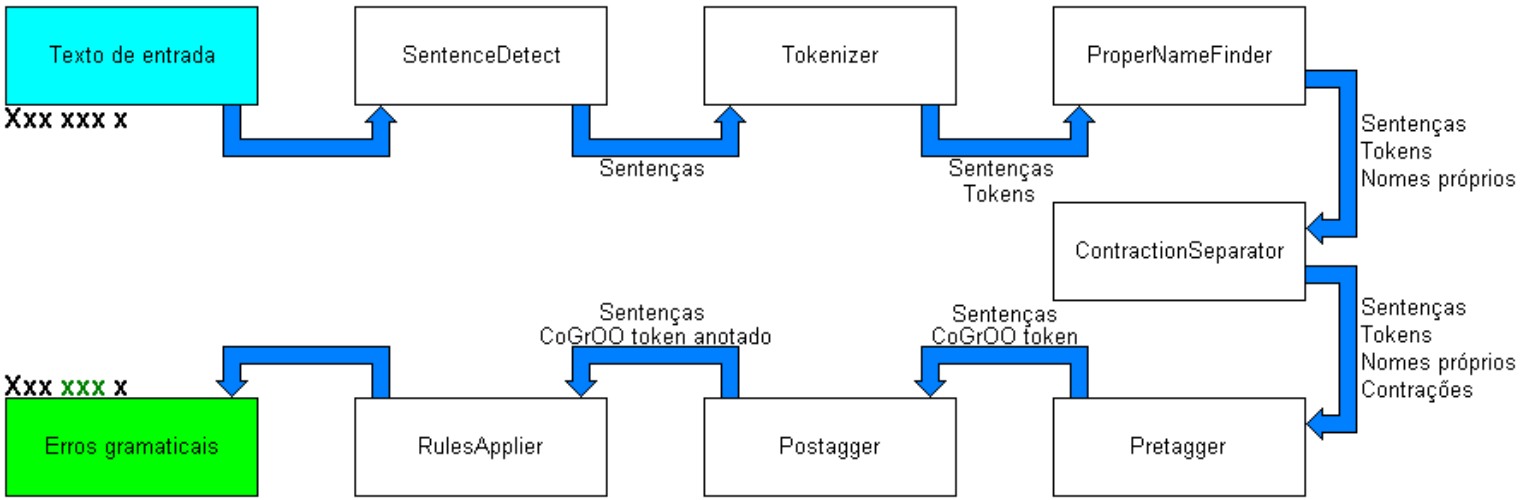
PROCESSA O TEXTO À PROCURA DE ERROS

DADOS PARA APRENDIZAGEM DE MÁQUINA E CONSULTA DURANTE O PROCESSAMENTO DO TEXTO

É COMPOSTO PELO CORPUS E POR UM DICIONÁRIO LÉXICO

Estrutura do núcleo

Principais módulos do núcleo do cogroo e o fluxo para processar textos



Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

Módulos do núcleo do cogroo

Função dos módulos

- *SentenceDetect*: Separa o texto em sentenças
- *Tokenizer*: Separa sentenças em tokens
- *ProperNameFinder*: Identifica nomes próprios no texto
- *ContractionSeparator*: Identifica contrações
- *Pretagger*: Prepara os tokens para receber classificação morfológica
- *Postagger*: Atribui a cada token uma classificação morfológica
- *RulesApplier*: Aplica as regras gramaticas e gera uma lista de erros

O que é?

- Editor de texto
- Poderoso ambiente de trabalho para diversos tipos de projeto
- Versão mais popular é o GNU Emacs
- Facilmente extensível e customizável por meio da linguagem emacs lisp
- Há varias extensões disponíveis

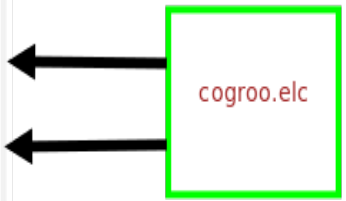
E agora?

Já temos o emacs e o CoGrOO...

E agora?

...mas e agora, como uni-los?

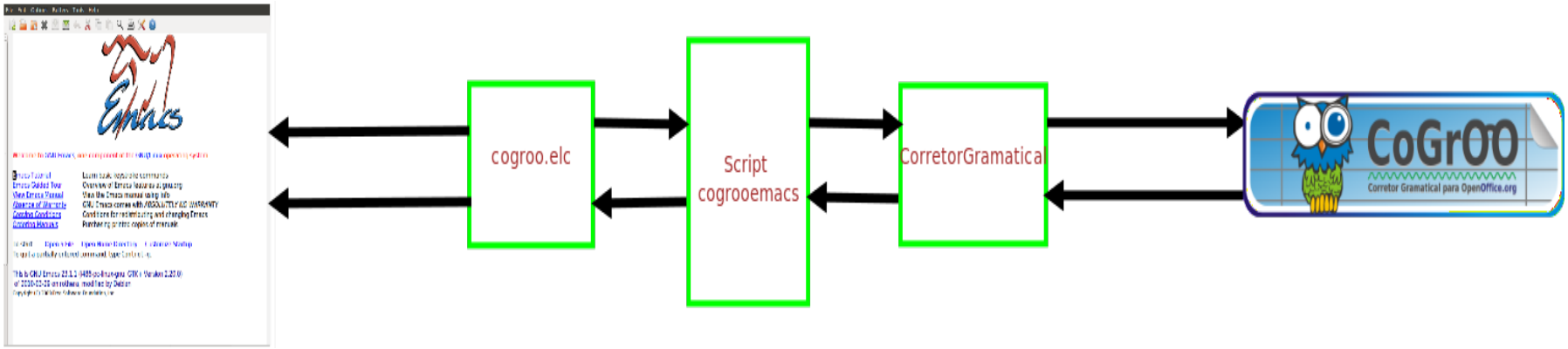
Interfaces



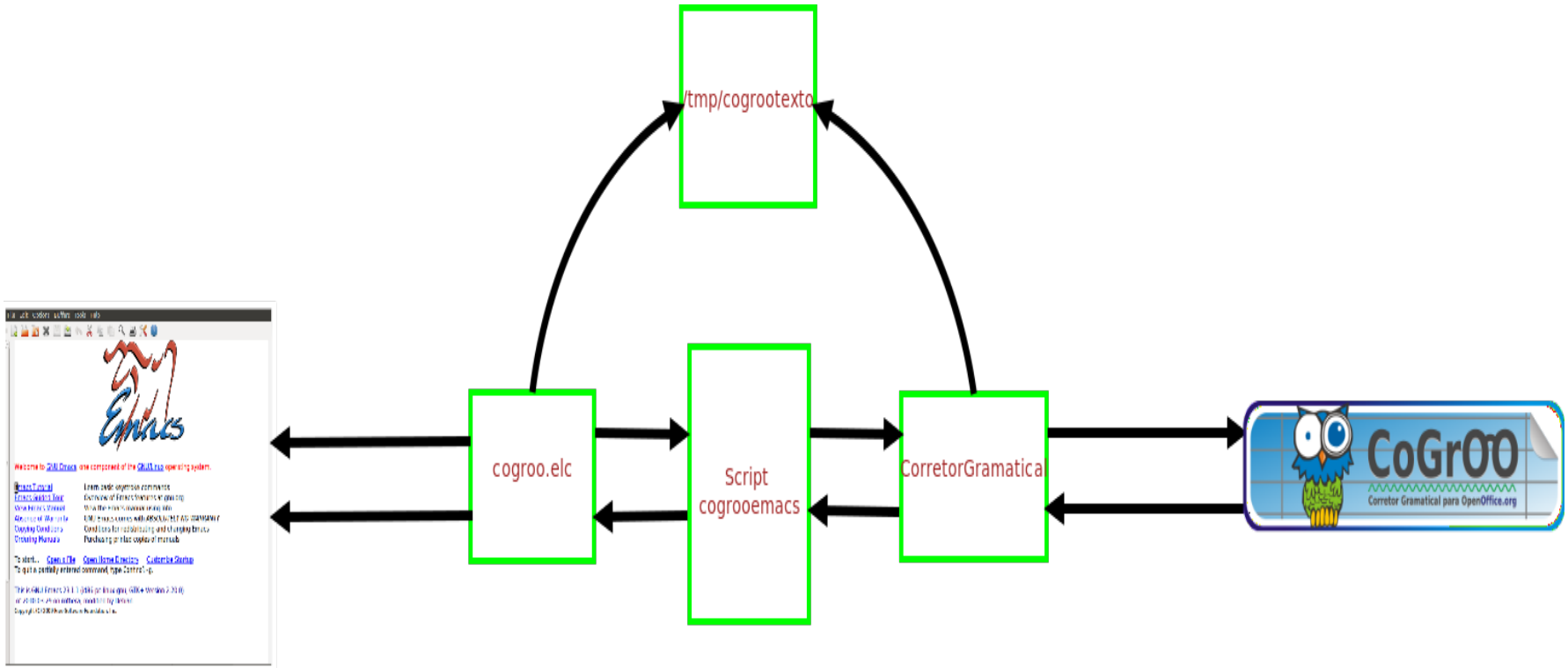
Mas...

...ainda não há interação

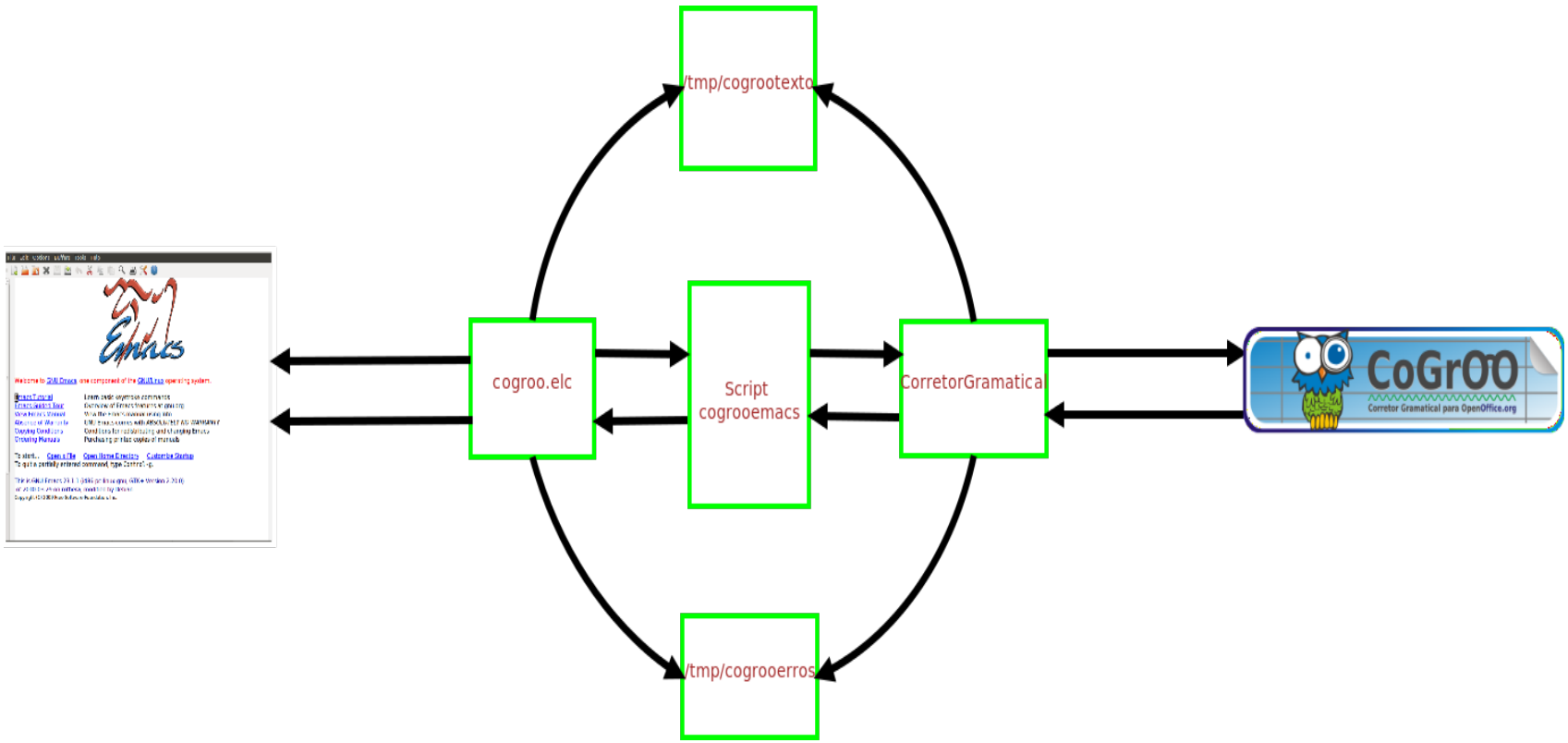
Script para a comunicação



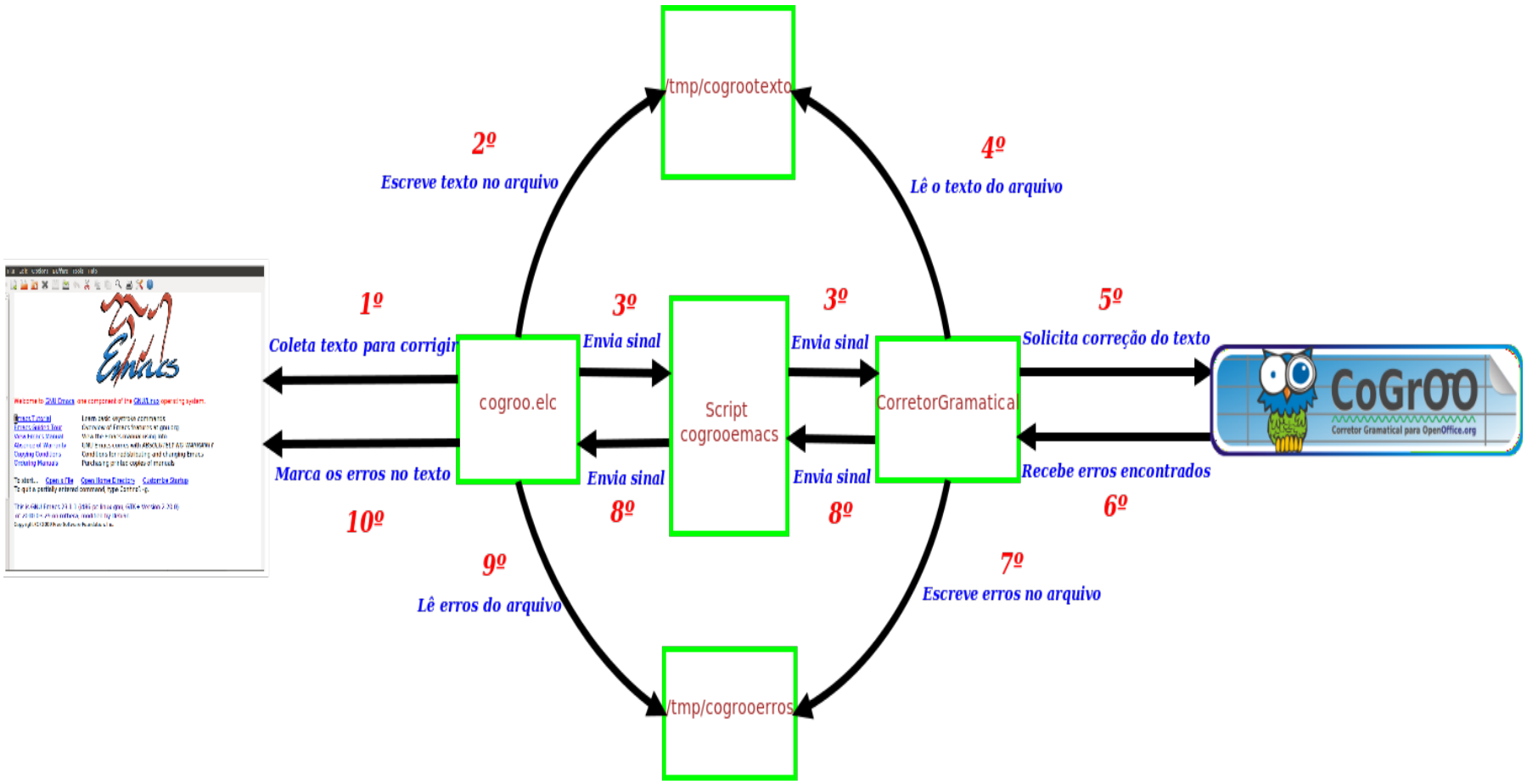
Arquivo para enviar o texto



Arquivo para enviar os erros

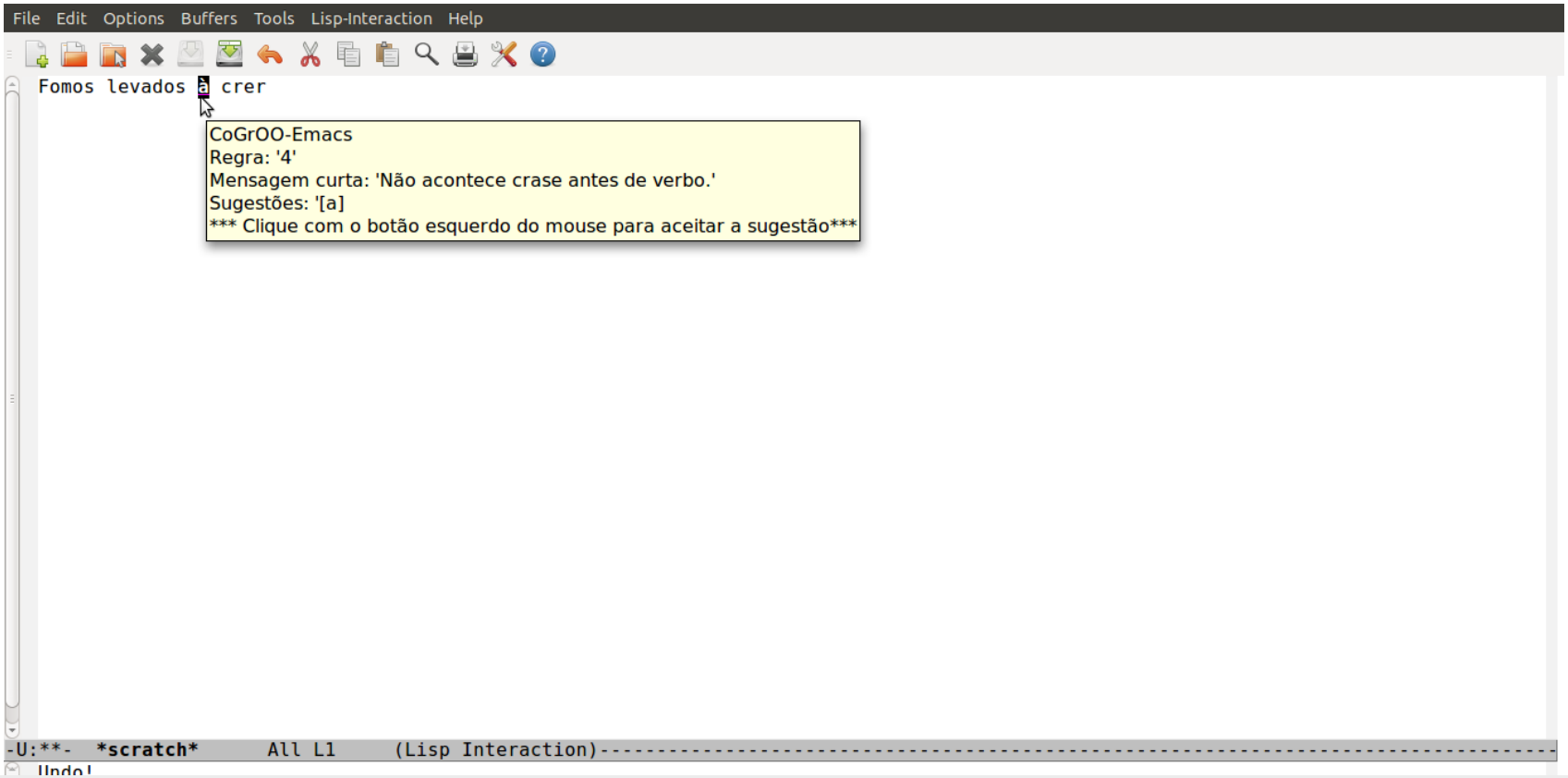


Fluxo para correção do texto



Resultado

O emacs já usa o cogroo para corrigir textos



Características

- Disponibiliza comandos para o usuário corrigir o texto
- Sublinha os trechos com erros
- Explica porque está errado
- Mostra sugestões de correção, quando possível
- Aceita a sugestão escolhida pelo usuário por meio do mouse

Trabalhos futuros

Tarefas

- Disponibilizar para os usuários
- Usar sugestões para melhorias
- Melhorar interação com os usuários
- Aperfeiçoar correção
- Reportagem automática de erros por meio do cogroo comunidade
- Aprendizado de máquina com os erros reportados

Referências

Diogo M. Pires , Fábio W. Gusukuma, Marcelo Suzumura, and William Colen.

Corretor gramatical acoplável ao openoffice.org cogroo 2.0.

2006.

GNU.

Gnu emacs.

Disponível em: `http://www.gnu.org/software/emacs/.`, Acesso em: 2010.

Robert Krawitz, Bil Lewis, Dan LaLiberte, Richard M. Stallman, and Chris Welty.

Gnu emacs lisp reference manual.

Disponível em: `http://www.gnu.org/software/emacs/manual/html_node/elisp/index.html.`, Acesso em: 2010.

Christopher D. Manning and Hinrich Schutze.

Foundations of Statistical Natural Language Processing.