CoGrOO – Um Corretor Gramatical acoplável ao *OpenOffice*

Jorge Kinoshita[1], Laís N. Salvador[2], Carlos E. D. Menezes[3]

[1]Departamento de Computação e Sistemas Digitais – Escola Politécnica da Universidade de São Paulo (USP)

[2]Nuperc – Universidade de Salvador (Unifacs)

[3]Centro de Ciências Exatas e Tecnológicas – Universidade Cruzeiro do Sul

jorge.kinoshita@poli.usp.br, lais@unifacs.br, cedmenezes@gmail.com

Abstract. *This paper describes a Portuguese language grammar checker project, called CoGrOO - Corretor Gramatical para OpenOffice (Grammar Checker for OpenOffice). This project aims to check grammatical errors in Brazilian Portuguese language.*

Resumo. *Este artigo descreve a construção de um corretor gramatical para a língua portuguesa, chamado CoGrOO - Corretor Gramatical para OpenOffice. O objetivo da ferramenta proposta é verificar inadequações gramaticais comuns na língua portuguesa falada no Brasil.*

1. Introdução

O pacote OpenOffice tem sido adotado em grande escala tanto por usuários pessoais quanto por usuários corporativos. Seu sucesso deve-se ao fato de ele ser uma suíte completa de escritório, ter funcionalidades e interface gráfica muito similares às de competidores proprietários, rodar em muitas plataformas e ser software livre, tendo, portanto, código-fonte aberto e um custo de licenciamento igual a zero. No entanto, este pacote possuía uma deficiência que era a falta de um corretor gramatical, suprida pela construção da primeira versão do CoGrOO [KINOSHITA 2006]. A versão 1.0 do CoGrOO, financiada pela FINEP, vem sendo amplamente utilizada principalmente por organizações que adotam o pacote OpenOffice, como, por exemplo, a Usina Hidrelétrica de Itaipu. O objetivo deste trabalho é apresentar a abordagem adotada na versão 2.0 (CoGrOO-2) que representou muitos avanços tanto na sua estrutura como também na flexibilidade em portar o projeto para outras línguas [SILVA 2006]. Seguem uma descrição da arquitetura do CoGrOO-2 e as conclusões relacionadas.

2. Arquitetura

O sistema CoGrOO é composto por módulos responsáveis pela análise da sentença e por

módulos de detecção de erros (inadequações) gramaticais. Os módulos de análise são: Separador de Sentenças, Separador de tokens, Detector de Nomes Próprios, Etiquetador Morfológico, Agrupador e Analisador Sintático Simples.

O CoGrOO está na segunda versão. Essa versão fez uso intenso de um software aberto: o OpenNLP [OPENNLP 2007]. O OpenNLP é um conjunto de módulos como etiquetador e agrupador que foram implementados tendo como base o algoritmo de entropia máxima, um dos algoritmos usados em Aprendizado Computacional. Para se usar esse algoritmo é necessário que o módulo seja treinado com base em um conjunto de textos previamente anotados. Um *corpus* em português brasileiro, chamado CETENFOLHA [LINGUATECA 2006], que contém anotações morfossintáticas, foi usado como base para a criação da maioria dos módulos do CoGrOO. Seu conteúdo é formado por ensaios jornalísticos, geralmente escritos em terceira pessoa. Segue uma descrição sucinta dos principais módulos do sistema.

2.1. Etiquetador

O papel do etiquetador é associar uma etiqueta morfológica para cada palavra da sentença. Exemplo: em “Maria casa com João”, o etiquetador atribui “verbo” à palavra "casa". Em geral, estes etiquetadores utilizam um *corpus* manualmente etiquetado para extrair as principais regras e generalizá-las para outros textos não etiquetados, conseguindo altas taxas de acerto [BRILL 92]. No CoGrOO-2, o etiquetador foi uma adaptação do existente no pacote OpenNLP, baseado em [RATNAPARKHI 98], para a língua inglesa. Foram feitas alterações em suas *features* para adaptá-lo ao português e obteve-se um desempenho da ordem de 95%.

2.2. Detector de sintagmas (Agrupador)

O objetivo deste módulo é localizar pequenos sintagmas nominais ou verbais, também chamados *noun phrase* (NP) e *verbal phrase* (VP) no texto analisado. Um sintagma é uma unidade da análise sintática composta por um núcleo e outros termos associados, formando uma locução que faz parte da estrutura da sentença. Os sintagmas se classificam de acordo com os seus elementos nucleares. O sintagma nominal (NP), quando o núcleo do sintagma é um nome; e o sintagma verbal (VP), quando o núcleo do sintagma é um verbo. No CoGrOO-2, o agrupador (chunker) do OpenNLP foi adaptado da língua inglesa para a portuguesa. Para que esta nova versão do agrupador funcionasse, foi criado um novo *corpus* especialmente para o treinamento do agrupador, pois as *tags* originais do corpus não são adequadas para o treinamento do OpenNLP. Cada palavra neste novo *corpus* foi etiquetada como: B-NP (Begin-NP), palavra que começa um sintagma nominal; I-NP (Inside-NP), palavra que está no meio de um sintagma nominal; B-VP ou I-VP [RAMSHAW 1995].

2.3. Analisador Sintático Simples

O objetivo deste módulo é levantar relações gramaticais. Na fase atual do projeto está sendo implementada a detecção da relação sujeito-verbo. Para se detectar o sujeito, observa-se um padrão, ou seja, uma seqüência de etiquetas e sintagmas que associa um NP como sujeito de um VP. Um exemplo de padrão sujeito-verbo é: (!, NP, VP) onde "!" denota "começo de sentença" associando NP como sujeito de VP. Os padrões de relações sujeito-verbo foram extraídos automaticamente do *corpus* CETENFOLHA que contém a etiqueta SUBJ que associa um NP a um VP. Por exemplo, caso a sentença "o garoto subiu no muro" estivesse no *corpus*, seria observado a seguinte seqüência (!, NP(o garoto), VP(subiu), PREP(em), NP(o muro)), e através da etiqueta SUBJ, seria extraído o padrão (!, NP, VP). Porém, existem casos no *corpus* onde a seqüencia (!, NP, VP) é observada mas NP não é sujeito de VP. No *corpus* observa-se que para o padrão (!, NP, VP), NP é sujeito de VP em 85% dos casos. Devido a esta alta freqüência, (!, NP, VP) é usado como um padrão sujeito-verbo. Os padrões encontrados no *corpus* servem de base para uma máquina de estados que é usada pelo corretor gramatical na detecção de erros de concordância entre sujeito e verbo. Não encontramos na literatura, algo semelhante à implementação que fizemos desse módulo e assim encaramos essa abordagem como uma contribuição. Porém, pretendemos verificar melhor a possibilidade do uso do OpenNLP para esse módulo.

2.4. Analisador de Desvios Gramaticais

O objetivo deste modulo é localizar possíveis erros gramaticais na sentença; para isso, são usadas regras que detectam esses erros. Este módulo aplica as regras - escritas manualmente - relativas a vários tipos de inadequações, sobre as sentenças analisadas. A entrada é uma expressão regular que denota uma seqüência de *tokens*, etiquetas, sintagmas e agrupamentos sujeito-verbo. A saída é uma estrutura de dados contendo a mensagem de erro, o trecho da sentença com problema e sugestões de como contornar o desvio gramatical. Este módulo é acionado em três momentos distintos durante o processamento da sentença: após a etiquetação morfológica, após o agrupador e após a análise sintática simples. Um exemplo de uma expressão regular que capta um erro de crase é aquela que detecta uma crase ("à") antes de um verbo. Essa regra pode ser aplicada logo após a etiquetação morfológica.

3. Conclusões

O projeto CoGrOO usa uma abordagem híbrida: estatística (aprendizado automático, com treinamento a partir de *corpus* anotado, usando os princípios da Máxima Entropia) e baseada em regras (processamento simbólico). Um dos diferenciais do trabalho apresentado é a abordagem usada na detecção de relações gramaticais, ou seja, a busca por padrões sintáticos no *corpus* anotado. Não foi encontrada abordagem similar na

literatura quanto à criação de regras para a detecção sujeito-verbo.

Um trabalho futuro é portar a segunda versão do CoGrOO para outras línguas (inglês e espanhol serão as escolhas iniciais). Como a arquitetura do *OpenNLP* é desvinculada de idiomas, o trabalho resume-se a especializar os algoritmos de aprendizado de máquina para levar em consideração particularidades dos dados de treinamento e do novo idioma, elaborar dados de treinamento e submetê-los aos algoritmos estatísticos.

Referências

BRILL, E. **A Simple Rule-Based Part Of Speech Tagger**. Proceedings of ANLP-92, 3rd Conference of Applied Natural Language Processing, Trento, Italy, 1992.

KINOSHITA, J.; SALVADOR, L.N.; MENEZES, C.E.D. **CoGrOO: a Brazilian-Portuguese Grammar Checker based on the CETENFOLHA Corpus**. "Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006", Gênova, Itália, 2006. (http://cogroo.incubadora.fapesp.br/portal/down/Doc/LREC_2006.pdf)

LINGUATECA. **CETENFolha**. *Corpus* em português, analisado morfo-sintaticamente. Disponível em: http://www.linguateca.pt/CETENFolha. Último acesso em: 02 dez. 2006.

NUNES, M.G.V.; OLIVEIRA Jr., O.N. **O processo de desenvolvimento do Revisor Gramatical ReGra**, Anais do XXVII SEMISH (XX Congresso Nacional da Sociedade Brasileira de Computação), Volume 1, p.6 (resumo). Artigo Completo na Versão em CD-Rom. PUC-PR, Curitiba, Brazil, 2000.

OPENNLP. *Framework open-source* para desenvolvimento de aplicações de processamento de linguagem natural. Disponível em: http://opennlp.sourceforge.net. Último acesso em: 05 de março de 2007.

RAMSHAW, L.A.; MARCUS, M.P. **Text chunking using transformation-based learning**. In Proceedings of the Third ACL Workshop on Very Large Corpora. Association for Computational Linguistics, 1995.

RATNAPARKHI, A. **Maximum Entropy Models for Natural Language Ambiguity Resolution**. Ph.D. Dissertation. University of Pennsylvania, Julho de 1998. (ftp://ftp.cis.upenn.edu/pub/ircs/tr/98-15/98-15.pdf)

SILVA, W.D.C.M.; SUZUMURA, M.; GUSUKUMA, F.W.; PIRES, D.A.M. **Corretor Gramatical Acoplável ao OpenOffice.org - CoGrOO 2.0**. Monografia de Conclusão do Curso de Engenharia Elétrica e Engenharia de Computação, Escola Politécnica, Universidade de São Paulo, Brazil, 2006.