

Gramática Adaptativa para Análise Sintática da Língua Portuguesa

J. M. N. dos Santos

Resumo—Este artigo apresenta uma proposta de geração de gramática da língua Portuguesa através de formalismo adaptativo. O dispositivo adaptativo que será apresentado neste artigo tem dois módulos principais. O primeiro módulo do dispositivo é um reconhecedor sintático, que indica se uma frase fornecida foi ou não reconhecida como uma frase válida pela Gramática representada no momento. O segundo é o módulo de aprendizagem onde novas regras são aprendidas pelo dispositivo, abrangendo a sua capacidade de reconhecimento. Este trabalho está em andamento e faz parte da tese de doutorado do autor.

Palavras Chave— Sistema em Linguagem natural, Tecnologia Adaptativa, Gramática, Análise Sintática em português, Processamento em Linguagem Natural.

I. NOMENCLATURA

PLN: significa processamento em linguagem natural.

Sintagma: “é qualquer constituinte imediato da oração, podendo exercer a função de sujeito, complemento, predicativo ou adjunto adverbial” [2]. Pode ser composto de uma ou mais palavras. Como exemplo podemos citar que a expressão “o poeta” é um sintagma substantivo, enquanto que a expressão “de muito valor” é um sintagma preposicional.

II. INTRODUÇÃO

O processamento de linguagem natural é uma área de pesquisa da computação e sua característica principal é o tratamento de uma língua nativa como por exemplo o português, inglês e francês. É uma área multidisciplinar por envolver aspectos da ciência da computação, lingüística, lingüística computacional e das ciências cognitivas.

Um dos grandes desafios da área de PLN é o desenvolvimento de sistemas e ferramentas que permitam que a comunicação do usuário com os aparelhos domésticos e computadores em geral possa ser efetuada na língua nativa de cada pessoa.

Uma das primeiras tarefas para entender o significado de uma frase em uma língua natural é realizada através da análise sintática, com base em uma descrição formal de uma gramática para a língua. A análise sintática é o processo de análise de um texto para determinar sua estrutura gramatical conforme uma determinada definição formal de gramática.

Este artigo apresentará uma forma de se obter uma

gramática para a língua portuguesa, utilizando tecnologia adaptativa, possibilitando-se efetuar a análise sintática.

III. PROCESSAMENTO EM LINGUAGEM NATURAL

Os avanços das pesquisas em *PLN* podem oferecer no futuro um conjunto de benefícios às pessoas. A seguir seguem alguns destes benefícios.

Maior facilidade no uso de aparelhos eletrônicos domésticos.

Com o avanço tecnológico e a diminuição do custo do *chip*, os aparelhos domésticos estão cada vez mais sofisticados: televisão, DVD, telefone celular, câmera fotográfica digital, etc. Os recursos destes aparelhos são muitas vezes subutilizado por uma parcela da população, devido à dificuldade de interação, que é feita através de menus das opções. Uma interação com os aparelhos utilizando uma língua nativa, ao invés de teclas e menus de funções, facilitaria o uso destes aparelhos.

Maior acessibilidade às informações da rede mundial de computadores (internet)

Outro aspecto a ser considerado é o crescimento mundial do uso da rede mundial (internet), inclusive nas camadas sociais menos favorecidas. Como boa parte das informações está em inglês, este fator dificulta o acesso à muita informação para a parcela da população com menos acesso à educação em países em desenvolvimento, como o Brasil. O uso de recursos de *PLN* poderá facilitar o uso da internet, sob os seguintes aspectos:

- tradução dos textos para a língua nativa das pessoas,
- buscas mais refinadas, facilitando o acesso à informação desejada e
- interface com usuário através da língua nativa, em substituição ao teclado e mouse utilizados atualmente.

Otimização no serviço de atendimento das empresas à clientes (“Call-Center”)

Atualmente é comum as empresas oferecerem serviços de atendimento aos seus clientes por telefone. Porém com a diversidade de produtos ou serviços oferecidos, para direcionar a ligação do usuário ao departamento desejado é necessário que o cliente responda a várias perguntas de um sistema através das teclas do telefone. Várias pesquisas realizadas mostram a insatisfação destes clientes, devido à demora e dificuldade de compreensão na utilização destes serviços. A comunicação de uma forma mais natural, através

da língua nativa, facilitará a utilização destes recursos.

Maior facilidade no uso de terminais de auto atendimento da rede bancária.

O auto atendimento bancário, feito através dos caixas eletrônicos, também está presente cada vez mais no cotidiano de todos, independentemente da escolaridade e classe social. Uma parcela das pessoas enfrenta muita dificuldade na interação com estas máquinas, devido à enorme quantidade de opções de serviços que são agrupados em opções dos menus, nem sempre de fácil entendimento.

A forma mais intuitiva para se identificar qual função o usuário quer acessar é através de um diálogo em uma língua nativa.

Tradução

Através da análise sintática e semântica de uma frase, pode-se realizar a tradução para uma outra língua.

IV. DISPOSITIVO ADAPTATIVO

Um dispositivo adaptativo pode ser representado por um conjunto de regras expressas em qualquer formato (que podem ser descritas do tipo “se-então”) e um conjunto de operações associado a cada uma das regras. Isto implica em se ter uma relação entre um conjunto de regras (condições) e um conjunto de ações.

Aplicamos a tecnologia adaptativa para representar fenômenos que possam ser representados por um modelo computacional na forma de regras, conforme a definição descrita acima. Algumas vezes, após a execução do modelo computacional representando várias situações do fenômeno estudado conclui-se que os conjuntos que representam as regras e ações podem ser modificadas para uma melhor adaptação ou aperfeiçoamento do modelo computacional. Esta alteração pode ser representada nas regras iniciais do dispositivo, como regras adaptativas. As regras adaptativas, quando executadas, alteram o conjunto de regras e ações, representando a experiência e aprendizado do dispositivo.

V. GRAMÁTICA ADAPTATIVA DA LÍNGUA PORTUGUESA

Este trabalho apresenta uma arquitetura de um dispositivo adaptativo para análise sintática de frases escritas na língua portuguesa.

Para se realizar a análise sintática de uma frase, utiliza-se um conjunto de regras gramaticais que definem a língua em questão. Baseado neste conjunto de regras de derivação, usualmente utilizada para representar uma gramática, o dispositivo será capaz de verificar se uma frase está gramaticalmente correta ou não, conforme sua gramática.

Como a elaboração de uma gramática representada por regras que abranja toda a língua portuguesa é complexa, adotou-se um modelo computacional que permite que novas regras gramaticais possam ser adicionadas a um conjunto inicial, aumentando assim a capacidade de representação da gramática. Desta forma o modelo criado será amplo mas

poderá ser aprimorado por um módulo de aprendizagem, aumentando sua capacidade de representação.

Esta característica de possibilitar a inclusão de novas regras gramaticais, torna este modelo computacional um dispositivo adaptativo, conforme conceito apresentado anteriormente no item IV.

As regras iniciais da gramática foram extraídas de [2]. Podemos separar essas regras em dois grupos. O primeiro abrange as regras de primeiro nível (o mais alto da árvore de derivação) e o segundo abrangendo as demais regras, ou seja, desde as intermediárias até as últimas contendo os símbolos terminais. Os símbolos terminais serão representados pelas categorias gramaticais das palavras, como: artigo, pronome, adjetivo ou advérbio. Os símbolos não terminais são utilizados para a definição de grupos de palavras, como por exemplo Sintagma Substantivo que pode ser um substantivo próprio, um substantivo comum, precedidos ou não de artigo. Para mais detalhes, observar as regras de derivações envolvendo estes elementos nas tabelas de regras.

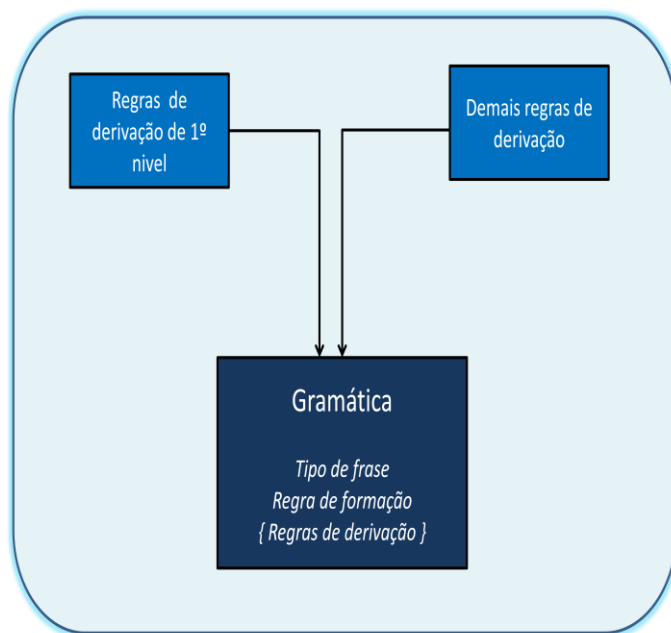


Fig 1 – Tabelas das regras da definição da Gramática.

A Fig. 1 ilustra as tabelas mencionadas na composição das regras da gramática. Baseado em [2], definiu-se o conjunto inicial da gramática, separando-as nos dois grupos de regras mencionados. Estas regras iniciais reconhecem vários padrões de frases, como por exemplo: oração de predicado nominal, oração de verbo transitivo indireto, oração de período composto por subordinação adverbial, etc

A seguir são apresentadas as duas tabelas de regras (a de 1º nível e a das demais regras) e uma tabela contendo as descrições dos símbolos (terminais e não terminais).

Regras de derivação – 1º nível (tabela 1)									
1.	SS	+	Vtd	+	SS				
2.	SS	+	Vlig	+	SS				
3.	SS	+	Vlig	+	SP				
4.	SS	+	Vti	+	SP				
5.	SS	+	Vtdi	+	SS	+	SP	+	Sadv
6.	SS	+	Vi	+	Conj	+	SS	+	Vi
7.	SS	+	Vtdi	+	SS	+	SP		
8.	SS	+	neg	+	Vti	+	SP	+	SAdv
9.	SS	+	Vtd	+	SS				
10.	Vlig	+	Sadj	+	SS				
11.	SS	+	Vti	+	SP	+	SP		
12.	SS	+	Vi	+	Sadv				
13.	Sadv	+	SS	+	Vti	+	SP		
14.	SS	+	Vlig	+	Sadj				

Demais regras de derivação (tabela 2)	
1	$SS \rightarrow (PrA + Sc Sp) \text{ ou } (Sp) \text{ ou } (Sc)$
2	$SP \rightarrow Prep + SS$
3	$Sadv \rightarrow Advt$
4	$SS \rightarrow Sb + (Artdef) + Vlig + Sadj$
5	$Sadj \rightarrow Adj$
6	$SS \rightarrow Sb + SS1 (1) + SS2 (1) + Vtd$
7	$SS \rightarrow PrA + Sc + OSAR$
8	$OSAR \rightarrow PrRel + Vi$
9	$OSAR \rightarrow PrRel + PrPes + Vtd$
10	$SS \rightarrow Card + Sc + Sadj$
11	$SS \rightarrow PrPes \text{ ou } \text{elipse}$
12	$OSAR \rightarrow Sadv + (PrPes) + Vi$
13	$OSAR \rightarrow Prep + PrRel + PrPes + PrPes + Vti$
14	$OSAR \rightarrow Rel + Sc + ArtDef + Sc + Vtd$
15	$SS \rightarrow Sadv + GV$
16	$Sadv \rightarrow PrRel$
17	$GV \rightarrow Vtd + SS + PrA + Sc$
18	$SS \rightarrow ArtDef + Sc$
19	$Sadv \rightarrow Prep + Sb + SS + Aux + Vi$
20	$PrA \rightarrow Artind$
21	$PrA \rightarrow Artdef$
22	$PrA \rightarrow dem$

Símbolos – terminais e não terminais (tabela 3)	
SS	Sintagma substantivo
Conj	Conjunção
Sadj	Sintagma (locução) adjetivo
Sadv	Sintagma adverbial
SP	Sintagma preposicional
Vlig	Verbo de ligação
Vtd	Verbo transitivo direto
Vtdi	Verbo transitivo direto e indireto
Vti	Verbo intransitivo
Neg	Negação
GV	Grupo verbal
OSAR	Oração subj. adjetiva restritiva
PrA	Pronome adjetivo
Advt	Advérbio de tempo
Artdef	Artigo definido
Artind	Artigo indefinido
Card	Numeral cardinal
Dem	Demonstrativo
Prep	Preposição
Prind	Pronome indefinido
PrPess	Pronome pessoal
PrRel	Pronome relativo
Sb	Subordinador (conj. subordinativa)
Sc	Substantivo comum
Sp	Substantivo próprio

VI. ARQUITETURA

O dispositivo adaptativo é composto de dois módulos: o de reconhecimento e o de aprendizagem.

Módulo de reconhecimento

Este modulo lê uma frase e verifica se ela é reconhecida por alguma regra da gramática. Caso ela seja reconhecida por alguma regra, o dispositivo emite uma mensagem indicando que a frase foi reconhecida e gera em um arquivo XML a árvore sintática correspondente.

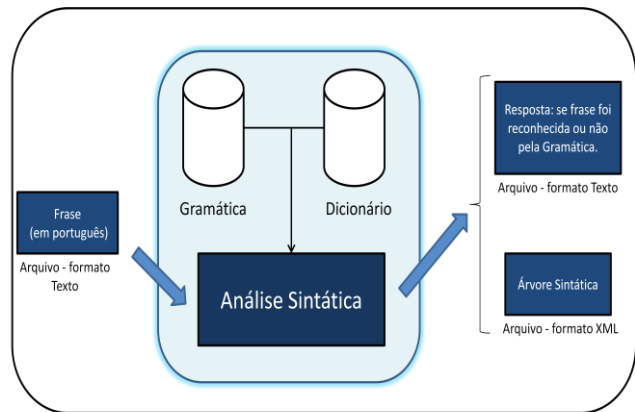


Fig 2 – Análise Sintática de uma frase

Modulo de aprendizagem

Este módulo contém a parte adaptativa da gramática. Seu funcionamento ocorre através da leitura de um texto (contendo uma ou mais frases). Este módulo poderá operar de duas maneiras: com ou sem supervisão.

Quando operar com supervisão, os padrões de frases identificados nos exemplos que não forem reconhecidos pela gramática serão submetidos a um especialista para aprovação ou não dessa nova regra.

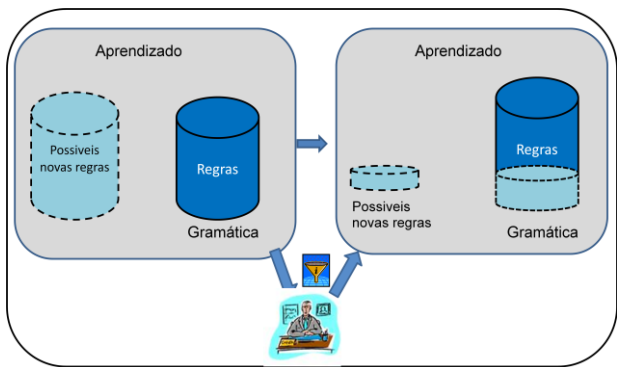


Fig 3 - Aprendizado – Com supervisão

Quando operar sem supervisão, o módulo de aprendizagem irá acatar como novas regras os padrões não reconhecidos pela gramática e que foram identificados através dos exemplos.

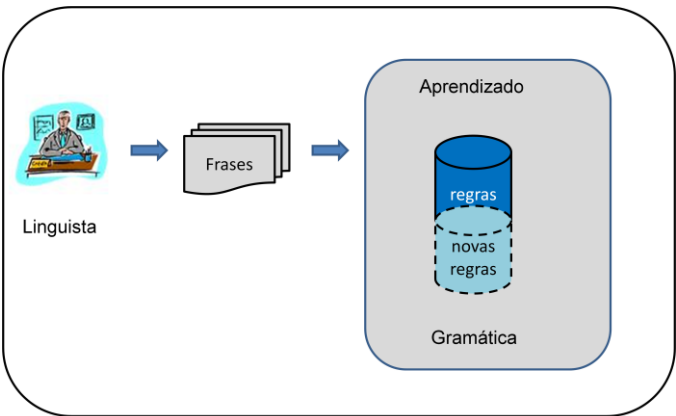


Fig. 4 - Aprendizado – Sem supervisão

Abaixo temos alguns exemplos de ilustração. O processo de identificação é o mesmo para todas as situações mencionadas. A diferença é apenas que o módulo de aprendizagem inclui uma nova regra de formação quando um novo tipo de frase é encontrado operando no modo sem supervisão, ou quando for confirmado pelo especialista quando estiver operando no modo com supervisão.

Oração Transitiva

Representação de uma oração transitiva direta através de sequência de não-terminais: SS1 - Vtd - SS2 (símbolos da tabela SNT-1). Regra 1 da tabela 1.

Regras de derivação dos termos da oração:

- SS1 → Regras1, 20 e
- SS2 → Regras 1 e 21 da tabela 2.

Exemplo: nenhum aluno conhece o livro
nenhum aluno: SS1,
conhece: Vtd,
o livro: SS2.

Aplicando-se a regra 1 e 20 em SS1 e regras 1 e 21 em SS2, temos:

- nenhum: Artind
- aluno: Sc
- conhece: Vtd
- o : Artdef
- livro: Sc

Oração de verbo transitivo direto e indireto

Representação de uma oração transitiva direta através de sequência de não-terminais: SS1 – Vtdi - SS2 – SP – Sadv. Regra 5 da tabela 1.

Regras de derivação (tabela 2) dos termos da oração:

- SS1 → Regra 1,
- SS2 → Regras 1 e 21,
- SP → Regras 2, 1,21 e
- Sadv → Regra 3.

Exemplo: Pedro colocou o livro na biblioteca hoje
Pedro: SS1,

colocou: *Vtdi*,
o livro: *SS2*,
em a (na) biblioteca: *SP*.

Aplicando-se as regras, temos:

Pedro: *Sp*,
colocou: *Vtdi*,
o: *Artdef*,
livro: *Sc*,
em: *Prep*,
a : *Artdef*,
biblioteca: *Sc*,
hoje: *Adv*.

- [4] J. J. Neto, “Contribuições à metodologia de construção de compiladores”, Tese de Livre Docência, USP, São Paulo, 1993.
- [5] J. J. Neto e M. K. Iwai, “Adaptive Automata for Syntax Learning”, CLEI 98 - XXIV Conferencia Latinoamericana de Informatica, MEMORIAS. pp. 135-149, Quito, Equador, 1998

José Maria Novaes dos Santos. É bacharel em Matemática Aplicada pela Universidade de São Paulo. É mestre em Engenharia da Computação pela Escola Politécnica da Universidade de São Paulo. Atualmente é Doutorando do Departamento de Engenharia de Computação e Sistemas Digitais da Escola Politécnica da Universidade de São Paulo. Suas áreas de interesse e pesquisa abrangem: Teoria da Computação, Tecnologia Adaptativa e Processamento em Linguagem Natural.

VII. TRABALHOS FUTUROS

Como futuro trabalho podemos citar a elaboração de um corretor gramatical para a língua portuguesa, baseado na gramática representada pelo dispositivo descrito neste artigo.

Outro trabalho interessante, é adaptar esta gramática e o corretor gramatical para outras linguas como o espanhol e o inglês.

Outro trabalho futuro pode ser realizado através do aprimoramento deste modelo para reconhecimento de padrão (estilo) de escrita. Pode-se fazer inferência em um texto longo, com várias frases de um autor, para tentar se reconhecer um padrão, conforme o tipo de frases utilizadas. Com base nestas informações, pode-se tentar verificar se um outro texto pertence ao mesmo autor ou não.

VIII. CONCLUSÃO

Foi apresentado um dispositivo adaptativo para representação de gramática para a língua portuguesa. A implementação deste modelo está em desenvolvimento, fazendo parte da tese de doutorado do autor.

Com base nos tipos de frases mapeados, através de [2], temos uma gramática com boa abrangência de representação da língua portuguesa. A sua característica adaptativa permitirá que ele aumente sua capacidade de representação. Porém, deve-se observar que este aumento de regras poderá diminuir significamente a performance de reconhecimento. Então, o ideal é escolher uma gramática representativa para a aplicação em questão, onde se tenha uma performance aceitável.

REFERÊNCIAS

- [1] J. J. Neto, Adaptive Rule-Driven Devices - General Formulation and Case Study. Lecture Notes in Computer Science. Watson, B.W. and Wood, D. (Eds.): Implementation and Application of Automata 6th International Conference, CIAA 2001, Vol. 2494, Pretoria, South Africa, July 23-25, Springer-Verlag, 2001, pp. 234-250.
- [2] C. P. Luft, Moderna Gramática Brasileira, São Paulo, Brasil, Editora Globo, 2ª. edição revisada e atualizada.
- [3] J. J. Neto, “Adaptive Rule-Driven Devices - General Formulation and Case Study. Lecture Notes in Computer Science”, Watson, B.W. and Wood, D. (Eds.): Implementation and Application of Automata 6th International Conference, CIAA 2001, Vol.2494, Pretoria, South Africa, July 23-25, Springer-Verlag, 2001, pp. 234-250.