Pontifícia Universidade Católica do
Rio Grande do Sul
Faculdade de Informática
Curso de Bacharelado em
Ciência da Computação

Parsing Probabilístico para a Língua Portuguesa

Trabalho de Conclusão I

Autores:

Rodrigo R M Kochenburger

Marlon Gomes Lopes

Orientador:

Carlos Augusto Prolo

Porto Alegre, Junho de 2009

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

CFG *Context Free Grammar* - Gramatica Livre de Contexto

PCFG *Probabilistic Context Free Grammar* - Gramatica Livre de Contexto Probabilística

Resumo

Abstract

1 Introdução

A linguagem é um dos aspectos mais fundamentais do comportamento humano e um componente crucial de nossas vidas. A linguagem em sua modalidade escrita serve como um registro a longo prazo do conhecimento que é passado de geração em geração. Na forma falada, ela serve diariamente como meio primário de coordenação do nosso comportamento e interatividade entre as pessoas.

A língua - ou linguagem - é estudada em diferentes meios ou áreas acadêmicas. Cada disciplina define os seus próprios problemas e possui seus próprios métodos para resolvê-los. A linguística, por exemplo, estuda a estrutura da própria linguagem, considerando perguntas como: a) porque certas combinações de palavras compõem uma sentença e outras não; ou b) porque algumas sentenças possuem significado e outras não. A psicolinguística estuda o processo de produção e compreensão da língua pelos seres humanos, levando em consideração perguntas como: a) como as pessoas escolhem, ou identificam, a estrutura apropriada das frases; ou b) como elas decidem os significados para cada palavra.

O objetivo da linguística computacional é utilizar os conhecimentos desenvolvidos nas áreas citadas - e outras relacionadas - para desenvolver teorias e aplicações utilizando as noções de algoritmos e estrutura de dados para processar e entender sintática e semanticamente a língua, escrita ou falada.

Nas ultimas duas décadas, o desenvolvimento de métodos estatísticos para o processamento de linguagem natural (PLN) vem sendo impulsionado pela evolução no poder de processamento dos computadores [MAN99, JUR00]. Esses métodos utilizam grande quantidade de dados, em conjunto com cálculos estatísticos, para tentar "entender"corretamente a estrutura e o significado da linguagem. Fundamental nesse processo foi o surgimento de *corpora* anotados (*treebanks*) [MAR93, MAR94, ABE03, SAR04].

Este trabalho pretende estudar o processo estatístico de *parsing* baseado em *corpus* aplicado à língua portuguesa. Iremos focar o processo de *parsing*, mais especificamente as

abordagens estatísticas baseadas em *corpus*, para solucionar esse problema e desenvolver, especificamente, o processo de *parsing* probabilístico aplicado à língua portuguesa. Iremos utilizar como base os estudos e o *parser* desenvolvido por Michael Collins [COL99, COL97] na versão posterior reimplementada por Dan Bikel [BIK04].

2 Motivação

A motivação pode ser dividida em dois aspectos principais: científico e tecnológico.

A motivação científica é a obtenção do conhecimento e o melhor entendimento a respeito de como as linguagens funcionam. Nenhuma das disciplinas tradicionais possui, isoladamente, ferramentas necessárias para decifrar completamente a produção e a compreensão linguísticas que os seres humanos possuem. Porém, é possível utilizar programas de computadores para implementar essa complexa teoria, de modo que seja possível testá-la, verificá-la e incrementalmente melhorá-la. Ao aprofundar o estudo deste processo, podemos desenvolver um entendimento a respeito de como os seres humanos processam as línguas.

Quanto à natureza tecnológica, a maior parte do conhecimento humano está armazenada de forma linguística, escrita ou falada, e computadores que conseguissem "entender"linguagem natural poderiam acessar toda essa informação. Outro aspecto relevante é a possibilidade de melhorsr a interação humano-computador, aumentando o nível de acessibilidade, o que tornaria mais simples a utilização de ferramentas computacionais por pessoas com necessidades especiais.

Atualmente, a evolução do poder computacional e a construção de grandes *treebanks* possibilitam a utilização de técnicas mais avançadas, que utilizam grande quantidade de informação e processamento para tentar resolver esses problemas. Técnicas como *parsing* probabilístico, que utiliza técnicas de aprendizado e cálculos estatísticos baseados em um banco de dados manualmente anotado - conhecido como *corpus* ou *treebank* - para identificar as informações sintáticas corretas, têm se mostrado bastante eficazes, na comparação com outros métodos, e suficientemente satisfatórias, por conta dessa evolução computacional.

Muitas pesquisas e trabalhos vêm sendo realizados, com foco em vários idiomas, notadamente para o inglês [PRO03, CHA97, COL97], entretanto verifica-se uma carência de pesquisas, ferramentas, recursos linguísticos e humanos para tratar computacionalmente

a língua portuguesa. Existem alguns trabalhos [WIN06, BIC00, BON03] mas é fato reconhecido pelos pesquisadores que ainda não se atingiu um resultado de nível desejável.

Michael Collins, no final da década de 1990, desenvolveu três modelos de *parsing* probabilístico, sendo os últimos extensões aos anteriores. Estes modelos e o seu *parser* são até hoje referência na área e continuam sendo utilizados. Posteriormente, em 2004, Dan Bikel reimplementou o *parser* de Collins, tornando-o mais parametrizável e extensível. Ambos os *parsers* foram amplamente testados para a língua inglesa e, na atualidade, as tentativas de se construir um *parser* probabilístico para a língua portuguesa não tem sido, até o momento, satisfatório.

Finalmente, outro motivador crucial para este trabalho de conclusão é o trabalho em desenvolvimento pelo projeto Semantic Share da Universidade de Lisboa, quanto a ao o desenvolvimento de um *corpus* para a língua portuguesa melhor anotado. Este *corpus* contem dados linguísticos de fala ou escrita, servindo de base de uso para o *parser* estatístico utilizado.

3 Objetivos

Este trabalho de conclusão terá como objetivos principais estudar e compreender as técnicas estatísticas de processamento de linguagem natural implementadas por Michael Collins, utilizar e analisar o *parser* reimplementado por Dan Bikel, que tornou possível a parametrização e extensão das suas bibliotecas para possíveis adaptações do código. Os objetivos específicos são os seguintes:

- Estudar as técnicas envolvidas no desenvolvimento de um *parser*.
- Estudar os modelos de *parsing* de Collins.
- Dominar o uso da ferramenta de *parsing* de Bikel e, se necessário, a original de Collins.
- Fazer um estudo detalhado dos parâmetros de implementação de Bikel, pois a eficácia dos algoritmos depende fundamentalmente dos ajustes desses parâmetros.
- Utilizar a ferramenta de Bikel para construção de um *parser* para a língua portuguesa. O treinamento do *parser* será feito utilizando-se o *corpus* anotado em desenvolvimento pelo projeto Semantic Share, da universidade de Lisboa [BRA09].
- Possivelmente, desenvolver módulos de processamento e as alterações de códigos que se mostrarem necessárias.

4 Metodologia

Este trabalho possui um forte componente experimental. Assim, em termos metodológicos, a cada experiência realizada, os resultados obtidos devem ser analisados quantitativa e qualitativamente, para orientar as correções nos parâmetros do *parser* ou indicar a necessidade de alterações como: a) de pré-processamento ou pós-processamento dos casos; ou b) no código do *parser*. Nesse sentido, a avaliação quantitativa é um componente importante e será feita de forma rigorosa. Pretende-se utilizar as metodologias tradicionais de precision/recall [BLA93] e possivelmente outras a definir.

Será usado um sistema de controle de versão que permite que se trabalhe com diversas versões dos arquivos de trabalho e versões do software durante nossos testes e implementações.

5 Referencial Teórico

O presente trabalho de conclusão situa-se na área de processamento de linguagem natural, conforme ilustrado na Figura 1.

No texto que segue, abordaremos alguns assuntos que fundamentam nosso trabalho.

5.1 Análise de Sentença

O processo de analise de sentença em linguagem natural é geralmente apresentado na literatura subdividido em vários níveis:

- Análise morfológica
- Análise sintática
- Análise semântica
- Análise pragmática

Este trabalho foca os dois primeiros níveis de análise acima citados. A seguir faremos uma breve introdução aos níveis acima citados.

Uma abordagem mais completa de todos os níveis pode ser vista em [ALL95, LIM01] entre outros.

5.1.1 Análise morfológica ou *Part-of-Speech Tagging*

O analisador morfológico identifica palavras ou expressões isoladas em uma sentença, sendo este processo auxiliado por delimitadores (pontuação e espaços em branco). As palavras identificadas são classificadas de acordo com seu tipo de uso ou, em linguagem natural, categoria gramatical.

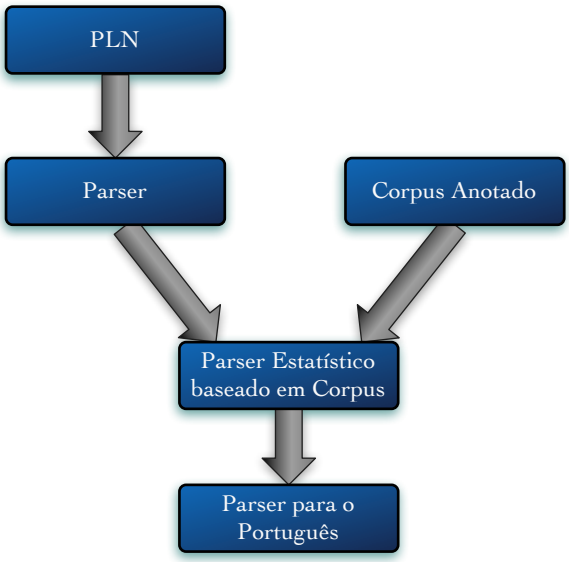

Figura 1: Estágios do processamento de linguagem natural

Neste contexto, uma instância de uma palavra em uma sentença gramaticalmente válida pode ser substituída por outra do mesmo tipo (exemplo: substantivos, pronomes, verbos, etc.), configurando uma sentença ainda válida. Para um mesmo tipo de palavra, existem grupos de regras que caracterizam o comportamento de um subconjunto de vocábulos da linguagem (exemplo: formação do plural de substantivos terminados em "ão", flexões dos verbos regulares terminados em "ar", etc.). Assim, a morfologia trata as palavras quanto à sua estrutura, forma, flexão e classificação, no que se refere a cada um dos tipos de palavras.

Esta fase é frequentemente chamada de *part-of-speech tagging*, pois seu principal resultado é a determinação da categoria sintática das palavras individuais como ocorrem na sentença, também conhecida como *part-of-speech* (POS). Entre essas categorias estão tipicamente as de nome (ou substantivo), verbo, preposição, etc. Outras características importantes podem ser obtidas nesta fase, como gênero (masculino ou feminino), número (singular ou plural), etc. Estas características secundárias, chamadas *features* ou traços, de certa forma estendem a POS. Cada POS tem um conjunto diferenciado de *features* apropriado que depende da aplicação do analisador e das concepções teóricas de quem a define. Na medida em que uma palavra é caracterizada pela sua categoria principal mais traços secundários, não é surpresa que haja uma razoável variabilidade na separação entre que características já devem estar embutidas na POS, e quais devem ser relegadas a *features*. Por exemplo, algumas propostas podem selecionar como POS nome e como *feature* número (singular ou plural). Outras podem atribuir POS *tag* (marcações de POS)

separados para nome-singular e nome-plural.

Os algoritmos para etiquetagem fundamentam-se em dois modelos mais conhecidos: os baseados em regras e os estocásticos. Os algoritmos baseados em regras, como o nome diz, fazem uso de bases de regras para identificar a categoria de um certo item lexical. Neste caso, novas regras vão sendo integradas à base à medida que novas situações de uso do item vão sendo encontradas. Os algoritmos baseados em métodos estocásticos costumam resolver as ambiguidades através de um *corpus* de treino, marcado corretamente (muitas vezes através de esforço manual), calculando a probabilidade que uma certa palavra ou item lexical terá de receber uma certa etiqueta em certo contexto. O etiquetador de Eric Brill [BRI95], bastante conhecido na literatura, faz uso de uma combinação desses modelos.

A escolha de um bom *tagset* é fundamental para o sucesso de um *parser*, embora não seja absolutamente claro como fazer este julgamento. Existem vários livros inteiros dedicados a este assunto. Em linhas gerais, um bom *tagset* para um *parser* é aquele que possui uma boa caraterística de equivalência distribucional"em termos sintáticos; isto é, palavras que ocorrem tipicamente nas mesmas posições nas sentenças têm mesmo POS, enquanto que as que têm características de distribuição diferentes na mesma sentença têm POS diferente.

O *corpus* que será usado no trabalho tem seu *tagset* definido em [BRA08].

5.1.2 Análise sintática ou *Parsing*

Através da gramática da linguagem a ser analisada e das informações do analisador morfológico, o analisador sintático procura construir árvores de derivação para cada sentença, mostrando como as palavras estão relacionadas entre si.

Durante a construção da árvore de derivação, é verificada a adequação das seqüências de palavras às regras de construção impostas pela linguagem, no processo de composição das sentenças. Dentre estas regras, pode-se citar a concordância e a regência nominal e/ou verbal, bem como o posicionamento de termos na frase.

A tarefa de um *parser* para a linguagem natural é construir a estrutura sintática da sentença, dividindo-a em subconstituintes de uma forma que reflita, segundo alguma teoria da linguagem, a estrutura composicional de análise da sentença. Esta estrutura é geralmente dada como uma árvore de constituintes, em que os nodos folhas são as POS, com as respectivas palavras, e os nodos internos os conhecidos como sintagmas ou

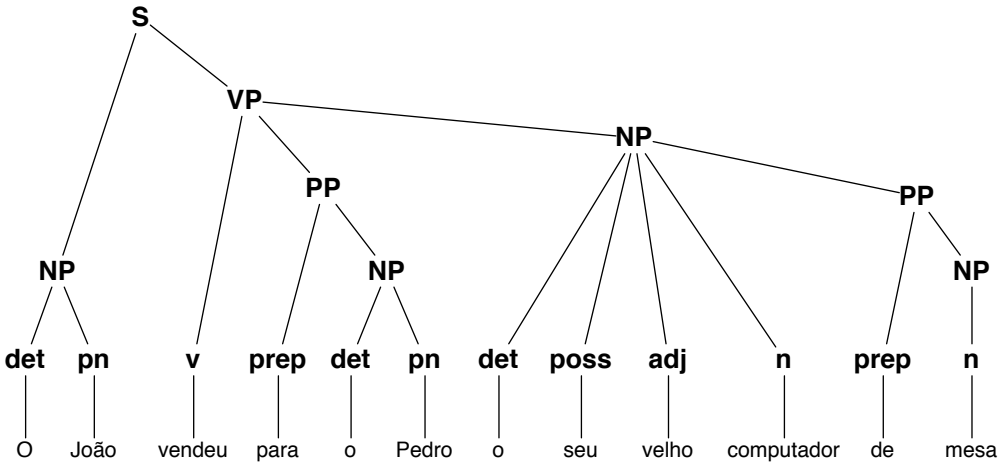

Figura 2: Árvore gramatical da frase *O João vendeu para Maria o seu velho computador de mesa*

categorias sintáticas de mais alto nível.

Por exemplo, a frase "O João vendeu para Maria o seu velho computador de mesa", seria anotada gramaticalmente da seguinte forma:

```
(S
   (NP (DET O) (PN João))
   (VP
      (V vendeu)
      (PP (PREP para) (NP (DET o) (PN Pedro)))
      (NP
          (DET o)
          (POSS seu)
          (SDJ velho)
          (N computador)
          (PP (PREP de) (NP (N mesa))))))
```

A Figura 2 ilustra a mesma árvore em formato gráfico.

5.2 *Corpus* Anotado

Segundo [SAR04], *corpus* é um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizado segundo alguns critérios, suficientemente

extenso em amplitude e profundidade, de maneira que seja representativo da totalidade do uso linguístico ou de algum de seus âmbitos, disposto de tal modo que possa ser processado por computador, com a finalidade de propiciar vários, e úteis, resultados para a descrição e análise.

Corpus anotados, também conhecidos como *treebanks* [ABE03], são - simplificadamente - bancos de dados de sentenças anotadas com informações sintáticas e semânticas que servem como fonte de aprendizado para os sistemas estatísticos. A qualidade do *treebank* influencia diretamente na qualidade do resultado obtido pelo *parser*. De acordo com [ABE03], a criação do primeiro *corpus* anotado data de aproximadamente 30 anos, e foi desenvolvido inicialmente para o inglês. O objetivo era prover um esquema de anotação mais completo possível, para ser utilizado por esses métodos empíricos, isto tendo em vista processamento dos corpora linguisticos. Para outras finalidades há coisas bem mais antigas

5.2.1 Extensão do *Corpus*

A extensão e diversidade dos *corpora* são definitivas na qualidade do aprendizado dos *parsers* estatísticos. Segundo [SAR04], pode-se definir três abordagens para a constituição de um *corpus*.

1. Impressionista: baseia-se em constatações derivadas da prática da criação e da exploração de *corpora*, em geral feitas por autoridades da área. Por exemplo, Aston [AST97] menciona patamares que caracterizariam um *corpus* pequeno (20 a 200 mil palavras) e um grande (100 milhões ou mais).

   Leech [LEE91] fala de 1 milhão de palavras com uma taxa usual (*going rate*), sugerindo o patamar mínimo. Outros são mais vagos, como Sinclair [SIN97], que postula que o *corpus* deva ser tão grande quanto a tecnologia permitir para a época, deixando subentender que a extensão de um *corpus* deva variar de acordo com o padrão corrente nos grandes centros de pesquisa, que possuem equipamentos de última geração.

2. Histórica: fundamenta-se na monitoração dos *corpora* eletivamente usados pela comunidade. Por exemplo, Berber Sardinha [SAR04] sugere uma classificação baseada na observação dos *corpora* utilizados, segundo 4 anos de conferências de *corpus*. Tabela de tamanho de *corporas*:

Tabela 2: *Tamanho de um Corpus.*

Tamanho	**Classificação**
Menos de 80 mil	Pequeno
80 mil a 250 mil	Pequeno-médio
250 mil a 1 milhão	Médio
1 milhão a 10 milhões	Médio-grande
10 milhões ou mais	Grande

3. Estatística: fundamenta-se na utilização de teorias estatísticas. Por exemplo, Biber [BIB93] emprega fórmulas matemáticas para identificar quantidades mínimas de palavras, gêneros e textos que se constituíram em uma amostra representativa. Algumas questões que norteiam essa abordagem são:

   (a) Dado um *corpus* preexistente que serve como amostra maior, qual o tamanho mínimo de uma amostra que mantem estáveis as características da amostra maior? Essa é uma perspectiva seguida por Biber [BIB90, BIB93].

   (b) Dada uma fonte externa de referência cuja dimensão é conhecida, qual o tamanho do *corpus* necessário para representar majoritariamente esta fonte? Essa vertente tem sido discutida pela comunidade de linguistas do *corpus*.

   (c) Quanto seria perdido se o *corpus* fosse de um tamanho x? Dados os recursos existentes, quais parâmetros utilizar para avalizar a decisão relativa ao tamanho de *corpus* que pode ser compilado? Uma proposta segundo essa perspectiva ainda não foi formalizada, mas está presente, por exemplo, em [GÓM97, GÓM97a], que estima matematicamente a quantidade do vocabulário presente em *corpora* de diversos tamanhos hipotéticos.

5.2.2 Corpus para processamento de linguagem natural

5.2.2.1 Corpus da Lingua Inglesa

Tabela 3: *Corpus da Lingua Inglesa.*

Corpus	**Lançamento, referência na literatura**	**Palavras**	**Composição**
Bank of English	1987***	459 milhões	inglês britânico
Longman Writen American Corpus	1997	100 milhões	inglês americano, escrito (jornais e livros)
DNC (British National Corpus)	1995	100 milhões	inglês britânico escrito e falado.
LLEC (Longman-Lancaster English Language Corpus)	1988	30 milhões	inglês de vários tipos , escrito e falado.
CHILDES (Child Language Data Exchange)	1990	20 milhões	inglês infantil, falado.
The Penn TreeBank	1989	10 milhões	inglês americano, escrito e falado.
Brown Corpus (Brown University Standart Corpus of Present-day American English)	1964	1 milhão	ingles americano, escrito

*** Data refere-se ao Birmingham Corpus, do qual o Bank of English Derivou

Existem outros corpus alem dos acima elencados, que possuem um numero menor de palavras. Três corpus da lista servem como marcos de referencia históricos:

Brown, BNC e Bank of English. O corpus Brown é um marco por razoes obvias: é o primeiro. O BNC é de destaque porque foi o primeiro a conter 100 milhões de palavras. Enquanto o Brown e o BNC são corpus de amostragem, planejados e fechados, O Bank of English é um corpus monitor, orgânico e em crescente expansão.

5.2.2.2 Penn TreeBank

O Penn TreeBank é um dos principais corpus disponíveis, contem aproximadamente 7 milhões de palavras com anotação de POS, 3 milhões de palavras of skeletally parser text, mais de 2 milhões de palavras de texto analisado para estrutura predicativa

(*predicate-argument structure*), e 1.6 milhões de palavras de transcrição de conversas. O material anotado possui diferentes origens e gêneros como manuais de computadores da IBM, anotações de enfermeiras, artigos do Wall Street Journal e transcrições de conversas telefônicas, entre outras.

A maioria das anotações do Penn Treebank consiste em anotação de POS e versões parentisadas dos textos escritos como os artigos do Wall Street Journal. Nos primeiros anos do projeto a parentisação foi utilizada usando apenas um simples "*skeletal parse*"ou parse estrutural, enquanto que mais tarde foram incluídas informações de "*richer predicate-argument bracketing schema*".

O Penn Treebank tagset, como muitos outros corpus, foi baseado no Brown Corpus. Vemos uma amostra do tagset do Penn Treebank na tabela abaixo, contendo 36 tags de POS e 9 outras tags para pontuação e 17 tags para anotação sintática. Uma detalhada descrição do tagset do Penn Treebank é encontrado no website do projeto Penn Treebank em http://www.cis.upenn.edu/ treebank/.

Tabela 4: *Tags de Part-of-Speech do Penn Treebank.*

CC	Conjunção Coordenativa	PP	Pronome Pessoal
CD	Numeral	PP$	Pronome Possessivo
DT	Determinante	RB	Advérbio
EX	Existencial lá	RBR	Advérbio, comparativo
FW	Palavra estrangeira	RBS	Advérbio, superlativo
IN	Preposição ou conjunção subordinada	RP	Partícula
JJ	Adjetivo	SYM	Símbolo
JJR	Adjetivo comparativo	TO	para
JJS	Adjetivo superlativo	UH	Interjeição
LS	Item, marcador	VB	Verbo infinitivo
MD	Modal	VBD	Verbo passado
NN	Nome, singular	VBG	Verbo gerúndio ou presente particípio
NNS	Nome, plural	VBN	Verbo passado particípio
NP	Nome próprio singular	VBP	Verbo, não-terceira pessoa singular
NPS	Nome próprio plural	VBZ	Verbo, terceira pessoa singular
PDT	Predeterminador	WDT	Determinador interrogativo
POS	Terminado possessivo	WP	Pronome interrogativo
		WP$	Pronome interrogativo possessívo
		WRB	Advervio interrogativo
#	Libra sinal	$	Dollar sinal
.	final	,	vírgula
:	dois pontos	(	parenteses abre
)	parenteses fecha	”	arpas dupla abre
’	aspas simples		

5.2.2.2.1 Tags de Part-of-Speech

5.2.2.2.2 Tags sintáticos

Tabela 5: *Tags sintáticos do Penn Treebank.*

ADJP	Sintagma Adjetivo	ADVP	Sintagma Adverbial
NP	Sintagma nominal	PP	Sintagma preposicional
S	Sintagma de clausula declarativa simples	SBAR	Sintagma subordinativo
SBARQ	Sintagma interrogativo	SINV	Sintagma declarativo com inversão de sujeito
SQ	SIntagma de questao sim/não e subconstituinte de SBARQ excluindo elemento interrogativo	VP	Sintagma verbal
WHADVP	Sintagma adverbial interrogativo	WHNP	Sintagma nominal interrogativa
WHPP	Sintagma preposicional interrogativa	X	Sintagma de constituinte desconhecido
*	Entendido como sujeito de infinitivo ou imperativo	0	Zero variante de clausula subordinada
T	Sintagma de constituinte interrogativo - trace		

5.2.2.3 Corpus da Lingua Portuguesa

Na língua portuguesa, há vários corpus eletrônicos de destaque, elecaremos alguns abaixo.

A lista apresenta um pequeno resumo dos corpus existentes para o português, selecionados por estar presente e ser fonte de pesquisa.

5.2.2.4 Floresta Sintática (Projeto Linguateca)

Um dos objetivos da Linguateca é melhorar significativamente as condições para o processamento do português, e prover recursos para pesquisa como os repositórios do Floresta Sintática , CETEMPublico e o CETEMFolha .

O CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público) é um corpus de aproximadamente 180 milhões de palavras em português de Portugal, criado pelo projeto Processamento computacional do português após a assinatura de um protocolo entre o Ministério da Ciência e Tecnologia português (MCT) e o jornal PÚBLICO.

O CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo) é um corpus de cerca de 24 milhões de palavras em português brasileiro, criado

Tabela 6: *Corpus da Lingua Portuguesa.*

Corpus	**Palavras**	**Composição**	**Localização**
Banco de Português	233 milhões	português brasileiro, escrito e falado	PUC/SP
CETEM (Corpus de extração de Textos eletrônicos MCT), publico	220 milhões	jornal português, "publico"	Projeto Linguateca
Corpus UNESP/Araraquara/ Usos do português	200 milhões	português brasileiro, escrito	UNESP / Araraquara
CRPC(COrpus de referencia do português contemporâneo)	152 milhões	português dos vários paises lusófonos, com predominância da variedade européia	CLUL - Centro de lingüística da Universidade de Lisboa.
NILC	35 milhões	português brasileiro escrito	NILC (USP, UFSCAR, UNESP Araraquara)

pelo projeto Processamento computacional do português com base nos textos do jornal Folha de S. Paulo que fazem parte do corpus NILC/São Carlos, compilado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC).

Projeto que deu origem à Linguateca, com base nos textos do jornal Folha de S. Paulo que fazem parte do corpus NILC/São Carlos, compilado pelo Núcleo Interinstitucional de Lingüística Computacional (NILC).

Conjunto de frases (corpus) analisadas (morfo)sintaticamente possuindo também indicação das funções sintáticas, a análise também explicita hierarquicamente informação relativa à estrutura de constituintes, dizemos que uma frase sintaticamente analisada se parece com uma árvore, donde um conjunto de árvores constitui uma floresta sintática (treebank).

O Projeto Floresta Sintá(c)tica é uma colaboração entre a Linguateca e o projeto VISL. Contém textos em português (do Brasil e de Portugal) anotados (analisados) automaticamente pelo analisador sintático PALAVRAS (Bick 2000) e revistos manualmente por lingüistas.

Atualmente, o corpus da Floresta Sintá(c)tica tem 4 partes, que diferem quanto ao gênero textual, quanto ao modo (escrito vs falado) e quanto ao grau de revisão lingüística: o Bosque, totalmente revisto por lingüistas; a Selva, parcialmente revisto, a Floresta Virgem e a Amazônia, não revistos. Junto, todo esse material soma cerca de 261 mil

frases (6.7 milhões de palavras) sintaticamente analisadas.

Uma floresta sintática serve para o treino e avaliação de analisadores morfossintáticos; para estudos baseados em corpus, e para uma investigação da língua, não apenas da sintaxe, mas também de aspectos semânticos e discursivos. Pode, ainda, ser um auxiliar no ensino.

Para nosso estudos de desenvolvimento de um parser probabilístico para a língua portuguesa e treino da ferramenta desenvolvida por Bikel , será utilizado o Bosque, parte da floresta sintática completamente revisada por lingüistas.

O Bosque é composto por 9.368 frases, retiradas os primeiros 1000 extratos (aprox.) dos corpora CETENFolha e CETEMPúblico. Desde 2007, o Bosque vem passando por um novo processo de revisão, em que foram corrigidas algumas pequenas inconsistências e acrescentadas novas etiquetas. A versão final, disponível para consulta e download, é o Bosque 8.0.

Este é o corpus mais correto da Floresta, e por isso o mais aconselhado para pesquisas em que não se prioriza tanto a quantidade, mas sim a precisão dos resultados.

Uma quantificação das etiquetas usadas no Bosque pode ser encontrada no anexo 4 da Bíblia Florestal, uma extensa documentação das opções lingüísticas tomadas durante o projeto.

5.2.2.4.1 Esquema de anotação Na Floresta, a cada palavra são associadas etiquetas, etiquetas principais (de função e de forma) e secundárias: Estas etiquetas aparecem como FUNÇÃO:forma. Em a menina gulosa, por exemplo, temos:

```
>N          a
H:n         menina
N<adj       gulosa
```

Em que, para o "a", > N é FUNÇÃO, e indica que a palavra em questão é dependente à esquerda (por isso o sinal ">") de um núcleo nominal (N). Já a forma de "a"é artigo definido. Por isso, o par >N:artd.

"menina"é o núcleo do sintagma nominal, por isso a FUNÇÃO é H. Como a palavra em questão é um nome, a forma é n.

Por fim, o adjetivo "gulosa". Como é um dependente (modificador) à direita do nome, recebe a etiqueta de FUNÇÃO (N<) e a etiqueta de forma adj.

A cada palavra também é associado o seu lema, e informações morfossintáticas (gênero, número, tempo, modo e pessoa para os verbos e, eventualmente, outras etiquetas indicativas de fenômenos como elipse, construções de foco etc). Aqui está um glossário com todas as etiquetas secundárias utilizadas na Floresta. Em a menina gulosa, o acréscimo das etiquetas secundárias leva ao seguinte formato:

```
>N:art('a' <artd> F S)      a
H:n('menina' F S)           menina
N<: adj('guloso' F S)       gulosa
```

Versão atual do Bosque: versão 8.0, de 13 de Outubro de 2008, 9.437 árvores revistas, correspondendo a 1962 extratos, 215.420 unidades, aprox. 183.619 palavras

5.2.2.4.2 Glossário com as etiquetas de função e de forma utilizadas na Floresta. Vemos em seguida uma amostra do tagset da Floresta Sintatica nas tabelas abaixo. Uma detalhada descrição do tagset da Floresta é encontrado no website do projeto Floresta Sintática em
http://linguateca.dei.uc.pt/Floresta/BibliaFlorestal/anexo1.html.

Tabela 7: *Tags de Part-of-Speech da Floresta Sinática.*

Símbolo	**Categoria**
N	nome, substantivo
PROP	nome próprio
ADJ	Adjectivo
N-ADJ	flutuação entre substantivo e adjectivo
V-FIN	verbo finito
V-INF	Infinitivo
V-PCP	Particípio
V-GER	Gerúndio
ART	Artigo
PRON-PERS	pronome pessoal
PRON-DET	pronome determinativo
PRON-INDP	pronome independente (com comportamento semelhante ao nome)
ADV	Advérbio
NUM	Numeral
PRP	Preposição
INTJ	Interjeição
CONJ-S	conjunção subordinativa
CONJ-C	conjunção coordenativa

5.2.2.4.3 Tags de Part-of-Speech

Tabela 8: *Tags Sintáticos da Floresta Sintática .*

Símbolo	**Categoria**
NP	sintagma nominal (H: nome or pronome)
ADJP	sintagma adjectival (H: adjectivo ou determinante)
ADVP	sintagma adverbial (H: advérbio)
VP	sintagma verbal (contém sempre MV e poderá exibir AUX)
PP	sintagma preposicional (H: preposição)
CU	sintagma evidenciador de relação de coordenação
SQ	sequência de funções discursivas; sequência de elementos identificadores do falante, tema, etc. e do discurso propriamente dito
FCL	oração finita
ICL	oração ão-finita
ACL	oração averbal

5.2.2.4.4 Tags sintáticos

5.2.2.5 Corpus do projeto Semantic Share

O Corpus definido no projeto Semantic Share será detalhado posteriormente neste trabalho.

5.3 Diagrama geral do processo de parsing estatístico baseado em *corpus*

Uma visão geral do processo de *parsing* estatístico pode ser observado na Figura 3.

No desenvolvimento de um *parser* estatístico baseado em *corpus*, o *corpus* anotado é dividido em 3 partes:

1. Treino

   Sentenças que o sistema usa para aprender.

2. Desenvolvimento (ou teste de desenvolvimento)

   Sentenças utilizadas para avaliar a qualidade do *parser* obtido a cada passo do desenvolvimento. Como a análise também é qualitativa, o processo de realimentação para correção do *parser* tem, fatalmente, um aspecto tendencioso. Ou seja, com o tempo, o *corpus* de desenvolvimento perde a isenção para representar resultados confiáveis, pois o desenvolvedor acaba adaptando o *parser* para corrigir especificamente os erros feitos naquelas sentenças. Isto é conhecido como *"overfitting"*.

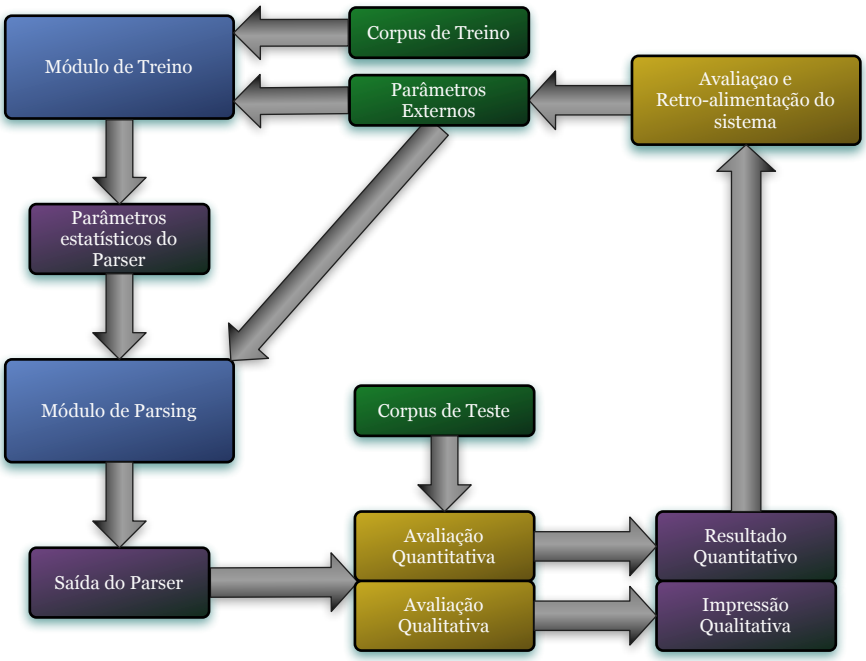

Figura 3: Estágios do processamento de linguagem natural

3. Teste final

   É semelhante ao de desenvolvimento, porem é proibida a análise qualitativa sobre o mesmo.

O modulo de geração do *parser* tem como entrada os exemplos do *corpus* de treino e gera os parâmetros estatísticos que serão utilizados pelo *parser* para tomar as decisões. Este módulo é parametrizável com informações linguísticas fornecidas pelo desenvolvedor, que guiam a interpretação do *corpus* de treino. Por exemplo, a informação de que, quando um constituinte tem dois nomes seguidos, o núcleo é o da esquerda para o português e é o da direita para o inglês.

O *parser* gerado é composto pelo módulo de *parsing* que recebe as sentenças de entrada e toma as decisões de análise guiado pelos parâmetros estatísticos aprendidos, gerando a sentença analisada.

Cada vez que uma nova versão (ou seja, um novo conjunto de parâmetros estatísticos) do *parser* é gerada , ele é testado e os resultados do teste usados para realimentar o processo. Este teste é feito sobre o *corpus* de desenvolvimento. Os resultados são analisados qualitativa e quantitativamente. Com base nestes valores, pode-se avaliar, por exemplo se a nova versão é melhor ou pior que as anteriores e o que se pode fazer para melhorar.

5.4 *Parsers* para Português

Conforme mencionado anteriormente, existem alguns trabalhos de construção de *parsers* para o português. Dentre eles, os de Eckhard Bick [BIC00], baseado em regras, e portanto difícil de ser expandido ou adaptado; [WIN06] e [BON03], que também são estatísticos, baseados no modelo de Collins. Entretanto, revisando o que já referimos, os resultados até agora obtidos ainda estão distantes dos desejados.

6 Projeto Semantic Share

6.1 Objetivos principais

Um objetivo principal do SemanticShare é o desenvolvimento para o Português de corpora anotados da mais recente geração e da próxima geração - um PropBank e um LogicalFormBank -, dos quais uma parte é paralela a bancos de dados similares que estão a ser produzidos para outros idiomas, noutros projetos.

6.2 Características principais

Estes corpora são diferentes materializações de um banco único de enunciados e correspondentes representações gramaticais, com as seguintes características principais:

6.2.1 Abrangência

Contêm informação morfológica, sintática e semântica integrada.

6.2.2 Acessibilidade

Podem ser apresentadas em uma ou mais de entre várias vistas:

1. Frases
2. Segmentos lexicais
3. Lemas
4. Traços de flexão
5. Etiquetas morfossintáticas

6. Entidades nomeadas e unidade multi-palavra
7. Árvores de constitutintes
8. Árvores de funções e papéis semânticos
9. Formas lógicas

6.2.3 Precisão

Cada representação é arquivada por seleção humana, depois de ter sido gerada por uma analisador gramatical;

6.2.4 Profundidade

São arquivadas num formato de representação interno que é lingüisticamente bem informado, seguindo um quadro gramatical de primeira linha para a lingüística computacional (HPSG);

6.2.5 Evolução

São apoiadas por ferramentas de desenvolvimento de corpora avançadas que asseguram uma extensão fácil das estruturas anotadas quando mais informação de mais dimensões lingüísticas possa ter de ser adicionada em extensões futuras (e.g. tempo, resolução de anáfora, etc), ou quando a cobertura da gramática seja aprofundada.

6.3 Infra-estrutura de investigação

Estes objetivos estão ao alcance do projeto na medida em que tiram partido da maioria das ferramentas e recursos desenvolvidos pela equipa do SemanticShare em projetos anteriores bem sucedidos, nomeadamente o projeto TagShare, do qual o SemanticShare é uma continuação. Eles constituem uma coleção única de ferramentas de última geração para o Português – segmentador de frases e lexemas, etiquetador morfossintático, lematizador, analisador morfológico, reconhecedor de entidades nomeadas –, juntamente com o respectivo corpus de 1 milhão de ocorrências, anotado com precisão de acordo com as dimensões 1.-6. acima (serviços online em http://nlxgroup.di.fc.ul.pt).

6.4 Apoio de uma comunidade internacional de investigadores

Tudo isto será realizado com o apoio do consórcio DELPH-IN, uma iniciativa de nível mundial que visa dinamizar investigação de ponta em processamento lingüístico profundo através da partilha de ferramentas de desenvolvimento "open source", de recursos e de boas práticas (http://wiki.delph-in.net) entre os seus participantes convidados (vd. carta de convite em http://www.di.fc.ul.pt/ ahb/semanticshare.htm).

A sua plataforma tecnológica e a sua ferramenta de anotação sem rivais permitem avanços rápidos na concretização dos objetivos do projeto, dando assim continuidade à cooperação anterior, nomeadamente no quadro do projeto GramaXing, em que uma gramática para o processamento lingüístico profundo foi desenvolvida e está a ser mantida.

Para além disso, parte do banco lingüístico a ser desenvolvido é a componente portuguesa de bancos paralelos que estão a ser desenvolvidos para outros idiomas por outros membros do DELPH-IN, segundo requisitos similares.

Estes corpora anotados representam recursos chave para o processamento do Português, incluindo:

- fornecer uma base empírica para o estudo lingüístico deste idioma e para o desenvolvimento de ferramentas elaboradas manualmente;
- treinar ferramentas de base estatística para o processamento superficial e profundo, incluindo parsers, etiquetadores de papéis semânticos, etc;
- avaliar ferramentas de processamento;
- apoiar a experimentação de abordagens inovadoras em PLN multilingue, incluindo tradução automática estatística ou meta-anotação automática para a web semântica, etc...

6.4.1 Esquema de Anotação do Projeto Semantic Share

O esquema de anotação do projeto Semantic Share são visões baseadas na anotação base do projeto que utiliza HPSG [BRA08].

A partir desta anotação são extraídas "visões"no formato de árvores de constituíntes. estas árvores formam o corpus de pesquisa utilizado neste trabalho.

Vemos em seguida uma amostra do tagset extraído do corpus definido no projeto Semantic Share.

6.4.1.1 Tags de Part-of-Speech

Tabela 9: *Tags de Part-of-Speech do projeto Semantic Share.*

Símbolo	**Categoria**
A	Adjetivo
ADV	Adverbio
C	Complementador (que)
CARD	Cardinal
CONJ	Conjunção
D	Determinador
DEM	Pronome demonstrativo
N	Nome
P	Preposição
PNT	Pontuação
POSS	Pronome possessivo
PPA	Particípio passado
QNT	Quantificador
V	Verbo

6.4.1.2 Tags sintáticos

Tabela 10: *Tags sintáticos do projeto Semantic Share.*

Símbolo	**Categoria**
ADVP	Sintagma adverbial
AP	Sintagma Adjetival
CONJP	Sintagma coordenativo
CP	Sintagma Complementizador
NP	Sintagma nominal
N'	
pp	Sintagma preposicional
PPA'	
PPAP	Sintagma de oração Passiva
S	Sintagma de sentença
S/ADVP	Sintagma de sentença adverbial
S/AP	Sintagma de sentença adjetival
S/NP	Sintagma de sentença nominal
S/PP	Sintagma de sentença preposicional
SNS	Sintagma de sentença sem sujeito
SNS/ADVP	Sintagma de sentença sem sujeito adverbial
SNS/NP	Sintagma de sentença sem sujeito nominal
VP	Sintagma verbal
VP/ADVP	Sintagma verbal adverbial
VP/AP	Sintagma verbal adjetiva
VP/NP	Sintagma verbal nominal
VP/PP	Sintagma verbal preposicional

7 Modelos Probabílisticos de Michael Collins

7.1 Gramatica Livre de Contexto Probabilística

Para descrever os modelos probabilísticos de parsing de Michael Collins antes precisamos entender um pouco Gramática Livre de Contexto Probabilística (que a partir de agora vamos nos referir como PCFG).

PCFGs são uma extensão de uma gramática livre de contexto , só que existe uma probabilidade associada a cada regra de produção.

O uso de técnicas estatísticas para o aprendizado de gramáticas foi inspirado no sucesso dessas técnicas para o processamento de fala (Charniak 1993). O modelo proposto em uma gramática livre de contexto probabilística (PCFG) faz uma suposição de independência que considera a probabilidade de cada regra de sintagma independente de todos os outros sintagmas na sentença. A ordem de derivação não afeta o modelo. As probabilidades nas PGFGs, atribuídas às regras, são encaradas como a probabilidade do sintagma-pai usando tal regra, nos subelementos descritos, em comparação a todas as outras regras que expandem o mesmo sintagma.

As gramáticas probabilísticas têm muitas vantagens. Sendo elas são extensões óbvias das gramáticas livres de contexto, os algoritmos usados para GLCs podem ser transportados para as PCFGs, permitindo que todas as possíveis análises possam ser encontradas num tempo de ordem n^3, em que n é o tamanho da sentença.

Um algoritmo bastante usado é o Inside-Outside (Charniak 1993). Usando um corpus grande e o algoritmo, o modelo pode ser treinado automaticamente de um modo completamente não supervisionado, considerando todas as análises possíveis da sentença no corpus de treinamento.

A Ambigüidade é talvez seja maior problema na analise de sentenças. Uma gramática

probabilística oferece solução para este problema, escolhendo a interpretação mais provável no momento da analise.

Outro importante uso de uma gramática probabilística esta em modelar uma linguagem para reconhecimento de fala.

Vamos exemplificar:

Na frase "vi o homem no monte com os binóculos",

Supondo a gramática:

1. $S \rightarrow SV|SVSP$
2. $SV \rightarrow vSN|vSNSP$
3. $SN \rightarrow SNSP|SNSPSP|SNSPSPSP$

Existe grande quantidade de árvores geradas, e seria possível relacionar "o homem"com "os binóculos"e "vi"com "no monte".

O caso anterior é genuinamente ambíguo, mas há muitos casos de ambiguidade que se devem apenas à gramática em si.

Outro exemplo :

Indique-me um hotel com piscina com água quente

1. $S \rightarrow SV|SVSP$
2. $SV \rightarrow SNSP|SNSPSP$

O hotel é que tem água quente?

Portanto, temos um problema quanto a descobrir qual a árvore de análise correta.

Como solução poderíamos deixar que as situações de ambigüidade sejam resolvidas pela análise semântica,usar regras de desambiguação manuais ou usar modelos probabilísticos para atribuir probabilidades às diferentes arvores.

Uma gramática livre de contexto probabilística (PGLC) é um tuplo (N, T, S0, R) composta por:

- N: Conjunto de símbolos não-terminais
- T: Conjunto de símbolos terminais
- S0: símbolos não-terminal, designado por símbolo inicial
- R: Conjunto de regras da forma $A \rightarrow \alpha[p]$, onde:
  - A é um símbolo não terminal;
  - α é uma cadeia de zero ou mais símbolos terminais e não terminais;
  - p é um número entre 0 e 1 que representa $P(\alpha|A)$

$p = P(\alpha|A)$ = probabilidade de um dado não terminal A ser expandido na expressão α.

$p = P(\alpha|A) = P(A \rightarrow \alpha) = P(A \rightarrow \alpha|A) =$

$P(RHS|LHS)$, em que:

$RHS = right - hand - side$

$LHD = left - hand - side$

Sendo:

$\sum_{\alpha} P(A \rightarrow \alpha) = 1$

Podemos usar uma PCFG para estimar a probabilidade associada a uma dada árvore, o que vai permitir arranjar uma solução para os casos de ambiguidade.

Mas antes, temos de calcular as probabilidades associadas a cada regra.

- Supondo n regras da forma $A \rightarrow \alpha i$
- Calcula-se $P(A \rightarrow \alpha i|A)$
- Podem-se calcular estes valores, fazendo contagens num TreeBank por exemplo.
- Estas contagens podem ser feitas através da seguinte fórmula:

$$P(A \rightarrow \alpha_i) = \frac{count(A \rightarrow \alpha_i)}{\sum_{j=1}^{N} count(A \rightarrow \alpha_i)} = \frac{(A \rightarrow \alpha_i)}{count(A)}$$

Consegue-se deste modo associar probabilidades às regras e construir uma gramática probabilística:

1. $SV \rightarrow Verbo$[.50]
2. $SV \rightarrow VerboSN$[.45]
3. $SV \rightarrow VerboSNSN$[.05]

Considerando que a probabilidade de cada constituinte é independente do contexto em que aparece na árvore global de análise, temos que a probabilidade do constituinte A, obtido a partir de $a1, ..., an$, através da regra $A \rightarrow a1...an$ é dada por:

$P(A) = P(A \rightarrow \alpha$ 1 ... n $|A)$ * P(α 1) * ...* P(α n)

Adicionalmente
Nas folhas da árvore, usam-se as probabilidades POS $P(\alpha i|wi)$

Dado que uma árvore é constituída pelo conjunto de regras que participam na sua derivação, a probabilidade associada a cada árvore é o produto das probabilidades das regras usadas na sua derivação.

7.2 Trabalhos Anteriores, historico de Parsing Probabilistico para PLN

O aprendizado estatístico se insere num contexto cuja linha de pesquisa é chamada de empírica, uma vez que se baseia em exemplos já prontos e aprende como lidar com aqueles ainda não vistos. De acordo com Manning e Schütze (1999), a linha empiricista, que entre as décadas de 60 e 80 ficou nas sombras de crenças racionalistas encabeçadas por Chomsky (1965), cujo reflexo dentro da Inteligência Artificial caracterizava-se pela criação

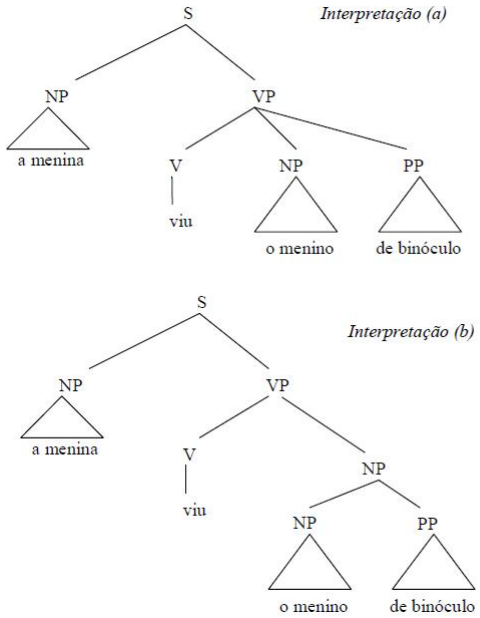

Figura 4: Imagem das árvores da frase, *A menina viu o menino*

de sistemas inteligentes com grande quantidade de conhecimento inicial codificado à mão, ressurgiu na década de 90, com a idéia de que o conhecimento pode ser induzido a partir de algumas operações básicas de associação e generalização (Mitchell 1980). Assim, segundo o empiricismo, uma máquina poderia aprender a estrutura de uma linguagem apenas observando uma grande quantidade de exemplos, usando procedimentos estatísticos gerais e métodos de associação e generalização indutiva, como aprendizado indutivo de regras ((Brill 1993), (Hermjakob e Mooney 1997).

A Ambigüidade foi Um dos maiores obstáculos encontrados pelos sistemas automáticos de parsing (parsers) surge quando estes se deparam com sentenças que possuem algum tipo de ambigüidade sintática, ou seja, sentenças com duas ou mais árvores possíveis. Por exemplo, na sentença "A menina viu o menino de binóculo", a atribuição dos sintagmas sintáticos pode mudar a interpretação que se faz da frase. duas árvores de representação sai possíveis para a mesma sentença, nas quais, de acordo com a estrutura da primeira, a menina é que estava de binóculo e observou o menino, e na segunda, a menina observou o menino que estava de binóculo conforme ilustrado na Figura 4. Nesses casos, é necessário que o parser opte por uma delas, e que, de preferência, seja a que se esteja buscando.

Assim, para que o parser pudesse fazer a atribuição procurada, ele precisaria de uma certa interpretação de significado que o ajudasse a fazer a escolha correta. No entanto, tais sistemas são totalmente desprovidos de quaisquer informações nesse sentido. Muitos concordam que essa não é uma função do parser, delegando tal responsabilidade a uma

unidade especial de desambiguação. Os parsers estatísticos utilizam medidas de probabilidades observadas em sentenças previamente analisadas como critério de desempate em prováveis ações de desambiguação. Portanto, para que funcione, é necessário que se tenha um conjunto bastante representativo de sentenças com suas respectivas árvores sintáticas. Desse modo, o parser atribuirá probabilidades às possíveis análises de uma sentença, apresentando como resposta aquela de maior probabilidade, sendo isso feito em três passos: (1) encontra todas as possíveis análises; (2) atribui-lhes probabilidades, e (3) seleciona a de mais alta probabilidade.

Um importante fator influenciador nas pesquisas no campo de PLN foi a disponibilização de grandes corpus (treebanks) anotados usados para treino dos analisadores probabilísticos.

A falta de treebanks no inicio das pesquisas com parsers estatísticos gerou um grande numero de abordagens quanto ao treino e avaliação dos parsers existentes.

7.3 Problemas encontrados na Gramática Livre de Contexto Probabilística

PCFGs foi o ponto de inicio natural para as pesquisas em desenvolvimento de parsers estatísticos para PLN. Suas propriedades formais foram bem compreendidas, algoritmos eficientes para parsing eram bem conhecidos e descritos.

Desafortunadamente, após algumas pesquisas descobriu-se que o uso apenas de PCFGs era insuficiente para PLN por vários aspectos. Nos casos de treino não supervisionados alguns algoritmos não tiveram sucesso quanto as estruturas lingüísticas.. No caso de treino supervisionado, PCFGs também se mostrou um modelo pobre com resultados não muito expressivos.

A percepção dessas falhas na utilização de PCFGs em PLN gerou estudos profundos em três áreas na tentativa se achar métodos apropriados e eficientes para PLN estatística.

1. Desenvolvimento de Modelos mais sensíveis quanto as estruturas de linguagem.

2. O desevolvimento de modelos contendo parâmetros correspondentes as dependências léxicas.

3. E o desenvolvimento dos chamados "history-based modes"ou abordagem a modelos históricos.

7.4 Métodos Probabilísticos com aumento de sensibilidade estrutural ou ao contexto

Muitas pesquisas foram voltadas ao objetivo de aumentar a sensibilidade ao contexto de uma PCFG, tendo resultados encorajadores, [pág 106 tese collins], Considerando um modelo baseado em regras, mais uma vez, com mais sensibilidade ao contexto que PCFG. Outras pesquisas como [pág 106 collins] considerando versões parcialmente supervisionadas do algoritmo Inside-Outside: Considerando a idéia de que o treebank como TBI, relativamente plano, com arvores não tão anotadas, e que o algoritmo de aprendizagem seria capaz de usar as informações desse treebank, enquanto aprendia mais detalhes de maneira não supervisionada.

Todos esses modelos continuavam com falta de sensibilidade lexical.

7.5 Formalismo incluindo dependência lexical

Existem pelo menos duas razoes para o desenvolvimento de modelos que incluem dependência de parâmetros. Primeiro, Pesquisas interessadas em modelagem para reconhimento da fala imaginavam que enquanto modelos "trigram"teriam modelos sintáticos pobres, as probabilidades associadas aos pares ou triplos de palavras era muito úteis quando tinham probabilidades associadas as sentenças na linguagem. Segundo, o mais importante apontado por Michael Collins para seus estudos, pesquisas sugeriram que a dependência de probabilidade seria poderosa para abordar o problema da ambigüidade.

7.6 History-Base Models

Uma terceira linha de pesquisa, history-based models, foi desenvolvida por pesquisadores da IBM[pág 107 collins]. Estes modelos foram caracterizados por duas diferenças de uma simples PCFG. Primeiro. A arvore de analise representada foi incrementada de duas maneiras: tags não terminais foram extendidos para incluir informação como items léxicos (heads, words), ou categoria semântica; o condicionamento ao contexto foi extendido para prever a potencialidade de todas as estruturas anteriores geradas, ao invés de apenas expandir os símbolos não terminais como na PCFG. Segundo , mais poderoso método de aprendizado de maquina, em particular, arvores de decisão, foram usadas para estimar os parâmetros. A idéia básica foi expandir as funcionalidades e o contexto, in-

cluindo todas as fontes de informação para desambiguação ; então usar arvores de decisão para aprender exatamente qual conjunto de combinações era importante para analisar. Um importante avanço nessa pesquisa foi a troca de uso de gramática "hand-crafted"por modelos de treino utilizando treebanks [107 collins]

history-based , é um modelo gerativo probabilístico que usa a vantagem de possuir informação lingüística detalhada para resolver ambigüidade. Os autores abordam o sentido de contexto da forma mais fiel possível com o que nós humanos consideramos, racionalizando a quantidade de informação da sentença que é necessária e suficiente para determinar sua análise.

7.7 Modelos de Michael Collins

Collins (1996) propõe um modelo baseado em dependências lexicais entre bigramas. Este modelo usa informações lexicais para modelar relações núcleo-modificador. Tambem introduz um conceito de distância nesse modelo baseado em dependências entre bigramas. Segundo ele, a distância é uma variável crucial quando se decide se duas palavras estão relacionadas.

Após isso , Collins (1997) propoe três novos modelos gerativos de parsing, que usam uma nova abordagem para melhorar o modelo de bigramas, todos eles baseados na noção head-centering, em que o núcleo é o elemento principal e direcionador de todo o processode geração de uma árvore sintática.

Define uma probabilidade conjunta P(AS; S) sobre pares árvores-sentenças. Usa um modelo baseado na história da análise: uma árvore sintática é representada como uma seqüência de decisões, a partir de uma derivação top-down e centrada no núcleo da árvore sintática. Segundo o autor, a representação da árvore sintática dessa forma permite que suposições de independência sejam feitas, levando a parâmetros condicionados a núcleos lexicais: parâmetros de projeção do núcleo, subcategorização, colocação de complemento/adjunto, dependência, distância, ente outros parametros.

A seguir são apresentados cada um dos modelos. O Modelo 2 representa uma evolução em relação ao Modelo 1; e o Modelo 3, em relação ao Modelo 2.

7.7.1 Modelo 1

Este modelo apresenta uma proposta de como estender uma Gramática Livre de Contexto Probabilística (PCFG) para uma gramática lexicalizada (que considera itens lexicais). O Modelo 1 tem ainda parâmetros que correspondem a dependências entre pares de núcleos; a distância também é incorporada como uma medida, generalizando o modelo para uma abordagem baseada na história da análise.

Vantagens do modelo em relação ao anterior:

- o modelo não é deficiente (i.e., $\sum P(AS, S) = 1$); regras unárias são manuseadas de uma forma bastante natural pelo modelo;
- a medida de distância é melhorada. Por exemplo, a variável de adjacência passa a corresponder diretamente a estruturas de ligação à direita;
- o modelo mostra uma melhora de 1.9
- o modelo pode condicionar sua estimativa usando qualquer estrutura previamente gerada;
- a etiquetagem morfossintática é naturalmente incorporada ao modelo;
- o modelo define uma medida de probabilidade conjunta $P(AS, S)$, podendo ser treinado de uma maneira não supervisionada pelo algoritmo Expectation Maximization.

A geração do lado direito da regra é quebrada em uma seqüência de pequenos passos. Cada regra passa a ter a forma

$$Pai(nuc) = E_n(pe_n)...E_1(pe_1)NUC(nuc)D_1(pd_1)...D_m(pd_m)$$

em que $NUC(nuc)$ representa o núcleo do sintagma, que recebe o item lexical nuc de seu pai Pai; $E_1...E_neD_1...D_m$ são seus modifiadores, à esquerda e à direita, com itens lexicais pe e pd, respectivamente. As seqüências à direita e à esquerda são aumentadas com um símbolo STOP, de forma que permita um processo de Markov para o modelo. Assim, $E_{n+1} = D_{m+1} = STOP$.

A regra de probabilidade pode ser reescrita usando a regra da cadeia de probabilidades:

$$P(E_{n+1}(pe_{n+1})...E_1(pe_1)NUC(nuc)D_1(pd_1)...D_{m+1}(pd_{m+1})|Pai(nuc)) =$$

$$P_{nuc}(NUC|Pai(nuc))\times$$

$$\prod_{i=1..n+1} P_{esq}(E_i(pe_i)|E_1(pe_1)...E_{i-1}(pe_{i-1}), Pai(nuc), NUC)\times$$

$$\prod_{i=1..m+1} P_{dir}(D_j(pd_j)|E_1(pe_1)...E_{n+1}(pe_{n+1}), D_1(pd_1)...D_{j-1}(pd_{j-1}), Pai(nuc), NUC)$$

Para um modelo ser Modelo Baseado na História da Análise (MBHA), cada modificador poderia depender de qualquer função F dos modificadores anteriores, categoria do núcleo/pai e núcleo.

$$P_{esq}(E_i(pe_i)|E_1(pe_1)...E_{i-1}(pe_{i-1}), Pai(nuc), NUC) =$$

$$P_{esq}(E_i(pe_i)|\Theta(E_1(pe_1)...E_{i-1}(pe_{i-1}), Pai(nuc), NUC))$$

$$P_{dir}(D_j(pd_j)|E_1(pe_1)...E_{n+1}(pe_{n+1}), D_1(pd_1)...D_{j-1}(pd_{j-1}), Pai(nuc), NUC) =$$

$$P_{dir}(D_j(pd_j)|\Theta(E_1(pe_1)...E_{n+1}(pe_{n+1}), D_1(pd_1)...D_{j-1}(pd_{j-1}), Pai(nuc), NUC))$$

Fazendo a suposição de independência de que os modificadores são gerados independentemente uns dos outros, ou seja, fazendo F ignorar tudo a não ser P, NUC e nuc, temos

$$P_{esq}(E_i(pe_i)|E_1(pe_1)...E_{i-1}(pe_{i-1}), Pai(nuc), NUC) = P_{esq}(E_i(pe_i)|Pai(nuc), NUC)$$

$$P_{dir}(D_j(pd_j)|E_1(pe_1)...E_{n+1}(pe_{n+1}), D_1(pd_1)...D_{j-1}(pd_{j-1}), Pai(nuc), NUC) =$$

$$P_{dir}(D_j(pd_j)|Pai(nuc), NUC)$$

A geração de um lado direito de um regra, dado o lado esquerdo, é então feita em três passos, sucessivamente, até que toda a árvore seja construída: (1) gera-se o núcleo (NUC); (2) geram-se modificadores à esquerda (E) e (3) geram-se modifcadores à direita (D).

7.7.1.1 Adicionando Distancia

Assim como na proposta anterior (Collins 1996), Collins também adiciona distância a esse modelo (Collins 1997). Essa adição é importante para capturar preferências por estruturas de ligação à direita (que quase sempre traduz a preferência por dependências entre palavras adjacentes) e a preferência por dependências que não cruzam um verbo. A distância pode ser incorporada adicionando uma quantidade de dependência entre os modificadores.

$$P_{esq}(E_i(pe_i)|Pai, NUC, nuc, E_1(pe_1)...E_{i-1}(pe_{i-1}) =$$
$$P_{esq}(E_i(pe_i)|NUC, Pai, nuc, distancia_{esq}(i-1))$$

$$P_{dir}(D_i(pd_i)|Pai, NUC, nuc, D_1(pd_1)...D_{i-1}(pd_{i-1}) =$$
$$P_{dir}(D_i(pd_i) - NUC, Pai, nuc, distancia_{dir}(i-1))$$

A distância é um vetor contendo duas informações: adjacência (que permite aprender preferências por ligações à direita) e existência de um verbo entre eles (que permite aprender a preferência pela modificação do verbo mais recente).

A distância é um vetor contendo duas informações: adjacência (que permite aprender preferências por ligações à direita) e existência de um verbo entre eles (que permite aprender a preferência pela modificação do verbo mais recente).

7.7.2 Modelo 2

O Modelo 2, proposto por Collins, introduz a distinção entre complemento/adjunto. Os complementos são acrescidos do sufixo "C". Assim, o modelo é estendido para fazer essa distinção e também para ter parâmetros que correspondam diretamente a distribuições de probabilidade sobre subcategorizações para núcleos. O processo gerativo passa então a incluir escolha probabilística de subcategorização à esquerda ou à direita:

1. Escolhe o núcleo com probabilidade $P_{nuc}(NUC|Pai, nuc)$

2. Escolhe subcategorizações à esquerda e à direita, E-C e D-C, com probabilidades $P_{esq}(E-C|Pai, NUC, nuc)$ e $P_{dir}(D-C|Pai, NUC, nuc)$. Cada subcategorização é um conjunto que especifica os complementos que o núcleo requer como modificadores à direita ou à esquerda.

3. Gera modificadores à esquerda e à direita com probabilidades

   $P_{esq}(E_i(pe_i)|NUC, Pai, nuc, distancia_{esq}(i-1), E-C)$ e

   $P_{dir}(D_i(pd_i)|NUC, Pai, nuc, distancia_{dir}(i-1), D-C)$

Conforme os complementos são gerados, eles são removidos do conjunto de subcategorização (SUBCAT) apropriado. A probabilidade de gerar o símbolo STOP é 1 quando SUBCAT estiver vazio, e a probabilidade de gerar um complemento será 0 quando ela não estiver no SUBCAT.

7.7.3 Modelo 3

O Modelo 3 é estendido da gramática de estrutura de frase generalizada para possibilitar tratamento de Wh-moviment. Introduz parâmetros TRACES e Wh-Moviment. Por exemplo, na frase "The store that IBM bought last week", o modelo usaria as regras para gerá-la:

1. $SN \rightarrow SNSBAR(+gap)$
2. $SBAR(+gap) \rightarrow Wh_{sn}S-C(+gap)$
3. $S(+gap) \rightarrow SN-CSV(+gap)$
4. $SV(+gap) \rightarrow VerboTraceSN$

SBAR é a representação para subcláusula; gap é a indicação de que falta algo naquele espaço

7.7.3.1 Casos especiais

Os SN_Base não sofrem suposições de independência. A probabilidade da regra $SN\_Base(cao) \rightarrow Determinante(o)Nome(cao)$, é estimada como:

$$P_{nuc}(Nome|SN_Base, cao) \times P_{esq}(Determinante(o)|SN_Base, Nome, cao) \times$$

$$P_{esq}(STOP|SN_Base, Nome, cao) \times P_{dir}(STOP|SN_Base, Nome, cao)$$

7.7.3.2 Coordenação

Ao invés de a coordenação ser gerada num processo normal, como mais um modificador, o processo gerativo foi alterado para gerar a coordenação e o sintagma seguinte num único passo; assim, um não-terminal e um flag binário coord são gerados; coord = 1 se há um relacionamento de coordenação. Por exemplo, a regra $SN(homem) \rightarrow SN(homem)C(e)SN(cao)$,teria a seguinte distribuição de probabilidade:

$$P_{nuc}(SN|SN(homem)) \times P_{esq}(STOP|SN, SN, homem)$$

$$P_{dir}(SN(cao), coord = 1|SN, SN, homem) \times P_{dir}(STOP|SN, SN, homem)$$

$$P_c(C, e|SN, SN, SN, homem, cao)$$

8 Dan Bikel Statistical Parsing Engine

8.1 Estudo dos parametros de implementação do Statistical Parsing Engine de Dan Bikel

9 Avaliação

As medidas de avaliação do *parser* seguirão a proposta de GEIC/Parseval [BLA91], possivelmente adaptado conforme [COL97] para ignorar pontuação e não considerar a marcação de POS na avaliação. Em particular, serão usadas as medidas de *Labeled Precision* (LP) e *Labeled Recall* (LR) e sua média harmônica ($F_{\beta=1}$), descritas abaixo:

$$LP = \frac{número\ de\ constituintes\ corretas\ na\ análise\ proposta}{número\ de\ constituintes\ da\ análise\ proposta}$$

$$LR = \frac{número\ de\ constituintes\ corretas\ na\ análise\ proposta}{número\ de\ constituintes\ do\ treebank\ analisado}$$

$$F_{\beta=1} = \frac{2 * LP * LR}{LP + LR}$$

O termo *Labeled* se refere ao fato de que uma constituinte, para contar como corretamente recuperado, deve acertar a extensão correta do texto bem como o rótulo do constituinte.

10 Resultados Parciais

Resultados parciais

Referência Bibliografia

[ABE03] (Abeillé, Anne,Eds.). **Treebanks**: building and using parsed corpora. Kluwer Academic Publishers, 2003.

[ALL95] Allen, James. **Natural language understanding**. The Benjamin/Cummings Publishing Company, Inc., 1995.

[AST97] Aston, Guy. **Small and large corpora in language learning**. *PALC Conference*, 1997.

[BIB90] Biber, Douglas. **Methodological issues regarding corpus-based analyses of linguistic variation**. *Literary and Linguistic Computing*, 1990.

[BIB93] Biber, Douglas. **Representativeness in corpus design**. *Literary and Linguistic Computin*, 1993.

[BIC00] Bick, Eckhard. **The parsing system palavras, automatic grammatical analysis of portuguese in a contraint grammar framework**. Aarhus University Press, 2000.

[BIK04] Bikel, Dan. **Intricacies of collins' parsing model**. *Computational Linguistics*, 2004.

[BLA91] Black, Ezra; Abney, Steven; Gdaniec, C.; Grishman, Ralph; Harrison, P.; Hindle, Don; Ingria, R.; Jelinek, Fred; Klavans, Judith; Liberman, Mark; Marcus, Mitchell; Roukos, Salim; Santorini, Beatrice; Strzalkowski, T. **A procedure for quantitatively comparing the syntactic coverage of english grammars**. In: *Proceedings of the DARPA Speech and Natural Language Workshop*, San Mateo, CA, USA., 1991.

[BLA93] Black, Ezra; Jelinek, Fred; Lafferty, John; Magerman, David M.; Mercer, Robin; Roukos, Salim. **Towards history-based grammars**: using richer models for probabilistic parsing. In: *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, OH, USA., 1993.

[BON03] Bonfante, Andréia Gentil. **Parsing probabilístico para o português do brasil**. 2003. Tese de Doutorado.

[BRA08] Branco, Antonio; Costa, Francisco. **A computational grammar for deep linguistic processing of portuguese**: lxgram, version a.4.1. 2008.

[BRA09] Branco, António. **Semantic share project**. 2009. (http://semanticshare.di.fc.ul.pt/. (Último acesso em Abril 2009)).

[BRI95] Brill, Eric. **Transformation-based error-driven learning and natural language processing**: a case study in part-of-speech tagging. *Computational Linguistics*, 1995.

[CHA97] Charniak, Eugene. **Statistical techniques for natural language parsing**. *AI Magazine*, 1997.

[COL97] Collins, Michael. **Three generative, lexicalised models for statistical parsing**. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain., 1997. p.16–23.

[COL99] Collins, Michael. **Head-driven statistical models for natural language parsing**. 1999. Tese de Doutorado.

[GÓM97] Gómez, Pascual Cantos; Pérez, Aquilino Sánchez. **Predicability of word forms (types) and lemmas in linguistic corpora. a case study based on the analysis of the cumbre corpus**: an 8-million word corpus of contemporary spanish. *International Journal of Corpus Linguistics*, 1997.

[GÓM97a] Gómez, Pascual Cantos; Pérez, Aquilino Sánchez. **El ritmo incremental de palabras nuevas en los repertorios de textos**: estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas. Atlantis: Revista de la Asociación Española de Estudios Anglo-Norteamericanos, 1997.

[JUR00] Jurafsky, Daniel; Martin, James H. **Speech and language processing**. Prentice-Hall, 2000.

[LEE91] Leech, Geoffrey. **The state of art in corpus linguistics**. 1991.

[LIM01] Lima, Vera Lúcia Strube de. **Linguística computacional**: princípios e aplicações. *IX Escola da Informática da SBC-Sul*, 2001.

[MAN99] Manning, Christopher D.; Schütze, Hinrich. **Foundations of statistical natural language processing**. The MIT Press, Cambridge, MA, 1999.

[MAR94] Marcus, Mitchell; Kim, Grace; Marcinkiewicz, Mary Ann; MacIntyre, Robert; Bies, Ann; Ferguson, Mark; Katz, Karen; Schasberger, Britta. **The Penn Treebank**: annotating predicate argument structure. In: *Proceedings of the 1994 Human Language Technology Workshop.*, 1994.

[MAR93] Marcus, Mitchell P.; Santorini, Beatrice; Marcinkiewicz, Mary Ann. **Building a large annotated corpus of English**: the Penn Treebank. *Computational Linguistics*, v.19, n.2, p.313–330, 1993.

[PRO03] Prolo, Carlos A. **LR parsing for Tree Adjoining Grammars and its application to corpus-based natural language parsing**. June, 2003. Tese de Doutorado.

[SAR04] Sardinha, Tony Berber. **Linguística de corpus**. Manole, 2004.

[SIN97] Sinclair, John. **Corpus evidence in language description**. 1997.

[WIN06] Wing, Benjamim; Baldridge, Jason. **Adaptation of data and models for probabilistic parsing of portuguese**. *PROPOR 2006*, 2006.