Universidade Federal do Pampa
Pablo Botton da Costa

Análise Sintática Semi-supervisionada por Constituência Aplicada ao Português e Inglês

Alegrete
2014

Pablo Botton da Costa

Análise Sintática Semi-supervisionada por Constituência Aplicada ao Português e Inglês

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Fabio Natanael Kepler

Alegrete

2014

Pablo Botton da Costa

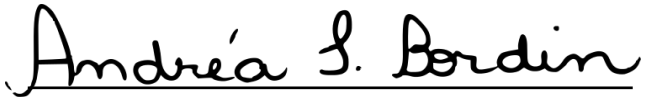
Análise Sintática Semi-supervisionada por Constituência Aplicada ao Português e Inglês

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

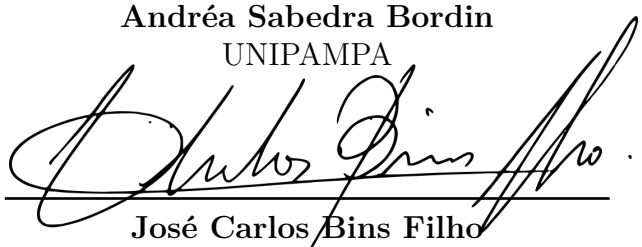
Trabalho de Conclusão de Curso defendido e aprovado em **17** de **março** de **2014**.

Banca examinadora:

Fábio Natanael Kepler
Orientador



Andréa Sabedra Bordin
UNIPAMPA



José Carlos Bins Filho
UNIPAMPA

Resumo

Este trabalho apresenta resultados em análise sintática semi-supervisionada para o português e inglês. O nosso modelo induz estruturas gramaticais através da modelagem do contexto, *span* e etiqueta sendo esta terna(contexto, *span* e etiqueta) utilizada para induzir os mais prováveis constituintes e etiquetas de uma sentença teste. O nosso modelo de analise sintática semi-supervisionada induz gramaticas através da constituência das palavras. O modelo apresenta também resultados para etiquetagem sintática, pois o nosso modelo induz os constituintes das frases e suas etiquetas se existirem. Como resultados obtivemos para o português uma medida-F não etiquetada em média geral de $17,30\%$ em relação ao melhor resultado possível para o córpus Tycho Brahe. Já para o inglês obtivemos resultados em média de $28,95\%$ abaixo do melhor resultado possível. Para a etiquetagem sintática nosso modelo obteve $20,79\%$ de alcance das frases de cada córpus do português, e dentre as que ele preenche obtém um acerto de $1,09\%$. Para etiquetagem aplicada ao inglês obtivemos $9,61\%$ das frases de cada córpus foram etiquetadas, e deste obtivemos um acerto de $1,71\%$ das etiquetas certas. Este trabalho apresentam resultados relevantes para córpus não etiquetado e também percebemos que o tamanho dos córpus tanto treinamento e de teste influenciam negativamente nos resultados, isto é se comprarmos o córpus Tycho Brahe e WSJ levando em consideração seu tamanho e resultados veremos que se tivermos uma grande quantidade de treinamento e teste o nosso modelo apresenta resultados baixos. Para resultados em etiquetagem sintática percebemos que o nosso modelo não foi muito efetivo, pois o modelo simplesmente seleciona a etiqueta com maior probabilidade ao invés de modelá-la como uma nova dimensão. Também apresentamos resultados dos modelos propostos por Klein e Manning (2004) aplicados ao português, quê não foram encontrados resultados na bibliografia.

Palavras-chave: Aprendizado de Máquina. Processamento de Linguagem Natural. Córpus com anotação sintática.

Abstract

This paper presents results on semi-supervised for Portuguese and English syntactic analysis. Our model induces grammatical structures through context modeling, span and label this triple(context,span and label) being used to induce the most likely constituents and labels of a test sentence. Our model of semi-supervised syntactic analysis induces grammars through the constituency of words. The model also presents results for syntactic tagging, because our model induces the constituents of sentences and their labels if they exist. As a result we obtained a measure-F for the Portuguese unlabelled overall average of 17.30% for the best results of corpus Tycho Brahe. Have we obtained results for the English average of 28.95% below the best possible result. For the syntactic tagging our model got 20.79% of the reach of sentences of each corpus of Portuguese, and among those he meets gets a hit of 1.09%. For labeling applied to the English got 9.61% of the sentences of each corpus were tagged, and this obtained a settlement of 1.71% of certain labels. This work presents results relevant to unlabeled corpus and also realize that the size of the corpus both training and testing negatively influence the results, ie if we buy the WSJ corpus Tycho Brahe and taking into account its size and results we see that if we have a large amount of training and testing our model shows lower results. For results in syntactic tagging realized that our model was not very effective because the model simply selects the label most likely rather than model it as a new dimension. We also present results of the models proposed by Klein e Manning (2004) applied to the Portuguese, what results were not found in the bibliography .

Key-words: Machine Learning. Computational Linguistics. Natural Language Processing. Corpus with syntactic annotation.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A análise sintática consiste em, dada uma frase qualquer, conseguir recuperar sua estrutura sintática, e para isso o analisador sintático analisa as estruturas internas das sentenças e suas relações lógicas de cada um de seus elementos. Uma das formas de analisar sintaticamente uma entrada, é através do computador onde existem várias formas de desempenhar esta tarefa automaticamente.

Um analisador sintático apresenta baixos resultados para linguagens humanas, como o português, pois as gramáticas humanas apresentam o problema da ambiguidade, que consiste em mais de uma regra gramatical para um determinada entrada. Para resolver o problema da ambiguidade utiliza-se a técnica de analise sintática probabilística que consiste em atribuir pesos a regras, gerando assim a possibilidade do analisador escolher ou tender a determinada regra. Para induzir estes valores bem como as regras que compõem esta gramática probabilística, o analisador utiliza-se de técnicas de aprendizado de máquina.

Os métodos de aprendizado de máquina como propõem Dempster, Laird e Rubin (1977) podem ser divididos em aprendizado supervisionado e não supervisionado. O aprendizado supervisionado apresenta bons resultados para a tarefa da análise sintática, mas necessita de um córpus relativamente grande com anotação sintática manual. Estes algoritmos de aprendizado supervisionado induzem as regras de uma gramática através destes córpus observando e aprendendo as estruturas das sentenças utilizando todos os recursos presentes nestes.

O aprendizado não-supervisionado é necessário quando os recursos dos córpus são limitados. A tarefa destes analisadores é induzir gramáticas através análise das estruturas das sentenças presentes nos córpus, induzindo assim regras que constituem aqueles elementos ou palavras.

O córpus é um conjunto de textos que são muito utilizados em tarefas de Processamento de Linguagem Natural (PLN), pois através deles conseguimos extrair informações sobre o tipo de texto, tamanho, regras gramaticais e entre outros. Para a tarefa de análise sintática seria interessante contarmos com um vasto conjunto destes textos, pois assim conseguiríamos extrair com maior facilidades as regras gramaticais de uma linguagem.

Para linguagens como o português seria interessante utilizarmos a técnica de análise sintática não supervisionada, pois os recursos linguísticos de anotação são escassos.

O analisador sintático não-supervisionado apresenta resultados não tão bons quanto o da análise supervisionada. Já Wang, Schuurmans e Lin (2008) demonstra em seu traba-

lho para análise de dependência onde utiliza aprendizado de máquina semi-supervisionado, combinando em seu modelo estruturas e etiquetas, com a possibilidade de melhoramento dos resultados das regras geradas desta gramática.

Através da motivação de o português necessitar de uma ferramenta de aprendizado não-supervisionado e este não apresentar resultados bons em comparação ao supervisionado, propomos a criação de uma método híbrido que induz estruturas constituintes e suas etiquetas sintáticas. Constituintes são palavras ou grupos de palavras que funcionam unitariamente em uma sentença.

1.1 Objetivos

Este trabalho tem como objetivo geral analisar e comparar resultados de análise sintática semi-supervisionada e não-supervisionada por constituência para o português e inglês.

1.2 Justificativa

Atualmente um analisador sintático é muito importante para diversas tarefas de Processamento de Linguagem Natural(PLN) como por exemplo extração e mineração de opinião, resumo de noticias e entre outras, pois as estruturas de constituência e classe de palavras que são gerados auxiliam a identificação e classificação das frases.

Para o português atualmente não existem trabalhos na área de analise sintática por constituência semi-supervisionada, agregando valor a este trabalho. Para ambas as línguas(Português e Inglês) o nosso trabalho e a bibliografia demonstram que a ainda a necessidade de se explorar nosvas propostas de analisadores sintáticos semi-supervisionados e não-supervisionados.

1.3 Organização do documento

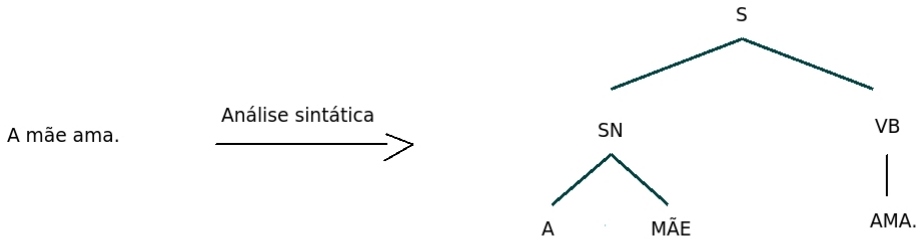
No Capítulo 2 será introduzido o que é análise sintática e como resolver seu problema de ambiguidade, o que são *bancos de árvores* e como induzir gramáticas através de técnicas de aprendizado de máquina. No Capítulo 3 apresentamos os modelos criados para analise sintática não-supervisionada e semi-supervisionada. O Capítulo 4 apresenta os métodos e métricas utilizados para medir os resultados da análise de constituência e dependência, e por fim explicamos o nosso modelo de análise semi-supervisionada e discutimos seus resultados. No ultimo Capítulo 5 apresentamos os trabalhos futuros e concluímos.

2 Fundamentação teórica

A análise sintática ou (*parsing*) consiste em dada uma sentença, extrair a melhor árvore sintática para a mesma. Esta tarefa de extrair a melhor árvore sintática para uma sentença é executável a partir da construção de uma gramática. A gramática é um método para a ordem e arranjo das palavras em uma sentença a partir de uma entrada e um conjunto de regras. Através da construção de uma gramática o analisador verifica as derivações possíveis para uma certa entrada.

A utilização de gramáticas em análise sintática tem como objetivo a extração ou recuperação de estruturas sintáticas. A Figura 1 exemplifica a tarefa de extração da estrutura sintática da frase "A mãe ama", apesar de apresentar bons resultados é necessário constituir muitas regras como é o caso da linguagem humana que é muito complexa e ambígua. Para facilitar a tarefa de análise sintática Manning e Schütze (1999) propõem a abordagem de indução gramatical onde a extração das regras de uma gramática é feita através de técnicas de aprendizado supervisionado, não-supervisionado ou semi-supervisionado em auxílio a um parser. Mas o problema da ambiguidade ainda persiste então estes ainda propõem a utilização de um analisador sintático probabilístico.

Figura 1 – Processo de análise sintática.



Um analisador sintático probabilístico tenta solucionar a ambiguidade de uma gramática atribuindo peso as produções. A ambiguidade induz o analisador a interpretar uma sentença erroneamente, levando assim a sentença ter mais de uma interpretação (derivação) possível, como por exemplo a frase "o menino viu a menina de binóculos", onde a sub-frase de binóculos pode ser associado ao menino ou a menina. Em termos gramaticais teremos mais de uma produção possível para a mesma frase. A ambiguidade em gramáticas gera indecisão comprometendo assim o desempenho de um analisador sin-

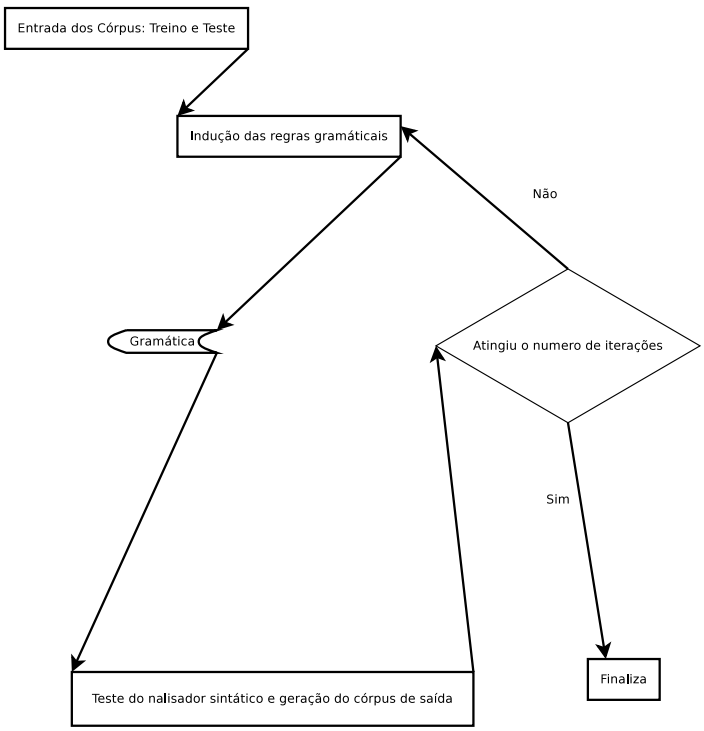
tático. Na gramática abaixo, podemos perceber que a ambiguidade geraria uma indecisão ao analisador sintático onde a produção B gera duas diferentes produções:

$$S \rightarrow exemploB$$
$$B \rightarrow ambiguidade1$$
$$B \rightarrow ambiguidade2$$

2.1 Análise sintática probabilística

A análise sintática probabilística que Manning e Schütze (1999) classificam como uma implementação da técnica de *chunking*, que é reconhecimento de alto nível de estruturas o que possibilita condensar nossa descrição da sentença. Uma forma de capturar a regularidade dos *chunk's* sobre um conjunto de palavras é aprender sua classe gramatical. Este problema em computação denomina-se indução de uma gramática. O problema central é como induzir empiricamente a estrutura ou árvore anotada sintaticamente a partir de uma entrada textual.

Figura 2 – Processo de análise sintática Probabilística.



O analisador sintático também tem como tarefa desambiguar sentenças através da indução de gramáticas e atribuição de pesos diferentes às regras. A gramática abaixo seria um exemplo de uma gramática utilizada em um analisador sintático probabilístico, onde

o analisador selecionaria a sentença com maior probabilidade, neste caso a de $P(s) = 0,8$. Mas para definir-se um analisador sintático probabilístico pode-se simplesmente escrever todas as regras gramaticais possíveis para uma linguagem, mas isso demandaria muito tempo e dificuldade em implementar todos os casos possíveis, então como aponta Manning e Schütze (1999) devemos criar um modelo de aprendizado, onde incrementalmente induzimos as regras gramaticais e seus pesos, através de uma etapa de treinamento e verificamos a sua eficacia na etapa de teste. A Figura 2 representa o funcionamento deste modelo, onde em primeira etapa temos as definições dos conjuntos de teste e treino, após isso o analisador fica encarregado de induzir a gramática através de algum modelo previamente definido durante um certo número de iterações, durante esta etapa também são realizados pequenos testes para verificar a eficiência da indução gramatical e por fim o analisador gera o conjunto de árvores finais.

$$S \rightarrow exemploB$$
$$B \rightarrow \text{ambiguidade1 } [0.8]$$
$$B \rightarrow \text{ambiguidade } [0.2]$$

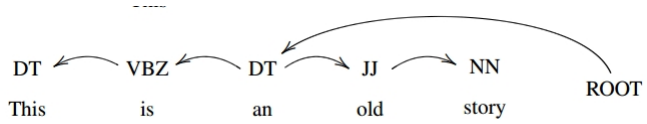
2.2 Indução gramatical

Como já foi dito para induzirmos a classe gramatical das palavras de forma supervisionada, é necessária a utilização de córpus de anotação sintática e isso despende de mão de obra qualificada. Então a utilização de técnicas de aprendizado de máquina não-supervisionado se faz necessário cada vez mais. Para o inglês existem ótimas propostas de analisadores sintáticos e um dos melhores resultados levantados pela bibliografia é a proposta de Klein e Manning (2004), onde ele analisa a estrutura das frases quanto a sua constituência e dependência através da combinação destes modelos, *Constituent-Context Model* CCM e *Dependency Model with Valence* DMV respectivamente, o analisador induz a a gramática. Apesar de o aprendizado de máquina supervisionado ter a necessidade de grandes quantidades de córpus, ele apresenta resultados superiores às demais formas de aprendizado para o processamento de linguagem natural.

De acordo com Manning e Schütze (1999), constituintes são palavras ou grupos de palavras que funcionam unitariamente em uma sentença. Em sua definição mais literária, constituinte é aquele que representa ou constitui um organismo ou uma classe.

Gramáticas dependentes como propõem Manning e Schütze (1999), descrevem uma estrutura linguística em termo de dependências de palavras. A construção da árvore sintática se dá através de uma palavra raiz e todas as outras são suas dependentes desta ou entre elas como mostra a Figura 3.

Figura 3 – A árvore de dependência da sentença *"This is an old story"*.



Fonte: Luque (2011)

2.3 Métodos de aprendizado

Para utilização de analisadores sintáticos probabilísticos, é necessária a inicialização e configuração do modelo de aprendizado de máquina. Segundo (MARSLAND, 2009) aprendizado pode ser categorizado em três termos: adaptação, generalização e recordação. O termo adaptação refere-se a capacidade do modelo de adaptar as mudanças, já os termos generalização e recordação é a capacidade de reconhecer similaridades entre diferentes eventos. Estas três características juntas tornam um modelo de aprendizado útil.

A configuração do modelo é a principal etapa no aprendizado de máquina, esta etapa pode também ser chamada de treinamento, onde é iterativa. Este *aprendizado* pode ser de três formas: supervisionado, não-supervisionado, e semi-supervisionado.

A técnica de aprendizado supervisionado depende de grandes informações prévias para induzir uma gramática automaticamente, dificultando assim o aprendizado para línguas como o português que possuem poucos córpus para treinamento e teste. Outra grande limitação é se aplicarmos como entrada na etapa de treinamento um córpus de um determinando domínio, e aplicarmos na etapa de teste outro com um domínio diferente, teremos baixo desempenho, pois técnicas de supervisão tem limitações para classificar palavras com domínios ou contextos diferentes. Um córpus de anotação sintática muito utilizado é o *Penn Treebank* para o inglês.

Já a técnica de aprendizado não-supervisionado utiliza-se de dados *puros* e busca a indução ou detecção de estruturas. O principal incentivo de utilizar aprendizado não-supervisionado em análise sintática é que existem poucos córpus com anotação sintática disponíveis para o português, tornando a utilização desta técnica muito útil para indução de gramáticas e ela também pode ser combinada com outras técnicas como a da aglomeração [1] e algoritmos de otimização de aprendizado.A técnica de semi-supervisão seria um junção de aprendizado supervisionado com o não supervisionado, em nosso caso queremos induzir as melhores estruturas para uma sentença (não supervisionado), sem utilizarmos as etiquetas de uma árvore, já na técnica de supervisão utilizaria as etiquetas presentes

[1] Do inglês *clustering*

no córpus *Penn Treebank* para induzirmos as estruturas gramáticas das frases. Então se unirmos estas duas técnicas conseguiríamos o melhor dos dois mundos.

2.3.1 Maximizando probabilidades

Em análise sintática como foi dito anteriormente é necessária a definição de um modelo de estruturas, este modelo pode ser treinado de forma não-supervisionada utilizando a técnica de Maximização de Expectativa (ME) proposto por Dempster, Laird e Rubin (1977). Como explica Manning e Schütze (1999) este algoritmo pode ser explicado como uma versão simplificada da técnica de *clustering* onde existem agrupamentos como estimativas de distribuições probabilísticas.

Como nosso foco é outro, explicaremos o (ME) aplicado em Processamento de Linguagem Natural (PLN). Como aponta Manning e Schütze (1999) este algoritmo é iterativo, a cada interação a probabilidade da máxima verossimilhança (ML) continua aumentando até convergir em algum valor, assim é possível estruturas de modelos e seus parâmetros. Abaixo explica-se estes dois passos:

- Passo-E ou **Expectativa**: Calcula expectativas de variáveis binárias, ou probabilidades de adesão de palavras (fragmentos).
- Passo-M ou **Maximização**: Estas probabilidades são recalculadas neste passo como uma estimativa de Máxima Verossimilhança (ML).

O algoritmo inicia no Passo-E e a cada iteração e chamado o Passo-M recebendo como parâmetro as probabilidades calculadas no passo anterior. Nas próximas seções explicaremos como o algoritmo ME é aplicado nos modelos utilizados neste projeto.

2.3.2 Treebanks e estruturas de representação

A Figura 4 mostra uma sentença no formato *Banco de árvores*[2], onde cada etiqueta representa um conjunto de sub-árvores anotadas sintaticamente. Atualmente, de acordo com Manning e Schütze (1999) existe um grande esforço na produção de córpus deste gênero no formato *Penn Treebank* por sua alta utilização na construção de analisadores sintáticos estatísticos tanto supervisionados quanto não-supervisionados. A estrutura para uma representação de análise sintática em níveis, é denominada árvore sintática, como mostra a Figura 1. Nesta figura podemos verificar que as folhas ou nós terminais são as palavras, os nós intermediários, os nós acima dos terminais, são as categorias sintáticas. Se existisse um nível acima ou seja um nó pai de um nó intermediário seria denominado de etiquetas sintáticas ou nós pré-terminais. As etiquetas presentes na tabela, seguem o formato do *Penn Treebank* que estão presentes na Tabela 1 .

Figura 4 – Uma árvore no córpus *Penn Treebank*

```
(IP-MAT (NP-VOC (NPR Senhor))
  (VB Ofereço)
  (PP (P a) (NP (PRO$-F Vossa) (NPR Majestade)))
  (NP-ACC (D-F-P as)
    (NPR-P Reflexões)
    (PP (P sobre)
      (NP (D-F a) (N vaidade) (PP (P de) (NP (D-P os) (N-P homens)))))))
```

Fonte: Galves e Faria (2010)

Tabela 1 – Etiquetas presentes no modelo Penn Treebank.

Etiqueta	**Descrição**
S	Senteça
ADJP	Adjetivo de frase
ADVP	Locução adverbial
NP	Sintagma Nominal
PP	Locução Proposicional
QP	Expressão Quantificada
VP	Verbo
WHNP	Wh-Sintagma Nominal
WHNPP	Wh-Locução Proposicional
CONJP	Frase de muitas palavras com conjunções
FRAG	Fragmento de frase
INTJ	Interjeição/Exclamação
LST	Marcador de lista
NAC	Não é um grupo constituinte
NX	Constituinte Nominal
PRN	Intercalação
PRT	Porção mínima de uma frase
RRC	Cláusula de redução relativa
UCP	Frase de coordenação
X	Não conhecido
WHADJP	Wh-Adjetivo de frase
WHADVP	Wh-Verbo

Fonte: Manning e Schütze (1999)

A literatura apresenta grandes referências a córpus como o WSJ (MARCUS MITCHELL P.; MARCINKIEWICZ, 1994) [3] para o inglês e Tycho Brahe (IEL-UNICAMP; IME-USP, 2010) para o português, ambos córpus se baseiam neste formato para construir suas árvores de sentença.

[2] Do inglês *Penn Treebank*
[3] *Wall Street Jornal*

2.4 Modelos de análise sintática

Na bibliografia levantada o melhor modelo de indução gramatical não supervisionada é o modelo CCM+DMV de Klein e Manning (2004) onde este obtêm ótimos resultados para o inglês. Mas a proposta de Klein e Manning não é a única para análise de constituência, existem outras como Clark (2001), Klein (2005), Smith e Eisner (2004). Clark utiliza a técnica de aglomeração distribucional para agrupar sentenças, mas o simples agrupamento não garante bons resultados, então eles utilizam o critério de informação mútua proposto por anning, Raghavan e Schütze (2008) para filtrar termos constituintes e destituintes. A proposta de Klein e Manning será explicada mais detalhadamente nas próximas seções. Já a proposta de Smith e Eisner (2004) apresenta melhorias em relação ao problema de máximo local do algoritmo de maximização de esperança.

Para análise por dependência, utilizamos neste projeto o modelo de dependência com valência (DMV) de Klein e Manning (2004) mas não é a única existente, também Blunsom e Cohn (2010) e Gillenwater, Pereira e etc. (2010). A proposta de Klein e Manning será explicada mais detalhadamente nas próximas seções. A proposta de Blunsom e Cohn (2010) utiliza uma técnica de *Tree Substitution Grammar* de Cohn e Blunsom (2009) que utiliza o modelo bayesiano de probabilidades em combinação com o DMV de Klein e Manning. Já o trabalho de Gillenwater, Pereira e etc. (2010) ao invés de utilizarem o algoritmo EM como os outros modelos citados anteriormente, este utiliza uma técnica de penalidades em distribuições pai-filho em uma árvore sintática e em pares de *etiquetas* de regularizações posteriores Graça, Ganchev e Taskar (2007), apresentando resultados de medida-F superiores ao DMV.

Todos estes trabalhos apresentam resultados para análise de dependência ou constituência em inglês, mas os trabalho de Gillenwater, Pereira e etc. (2010) apresenta também resultados para o português.

2.4.1 Modelo de constituência

O desempenho do modelo de constituintes, proposto por Klein e Manning (2004), apesar de ocorrer sob um conjunto de fortes restrições, ainda é o melhor que usa aprendizado não supervisionado que supera os modelos de *baseline* de ramificação a direita, que simplesmente agrupa árvores aleatórias a direita. Esse modelo baseado em constituintes[4] é chamado de CCM (*Constituent-Context Model*), e utiliza a técnica de aglomeração (*clustering*) para induzir a classe gramatical das palavras. A ideia é resolver o problema através da construção de decisões sobre constituência diretamente no modelo probabilístico, e também agrupando termos destituintes (não constituintes).

[4] De acordo com (MANNING; SCHÜTZE, 1999), constituintes são palavras ou grupos de palavras que funcionam unitariamente em uma sentença. Em sua definição mais literária, constituinte é aquele que representa ou constitui um organismo ou uma classe.

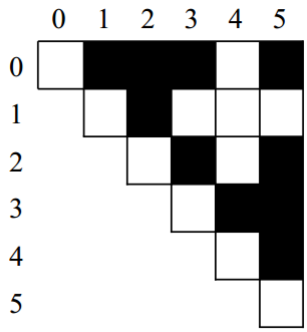
Para isso o modelo calcula a probabilidade conjunta de uma sentença S e uma parentetização B sobre S, que pode ser obtida pelo produto da probabilidade marginal de B e da probabilidade de S dado B, ou seja, $P(S, B) = P(B)P(S|B)$. A probabilidade conjunta $P(S|B)$ é obtida por:

$$P(S|B) = \prod_{i,j} P(\alpha_{i,j}|B_{i,j})P(x_{i,j}|B_{i,j}).$$

Dada uma constituência $B_{i,j}$, a distribuição $P(\alpha_{i,j}|B_{i,j})$ é uma probabilidade condicional das distribuições sobre uma produção, para constituintes e destituintes. Já $P(x_{i,j}|B_{i,j})$ é uma probabilidade condicional sobre o contexto dos termos de i a j em S. Essas duas probabilidades condicionais são usadas no algoritmo de Maximização de Esperança para aumentar a máxima verossimilhança do modelo. Para a probabilidade marginal, Klein e Manning (2002) afirmam que pode-se utilizar uma distribuição $P_{bin}(B)$, que seleciona e procura parentetizações uniformes binárias. Entretanto, eles não utilizam essa distribuição, pois ela tende a atribuir maior peso a estruturas desbalanceadas do que balanceadas, o que não se deseja no modelo. O modelo mais desejável de seleção seria um que atribuísse peso razoável a estruturas desbalanceadas, mas não de forma tendenciosa. Assim, os autores utilizam uma distribuição empírica que não é descrita.

A Figura 5 representa uma tabela aleatória, onde representa a probabilidade $P(B)$ para uma sentença. Os elementos em escuro representam se um determinado span $\alpha_{i,j}$ é um constituinte ou não isto é se existe um parêntesis associado aos índices i, j, i é a vertical já j representa a horizontal.

Figura 5 – Tabela de um modelo de parentetização no modelo CCM



2.4.2 Modelo de dependência com valência

O modelo proposto por (KLEIN; MANNING, 2004) calcula um processo generativo de estruturas de dependências, onde as palavras de um sentença criam dependência entre si em ambas as direções (esquerda, direita).

Em sua proposta original existem duas variantes, o modelo depende de utilizar ou não a restrição *one-side-first*. A segunda proposta não é muito bem explicada por Klein e Manning em nenhum de seus trabalhos. Então na primeira variante do modelo escolhe-se o lado que se deseja gerar dependências baseado na probabilidade para cada palavra $w \in S$, a distribuição $P_{order}(dir|s)$ onde $dir \in l, r$ define qual direção será tomada. Mas deve-se tomar uma decisão de quando parar que é calculada com base na distribuição $P_{stop}(true|w, dir, adj)$, onde adj é um simbolo *booleano* que indica se já gerou algum termo dependente. Se a distribuição P_{stop} indica que não se deve parar então se gera um dependente dep, que se obtêm através da distribuição $P_{choose}(dep|w, dir)$. Quando se decide parar w gera dependentes em outra direção.

Utilizamos a sequência D(w,dir) para gerar todas as dependências de w para alguma dir possível, a probabilidade de algum fragmento ou sequência D(w) é dada pela fórmula:

$$P(D(w)) = P_{order}(o|w) \prod_{dir \in (l,r)} \left(\prod_{dep \in deps_{D(w,dir)}} P_{stop}(false|w, dir, adj) P_{choose}(dep|w, dir) \right) P_{stop}(true|w, dir, adj)$$

A distribuição $P(D)$ de uma árvore de dependências D sobre uma sentença S é dada abaixo pela fórmula:

$$P(D) = P(D(Raiz)) \prod_{w \in S} P(D(h))$$

Klein e Manning utilizam também o algoritmo de Maximização de Esperança para maximizar probabilidades, mas como estrada este algoritmo calcula a probabilidade de uma dada etiqueta, um nó não terminal de uma árvore sintática, gere uma nova sentença com as dependências obtidas do modelo maximizando a máxima verossimilhança.

2.4.3 Modelo CCM+DMV

Os dois modelos podem ser combinados, já que ambos os modelos podem ser transformados no problema de árvores lexicalizada Manning e Schütze (1999). Com a transformação das árvores em árvores lexicalizadas podemos utilizar algoritmos de clusterização, otimizando assim os resultados de Medida-F.

3 Modelos Implementados

Neste capítulo apresentamos quais os modelos que desenvolvemos, então ele está estruturado em duas partes, onde primeiro apresentamos o modelo CCM+DMV onde contribuímos com resultados para o português, e na segunda parte apresentamos o nosso modelo onde induzimos estruturas sintáticas através da semi-supervisão.

3.1 Análise sintática não supervisionada

O analisador sintático não-supervisionado utiliza-se da indução de estruturas gramaticais, para definir qual a melhor representação possível para uma entrada.

A proposta de Klein e Manning (2004), que utiliza como modelos de indução gramatical o CCM+DMV foi utilizada como base para medirmos e compararmos seu desempenho para o português e inglês. O modelo CCM induz gramáticas através da constituência das palavras e o modelo DMV induz gramáticas através da dependência gramatical das palavras. Se unirmos os modelos teremos um modelo conjunto que induz gramáticas através da análise dos termos constituintes e dependentes de uma sentença. Este modelo inicialmente não apresentava resultados para o português, então implementamos e realizamos testes deste modelo para o português, com base na versão disponibilizada por Luque (2011).

3.2 Análise sintática semi-supervisionada

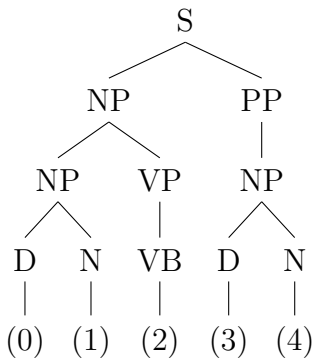
O analisador sintático que desenvolvemos (CCM+supervisão), consiste no modelo de constituência de Klein e Manning (2004) mas com adição de supervisão.

3.2.1 Modelo CCM+supervisão

O modelo propõe a utilização de frações de supervisão na etapa de indução da gramática do modelo CCM. A adição da supervisão tem como objetivo aumentar os acertos das medidas de precisão e cobertura no conjunto de árvores de teste. A edição desta supervisão implica também que o modelo passara à utilizar etiquetas sintáticas na etapa de indução, passando a calcular para cada $P(S|B)$ não somente a multiplicação de cada $P(\alpha_{i,j}|B_{i,j})$ por $P(x_{i,j}|B_{i,j})$ mas sim a melhor etiqueta para cada $split_{i,j}$[1], sendo ele constituinte ou não. Um exemplo de split pode ser observado na árvore abaixo onde a

[1] O split que está representado no modelo CCM como $\alpha_{i,j}$ é uma quebra na sentença.

etiqueta sintática NP, está dentro do split (3,4). Já o $contexto_{i,j}$[2], que pode ser observado na árvore para o $contexto_{2,4}$ (tendo como etiquetas morfossintáticas: VB,D,N), e este contexto é considerado no modelo como sendo um vizinho do split (2,4). No nosso modelo esta etiqueta será denominada de et.



A fórmula correta do nosso modelo para cacularmos a distribuição $P(S|B)$ é

$$P(S|B) = \prod_{i,j} P(\alpha_{i,j,et}|B_{i,j})P(x_{i,j,et}|B_{i,j}).$$

Poderiamos reescrever a fórmula da forma

$$P(S|B) = \prod_{i,j} \phi_{i,j,s,et}.$$

onde

$$\phi_{i,j,s,et} = \frac{P(\alpha_{i,j,s}|et)P(x_{i,j,s}|et)}{P(\alpha_{i,j,s}|false)P(x_{i,j,s}|false)}$$

Então o nosso modelo retorna a melhor etiqueta sintática, se existir, para cada constituinte, possibilitando ao nosso modelo também expor resultados para etiquetagem. Pois em contra partida ao modelo CCM o nosso modelo apresenta resultados etiquetados e os melhores constituintes para cada sentença S. A adição de etiquetas na etapa de indução gramatical tornaria o nosso modelo em um modelo semi-supervisionado.

[2] O contexto está representado no modelo CCM como $x_{i,j}$ e representa quais etiquetas morfossintáticas, o split tem como vizinho

4 Experimentos e resultados

Este capítulo tem como objetivo, apresentar os resultados que obtivemos para ambos os modelos CCM+DMV e CCM+supervisão. Primeiro apresentamos quais os métodos e métricas utilizados, após isso apresentamos os resultados dos modelos desenvolvidos.

4.1 Métodos e métricas

Uma forma de medir a quantidade de acertos em analisadores sintáticos não supervisionados e supervisionados é através de medidas de cobertura, precisão e medida-F não rotuladas e rotuladas. A medição é feita através de um córpus de teste. Abaixo descrevemos como são calculadas as medidas:

$$Precisão = \frac{Nós\ corretos\ da\ árvore\ sintática\ gerada}{Numero\ de\ nós\ totais\ da\ árvore\ sintática\ gerada}$$

$$Cobertura = \frac{Nós\ corretos\ da\ árvore\ sintática\ gerada}{Numero\ de\ nós\ totais\ da\ árvore\ sintática\ original}$$

$$Medida - F = 2 \cdot \frac{Precisão \cdot Cebertura}{Precisão + Cobertura}$$

As medidas de precisão, cobertura e medida-F são obtidas através da média dessas medidas para cada frase. Em sua proposta original Klein e Manning (2004) explicam que os modelos CCM e DMV, utilizam apenas árvores binárias para avaliação de estruturas, o modelo UBound calcula a eficiência máxima que pode ser alcançada utilizando árvores binárias. Outro modelo muito utilizado em sistemas não supervisionados é o linha base[1] que permite avaliar o desempenho de um sistema este é um modelo de geração aleatória de resultados básicos. No caso da análise sintática ocorre a geração de árvores para um sentença, mas este modelo somente gera um limiar base que permite compararmos o mínimo de acertos que um analisador sintático deve atingir.

Para medirmos o desempenho do modelo semi-supervisionado utilizamos as mesmas métodos e métricas utilizadas pelos outros modelos, mas geramos um resultado diferente que é a assertividade das etiquetas morfossintáticas. Estas etiquetas são obtidas através do nosso modelo, onde ele encontra a melhor etiqueta, ou a etiqueta com maior probabilidade que esteja dentro de um constituinte gerado.

[1] Do inglês *Baseline*

4.2 Analise sintática Não-supervisionada

Esta seção tem como objetivo demonstrar os resultados obtidos para analise sintática não supervisionada, onde apresentamos quais foram os métodos e métricas utilizados para obtenção das medidas.

4.2.1 Experimentos e Resultados

Fizemos a adaptação dos modelo CCM e DMV de uma versão desenvolvida por Luque (2011), que criou uma adaptação totalmente desenvolvida em Python, utilizando a biblioteca NLTK[2]. Esta adaptação foi feita para analisarmos e testarmos estes modelos no português para córpus binários. Como Luque (2011) não criou classes de leitura para os córpus de anotação sintática Bosque e Tycho Beahe, necessitava a alteração de seu projeto inicial para que também contemplasse estes córpus, e os resultados destas alterações serão discutidos ainda nesta seção.

Pela análise da Tabela 2 percebe-se que Luque obteve resultados muito próximos de Klein e Manning (2004) para os modelos de constituência e dependência aplicados ao córpus WSJ com limitante de até dez palavras por sentença (WSJ10). Para esta análise comparativa Luque treinou o seu modelo com 40 iterações (neste ponto de acordo com Klein e Manning (2004) os modelos convergem).

Tabela 2 – Resultados do modelo CCM+DMV.

Modelo	**Precisão**	**Cobertura**	**Medida-F**	**Córpus**
CCM-KM	64.2	81.6	71.9	WSJ10
DMV-KM	46.6	59.2	52.1	WSJ10
CCM+DMV-KM	69.3	88.0	77.6	WSJ10
CCM-L	64.3	81.6	71.9	WSJ10
DMV-L	58.3	74.1	65.3	WSJ10
CCM+DMV-L	67.9	86.2	75.9	WSJ10
LBranch	25.6	32.6	28.7	WSJ10
RBranch	55.1	70.0	61.7	WSJ10

Fonte: Luque (2011)

Os modelos CCM-KM, DMV-KM e CCM+DMV-KM são de Klein e Manning (2002) e os CCM-L, DMV-L e CCM+DMV-L são de Luque (2011). Os modelos LBranch e RBranch são modelos triviais de *baseline*. O LBranch[3] e RBranch[4] apenas criam árvores com ramificações a esquerda e direita, respectivamente.

[2] <http://nltk.org/>
[3] Do inglês *Left Branch*
[4] Do inglês *Rigth Branch*

Pela análise da Tabela 3 percebemos que os modelos CCM-L e DMV-L, aplicados ao português, apresentam baixos resultados em comparação com o inglês. Esse baixo desempenho deve-se que a quantidade de árvores para treinamento é menor que a do inglês, foram utilizadas 3938 para o córpus Tycho Brahe, já para o inglês foram 7422 árvores. Ambos os córpus Tycho Brahe e WSJ estão em formato binário, isto suas sentenças apresentam somente dois filhos por pai. O modelo proposto ficou muito próximo do UBound para o córpus Tycho Brahe não segmentado.

Tabela 3 – Comparação de resultados entre português e inglês.

Modelo	**Cobertura**	**Precisão**	**Medida-F**	**Córpus**
CCM-KM	81,60	64,20	71,90	WSJ10
DMV-KM	59,20	46,60	52,10	WSJ10
CCM+DMV-KM	88,00	69,30	77,60	WSJ10
CCM-L	64,20	41,60	50,50	Tycho Brahe10
DMV-L	58,80	38,10	46,20	Tycho Brahe10
CCM+DMV-L	68,80	48,10	60,10	Tycho Brahe10
UBound	52,82	60,85	60,40	Tycho Brahe10

4.2.2 Conclusões dos resultados

Atualmente existem limitações em questão de córpus anotados sintaticamente para o português, apesar do córpus Tycho Brahe[5] estar muito bem constituído. A baixa capacidade dos analisadores sintáticos supervisionados em atender diversos domínios de textos aumentam a necessidade de se investir em técnicas de análise sintática não supervisionada.

Os melhores modelos não supervisionados, levantados pela bibliografia para o inglês obtém baixo desempenho para o português. Uma das explicações para isso pode ser o menor número de sentenças disponível em comparação com o inglês. Outra explicação, provavelmente complementar, é sintaxe mais flexível, e portanto mais ambígua, do português.

Percebe-se pela análise da Tabela 3 e dos trabalhos levantados pela análise bibliográfica que seria interessante juntarmos a técnica de supervisão, passando a analisar também as etiquetas presentes na estruturas das árvores sintáticas, e não somente estados pré-terminais como o aprendizado não supervisionado opera, criando assim um modelo hibrido como propõem Wang, Schuurmans e Lin (2008), aplicando técnica de supervisão em conjunto com à não-supervisão em análise por constituência.

[5] <http://www.tycho.iel.unicamp.br>

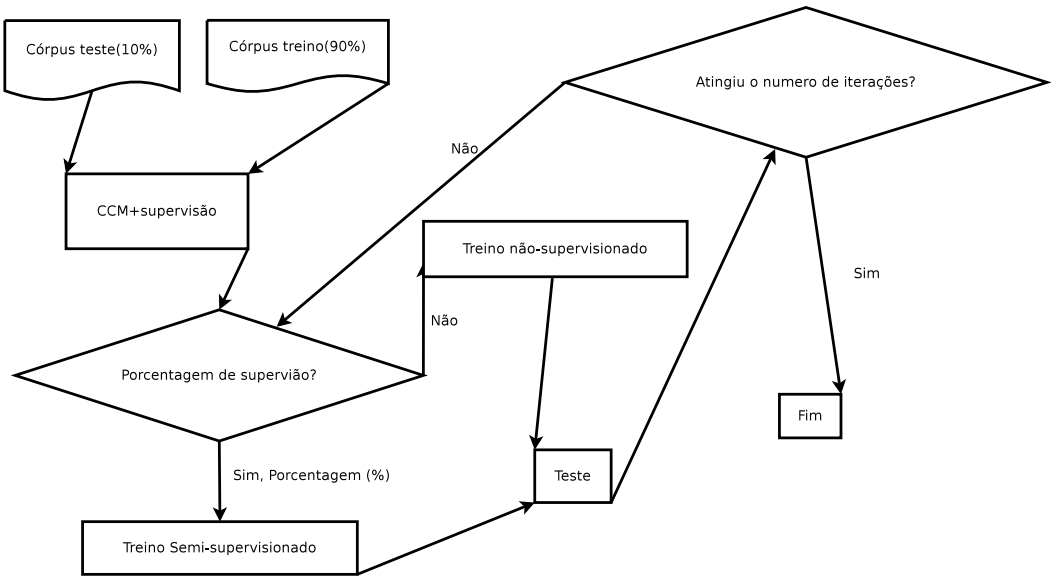
4.3 Analise sintática Semi-supervisionada

Esta seção tem como objetivo demonstrar os resultados obtidos para analise sintática semi-supervisionada, onde apresentamos quais foram os métodos e métricas utilizados para obtenção das medidas. Para os testes em inglês utilizamos o córpus WSJ10 com 7422 sentenças com até 10 palavras por sentença, e para o português utilizamos o córpus Tycho Brahe com 3938 também com sentenças de tamanho até 10. Ambos córpus são dados como entrada de forma binária.

4.3.1 Experimentos e Resultados

As etapas do modelo CCM+supervisão podem ser observadas na Figura 6, onde de forma iterativa realizamos o treino em porcentagens de supervisão[6], esta supervisão é reparametrizada a cada chegada no estado Fim.

Figura 6 – Processo do modelo CCM+supervisão



Para realizarmos os experimentos no modelo CCM+supervisão, dividimos os córpus em teste e treinamento. Utilizamos o método *Ten-cross validation* para realizarmos a divisão dos córpus, onde este consiste em dividir uma distribuição de dados em fatias de 10% de teste e 90% de treino, e de forma iterativa gera 10 córpus diferentes de teste e treinamento. O *Ten-cross* foi utilizado nos córpus Tycho Brahe para testes do português e WSJ para o inglês.

Como o CCM+supervisão gera resultados etiquetados e não etiquetados, teremos que avaliar os córpus gerados de duas maneiras, que é quanto ao acerto das etiquetas

[6] Dividimos os percentis entre intervalos de 10%, o primeiro percentil é 0%.

chutadas[7] e ao acerto dos constituintes (não etiquetado), para medirmos os córpus não-rotulados necessitamos ignorar as etiquetas morfossintáticas e sintáticas. Para ambos testes utilizamos o programa EVAL-B[8] que realiza testes para córpus etiquetados e não etiquetados.

Esta seção será divida em resultados etiquetados, onde serão analisados e comentados seus resultados, e os resultados não-etiquetados. Par fins de ordem começaremos expondo os resultados não-etiquetados.

4.3.1.1 Resultados não etiquetados

O objetivo desta seção é expor resultados não etiquetados, abaixo discutiremos estes resultados. Os teste foram organizados seguindo a métrica *Ten Cross validation*, então apresentaremos a média e o desvio padrão da média das execuções sobre um conjunto de vinte córpus diferentes, estes córpus são para o português e inglês. Para realizarmos os teste para o português, utilizamos o córpus Tycho Brahe onde este foi dividido em dez córpus diferentes, e dividimos cada execução em dez diferentes parcelas de supervisão.

Tabela 4 – Média Geral não etiquetado do Córpus Tycho Brahe

Modelo	**Cobertura**	**Precisão**	**Medida-F**
CCM+supervisão	58,25	71,85	64,32
UBound	96,56	72,04	81,62

Pela análise da Tabela 4 percebemos que a supervisão deixou nosso modelo pouco abaixo em relação a medida-F do Ubound entorno de $17,30\%$, apesar de nosso modelo somente selecionar a maior probabilidade associada a uma etiqueta, esta supervisão somente leva em consideração etiquetas sintáticas associadas a um contexto e um split. Como podemos perceber pela Tabela 5 que existe um desvio padrão da medida-F relativamente alto em torno de $7,19$ isso, deve-se ao fato de córpus de teste apresentar frases escritas com um entendimento fácil e outras de modo difícil, e isso acarretou está variação em nossos testes, e também deve-se ao fato de nosso modelo de segmentação dos córpus ser o *Ten-cross validation* e ele dividir o córpus original em 10 outros de forma incremental sem levar em conta o grau de dificuldades das frases.

A Figura 7 demonstra o crescimento da medida-F a medida que aumentamos a porcentagem da supervisão, mas também percebe-se que quando sobrecarregamos o nosso modelo com supervisão ele tende a cair.

[7] O modelo somente etiqueta sintaticamente se o split que estamos avaliando é um constituinte válido, se for é selecionado a maior probabilidade e a sua etiqueta.

[8] a configuração é feita de forma parametrizável, onde parametrizamos o córpus *gold*, ou o córpus de teste(com etiquetas corretas), e o córpus gerado pelo modelo.

Figura 7 – Medida-F CCM+supervisão não etiquetado Tycho Brahe.

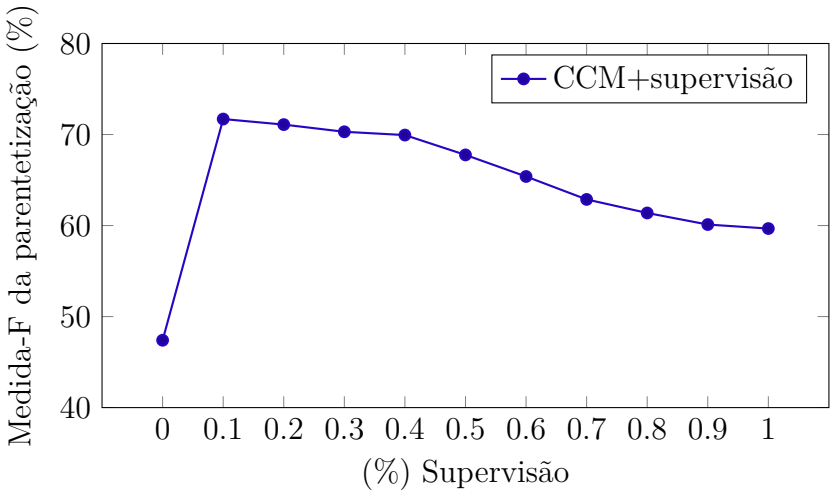


Tabela 5 – Média para cada iteração do Córpus Tycho Brahe não etiquetado

Supervisão (%)	**Cobertura**	**Precisão**	**Medida-F**
0,0	42,91	52,95	47,40
0,1	64,91	80,11	71,70
0,2	64,36	79,42	71,09
0,3	63,65	78,53	70,30
0,4	63,32	78,13	69,94
0,5	61,36	75,68	67,76
0,6	59,23	73,01	65,39
0,7	56,95	70,19	62,87
0,8	55,59	68,55	61,38
0,9	54,43	67,15	60,11
1,0	54,03	66,66	59,67
Média Geral:	58,25	71,85	64,32
Desvio Padrão:	6,51	8,03	7,19

Para o inglês também utilizamos a técnica *Ten-cross validation* para segmentação de dados dos córpus, mas o córpus utilizado foi o WSJ.

Tabela 6 – Média Geral não etiquetado do Córpus WSJ

Modelo	**Cobertura**	**Precisão**	**Medida-F**
CCM+supervisão	53,89	66,85	59,67
UBound	95,92	82,87	88,92

Pela análise da Tabela 6 verificamos que o nosso modelo para inglês foi inferior também ao Ubound entorno de $28,95\%$. Essa baixa deve-se ao fato de o córpus WSJ apresentar uma quantidade muito grande de textos e isso estatisticamente mostra que o

nosso modelo perde desempenho para córpus muito grandes. Já pela análise da Tabela 7 percebe-se que o desvio padrão entre as diferentes amostras dos córpus é alto e isso denota que a segmentação dos córpus de teste e treino não equilibrou os córpus em relação a dificuldade da escrita dos textos.

Figura 8 – Medida-F CCM+supervisão não etiquetado WSJ.

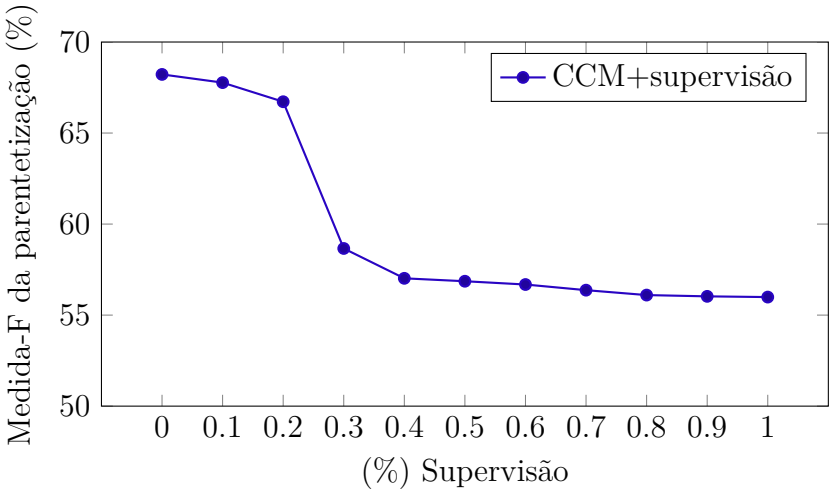


Tabela 7 – Média para cada iteração do Córpus WSJ não etiquetado

Supervisão (%)	**Cobertura**	**Precisão**	**Medida-F**
0,0	61,61	76,43	68,22
0,1	61,20	75,92	67,77
0,2	60,25	74,74	66,72
0,3	52,97	65,71	58,66
0,4	51,49	63,88	57,02
0,5	51,35	63,70	56,86
0,6	51,19	63,50	56,68
0,7	50,91	63,15	56,37
0,8	50,67	62,85	56,10
0,9	50,60	62,77	56,03
1,0	50,57	62,73	55,99
Média Geral:	53,89	66,85	59,67
Desvio Padrão:	4,63	5,75	5,13

Como pode-se perceber o nosso modelo atingiu resultados muito próximos do UBound, e pela análise das tabelas (5 e 8), quanto maior for o seu córpus de treinamento o desempenho do nosso modelo tente a cair e também a complexidade dos textos nos córpus de teste. Isso levaria a uma limitação do nosso modelo, que ficaria limitado a córpus pequenos e com textos simples do ponto de vista da dificuldade do vocabulário.

Para tentarmos reduzir esta queda poderíamos adaptar o nosso modelo para não somente selecionar a etiqueta com maior probabilidade. Gerando assim uma nova mo-

delagem que criaria uma nova dimensão em nosso modelo passando agora a selecionar a melhor etiqueta dado que você tem um conjunto da possível parentetização e das possíveis etiquetas.

4.3.1.2 Resultados Etiquetados

O objetivo desta seção é expor resultados etiquetados, o nosso modelo também tem como resultado córpus etiquetados sintaticamente, abaixo explicaremos estes resultados. Os teste foram organizados seguindo a métrica *Ten-Cross validation*, então apresentaremos a média e o desvio padrão da média das execuções sobre um conjunto de vinte córpus diferentes, estes córpus são para português e inglês, e são os mesmos utilizados para os resultados do não etiquetado.

Para realizarmos os teste do português utilizamos o córpus Tycho Brahe onde este foi dividido em dez córpus diferentes, e dividimos cada execução em dez diferentes parcelas de supervisão igualmente aos testes anteriores.

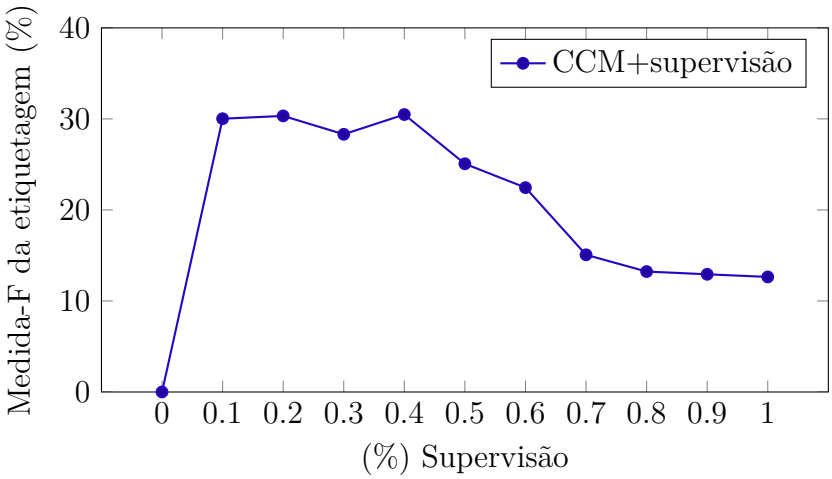
Tabela 8 – Média Geral do Córpus Tycho Brahe etiquetado

Supervisão (%)	**Cobertura**	**Precisão**	**Medida-F**
0,0	0,00	0,00	0,00
0,1	0,30	30,02	0,59
0,2	0,30	30,33	0,59
0,3	0,35	28,31	0,68
0,4	0,45	30,48	0,86
0,5	0,47	25,07	0,90
0,6	0,65	22,45	1,23
0,7	0,84	15,07	1,56
0,8	0,97	13,23	1,79
0,9	1,01	12,93	1,86
1,0	1,04	12,64	1,92
UBound	96,77	69,59	80,95
Média Geral:	0,53	20,79	1,01
Desvio Padrão:	0,33	10,17	0,60

A Tabela 8 apresenta resultados sobre o resultados etiquetados sintaticamente do nosso modelo, estes resultados expõem que o nosso modelo em média seleciona as etiquetas sintáticas com maior probabilidade em $20,79\%$ (precisão) das frases de cada córpus, e dentre as que são selecionadas ele acerta $1,01\%$, diferente mente do que o não etiquetado, a variação da dificuldade do vocabulário não tem tanta relevância quando se trata em resultados etiquetados como aponta o desvio padrão.

Pela análise do Figura 9 percebeu-se que a medida que aumentamos a taxa de supervisão o nosso modelo baixa seus resultados, levando assim a apresentarmos menos etiquetas nos córpus resultantes.

Figura 9 – Medida-F CCM+supervisão etiquetado Tycho Brahe.



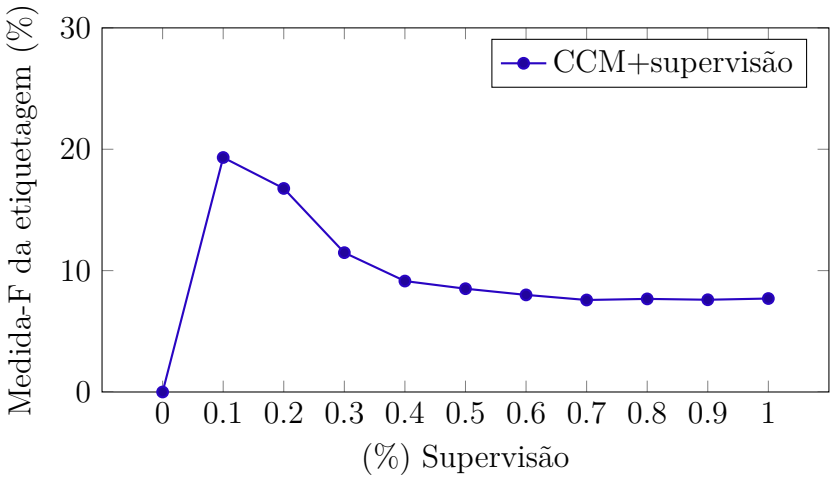
Pela analise da Tabela 9 percebemos que o nosso modelo para o inglês em média seleciona as etiquetas sintáticas com maior probabilidade em $9,61\%$ das frases de cada córpus, e obteve um acerto de $1,71\%$ das etiquetas possíveis. Podemos perceber pelo desvio padrão que também para o inglês a seleção da dificuldade do vocabulário no método *Ten-cross validation* não tem tanto peso e o tamanho de seu córpus que é superior a Tycho Brahe também não influência.

Tabela 9 – Média Geral do Córpus WSJ etiquetado

Supervisão (%)	**Cobertura**	**Precisão**	**Medida-F**
0,0	0,00	0,00	0,00
0,1	1,25	19,32	2,34
0,2	1,17	16,77	2,19
0,3	1,23	11,48	2,21
0,4	1,07	9,14	1,92
0,5	1,00	8,51	1,79
0,6	0,95	8,00	1,69
0,7	0,92	7,58	1,63
0,8	0,95	7,67	1,69
0,9	0,94	7,60	1,68
1,0	0,93	7,70	1,70
UBound	96,77	76,59	85,95
Média Geral:	0,95	9,61	1,71
Desvio Padrão:	0,36	5,35	0,66

Pela análise das tabelas (9 e 8) podemos que o nosso modelo apresenta dificuldades para etiquetar completamente um córpus, ou "chutar"mais etiquetas para os córpus de teste, isso ocorre porque o nosso modelo se prende a etiquetar sintaticamente uma sentença se este for um possível *span* e se existe uma etiqueta já modelada associada a este *span* ou contexto.

Figura 10 – Medida-F CCM+supervisão etiquetado WSJ.



Para o nosso modelo do ponto de vista da etiquetagem sintática, como apontam nossos resultados não faz diferença selecionarmos córpus de treino/teste mais complexos, mas como nos resultados não-etiquetados o nosso modelo é muito dependente do tamanho do córpus, então o córpus como Tycho Brahe terá resultados mais altos em comparação ao córpus WSJ que são mais extensos.

5 Conclusão e Trabalhos Futuros

A ambiguidade é uma tarefa muito importante de um analisador sintático probabilístico, onde através de métodos de aprendizado de máquina tenta induzir regras e e definir peso a estas regras. Para o analisador definir corretamente estas regras Klein e Manning (2004) propõem um modelo que induz estas regras através da constituência das palavras, mas o seu modelo não leva em consideração as etiquetas sintáticas, em contra partida nós propomos a utilização das etiquetas sintáticas em nossa modelagem, para a definição de quais os melhores constituintes e suas etiquetas possíveis de uma frase.

Como resultados para o português sem levarmos em consideração etiquetas, que era um de nossos objetivos principais, em média obtivemos uma diferença de $17,30\%$ em relação ao UBound binário que é o máximo que um analisador sintático probabilístico pode alcançar em um determinado córpus, já para o inglês obtivemos uma diferença em média de $28,95\%$.

Para a etiquetagem sintática nosso modelo obteve resultados de $20,79\%$ (precisão) das frases de cada córpus do português, e dentre as que ele preenche, obtém um acerto de $1,01\%$. Para etiquetagem aplicada ao inglês obtivemos $9,61\%$ das frases de cada córpus foram etiquetadas, e deste obtivemos um acerto de $1,71\%$ das etiquetas certas.

Como conclusão podemos verificar que o nosso modelo para o português se aproxima muito do UBound binário, o nosso modelo com pequenas porcentagens de supervisão atingiu resultados muito próximos ao UBound, mas após isso nosso modelo converge e apresentar um aqueda. Mas em contra partida a esse bom desempenho para o português, temos que o nosso modelo não apresenta bom desempenho com um córpus de maior tamanho que o Tycho Brahe, mas para córpus de tamanho médio/pequeno podemos obter um bom desempenho.

O baixo desempenho na etiquetagem dos córpus nos leva a concluir que podemos melhorar o nosso modelo, no sentido de modelarmos a etiqueta morfossintática, isto é incluí-la como uma nova dimensão em nosso modelo passando assim a escolhermos melhor uma etiqueta, não somente selecionarmos a com maior probabilidade como é feito atualmente, a baixa cobertura da-se pela "fraca"seleção das etiquetas, já se conseguíssemos aumentar a cobertura conseguiríamos aumentar os nós gerados nas árvores sintáticas do analisador, similarmente ao algoritmo CKY que combina regras gramaticais presentes em níveis inferiores gerando assim regras diferentes e com a probabilidade combinada. Definindo assim qual a melhor etiqueta, em contra partida poderíamos definir para serem gerados somente etiquetas dos *brackets* possíveis.

Então, as principais contribuições deste trabalho são que o nosso analisador sin-

tático semi-supervisionado apresenta bons resultados para o português que foi um dos principais objetivos deste trabalho, também pelo nosso levantamento bibliográfico notou-se que nossa proposta foi a única para o português. Já para o inglês apresentamos uma baixa nos resultados em comparação ao UBound, e também percebemos que a medida que temos maiores córpus de treinamento e teste o nosso modelo tende a apresentar baixos resultados.

Como trabalhos futuros poderíamos implementar um modelo de analise sintática semi-supervisionada que modele a etiqueta como uma nova dimensão, talvez aumentando assim o nosso acerto das etiquetas sintáticas e gerando melhores resultados para etiquetagem. Também poderíamos verificar diferentes contextos e suas etiquetas que estão presentes nas frases e não somente o contexto mais próximo à o *span* (esquerda, direita), aumentando assim as possibilidades de encontrarmos a etiqueta mais provável.

Algo interessante a ser feito seria propagarmos as modificações aos modelos de dependência (DMV), passando a verificar a etiqueta, e a o modelo combinado de constituência e dependência (CCM+DMV) ambos de Klein e Manning (2004).

Referências

ANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. Introduc- tion to information retrieval. In: *Cambridge University Press*. [S.l.: s.n.], 2008. Citado na página 23.

BLUNSOM, P.; COHN, T. Unsupervised induction of tree substitution grammars for dependency parsing. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2010. Citado na página 23.

CLARK, A. Unsupervised induction of stochastic context-free gram- mars using distributional clustering. In: *ConLL '01: Proceedings of the 2001 workshop on Computational Natural Language Learning*. [S.l.: s.n.], 2001. Citado na página 23.

COHN, S. G. T.; BLUNSOM, P. Inducing compact but accurate tree-substitution grammars. In: *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. [S.l.: s.n.], 2009. Citado na página 23.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. In: *Journal of the Royal Statistical Society*. [S.l.: s.n.], 1977. Citado 2 vezes nas páginas 15 e 21.

GALVES, C.; FARIA, P. *Tycho Brahe Parsed Corpus of Historical Portuguese*. 2010. Http://www.tycho.iel.unicamp.br/ tycho/corpus/en/index.html. Citado na página 22.

GILLENWATER, J.; PEREIRA, F.; ETC. Sparsity in dependency grammar induction. In: *ACLShort '10 Proceedings of the ACL 2010 Conference Short Papers*. [S.l.: s.n.], 2010. Citado na página 23.

GRAçA, J.; GANCHEV, K.; TASKAR, B. Expec- tation maximization and posterior constraints. In: *Proc. NIPS*. [S.l.: s.n.], 2007. Citado na página 23.

IEL-UNICAMP; IME-USP. Córpus histórico do português anotado tycho brahe. url: http://www.tycho.iel.unicamp.br/. In: *IEL-UNICAMP e IME-USP*. [S.l.: s.n.], 2010. Citado na página 22.

KLEIN, D. The unsupervised learning of natural language structure. In: *Tese de doutorado, Stanford University*. [S.l.: s.n.], 2005. Citado na página 23.

KLEIN, D.; MANNING, C. D. A generative constituent-context model for improved grammar induction. In: *Association for Computational Linguistics*. [S.l.: s.n.], 2002. Citado 2 vezes nas páginas 24 e 30.

KLEIN, D.; MANNING, C. D. Corpus-based induction of syntactic structure: Models of dependency and constituency. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. [S.l.: s.n.], 2004. Citado 10 vezes nas páginas 5, 7, 19, 23, 24, 27, 29, 30, 39 e 40.

LUQUE, F. M. Una implementacion del modelo dmv+ccm para parsing no supervisado. In: *2do Workshop Argentino en Procesamiento de Lenguaje Natural*. [S.l.: s.n.], 2011. Citado 3 vezes nas páginas 20, 27 e 30.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. 1th. ed. [S.l.]: The MIT Press, 1999. Citado 7 vezes nas páginas 17, 18, 19, 21, 22, 23 e 25.

MARCUS MITCHELL P., B. S.; MARCINKIEWICZ, M. A. Building a large annotated corpus of english: the penn treebank. In: *Association for Computational Linguistics 19.2, pp. 313–330*. [S.l.: s.n.], 1994. Citado na página 22.

MARSLAND, S. *Machine Learning: An Algorithmic Perspective*. 1th. ed. [S.l.]: CRC Press, 2009. Citado na página 20.

SMITH, N. A.; EISNER, J. Annealing techniques for unsupervised sta- tistical language learning. In: *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. [S.l.: s.n.], 2004. Citado na página 23.

WANG, Q. I.; SCHUURMANS, D.; LIN, D. Semi-supervised convex training for dependency parsing. In: *Proceedings of ACL-08: HLT*. [S.l.: s.n.], 2008. Citado 2 vezes nas páginas 15 e 31.