ANOTAÇÃO DE PAPÉIS SEMÂNTICOS PARA O PORTUGUÊS POR *CONDITIONAL RANDOM FIELDS*

Luan Barbosa Garrido

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Geraldo Bonorino Xexéo

Rio de Janeiro
Março de 2017

ANOTAÇÃO DE PAPÉIS SEMÂNTICOS PARA O PORTUGUÊS POR *CONDITIONAL RANDOM FIELDS*

Luan Barbosa Garrido

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Geraldo Bonorino Xexéo, Ph.D.

Prof. Jano Moreira de Souza, Ph.D.

Prof. Giseli Rabello Lopes, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2017

Garrido, Luan Barbosa

Anotação de Papéis Semânticos para o Português por *Conditional Random Fields*/Luan Barbosa Garrido. – Rio de Janeiro: UFRJ/COPPE, 2017.

XIV, 101 p.: il.; 29, 7cm.

Orientador: Geraldo Bonorino Xexéo

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2017.

Referências Bibliográficas: p. 76 – 87.

1. papéis semânticos. 2. CRF. I. Xexéo, Geraldo Bonorino. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

Agradecimentos

Agradeço aos meus familiares, que independente da situação em que nós encontrávamos, apoiaram minhas escolhas e me propiciaram chegar ao término desta dissertação. Principalmente à minha mãe, Maria Ozenete Barbosa Garrido, personificação da dedicação, que ao longo da minha vida, realizou tanto ou mais esforço que eu, para me colocar onde estou. Assim como a minha companheira Sara Beatriz Philomeno Costa, pela compreensão, auxílio e carinho, que me servem de porto seguro em momentos difíceis.

Agradeço ao professor Jano Moreira e a professora Giseli Rabello, professores de renomes e consagrados no cenário acadêmico, que formam a banca desta dissertação e se dispuseram a avaliar meu trabalho.

Agradeço ao meu orientador Geraldo Xexéo, que ao longo desta caminhada me guiou e se prontificou em me ajudar em todos os momentos, seja para me parabenizar pelos acertos, seja para ser corrigir quando errado.

Agradeço também a Sérgio Rodrigues, que antes de mais nada, é um amigo, que me ajudou nas horas de necessidade e continua fazendo-o até hoje. Sem sua ajuda, não poderia ter chegado ao final da dissertação.

Também agradeço especialmente a Tiago Santos, que mais que um amigo, tenho para mim como um irmão, e me auxilia em todos os momentos, independentemente da situação que exponho. Sem Tiago, o tema da minha dissertação poderia ser outro, pois foi analisando um de seus trabalhos que decidir optar pela anotação de papéis semânticos. Tiago ajuda-me até quando não o quer, e por isso, não poderia deixar de citá-lo em caráter especial. Assim como, agradeço à sua companheira Daiane Evangelista, que, da mesma forma que Tiago, sempre me ajuda quando necessário.

Agradeço a Erick Fonseca, que por seu trabalho *NLPNET*, gerou os insumos necessários para uma boa construção deste trabalho.

Agradeço a todos os meus amigos, mesmo que não possa citar todos aqui, pois seriam muitos, e não quero cometer injustiças pela falta de seus nomes. Porém, dedico um espaço especial para Vitor Machado, Gabriel Almeida, André Albuquerque e Jorge Rama, que seja trabalhando comigo, ou revisando meus textos, sempre se prontificaram a me ajudar.

Por fim, peço desculpas a quem não foi citado. Tenham em mente que os agrade-

cimentos se estenderiam demasiadamente caso citasse à todos que eu deveria fazê-lo. Portanto, sintam-sem lembrando neste momento especial.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ANOTAÇÃO DE PAPÉIS SEMÂNTICOS PARA O PORTUGUÊS POR *CONDITIONAL RANDOM FIELDS*

Luan Barbosa Garrido

Março/2017

Orientador: Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

A anotação de papéis semânticos (APS) pode ser descrita como um meio para diversos fins. Muitas são as áreas dentro do processamento de linguagem natural (PLN) que se beneficiam das etiquetas semânticas dos constituintes da sentença para enriquecer os dados em seus próprios objetivos. Relatado na literatura a diversos séculos, a tarefa de APS renova sua popularidade a partir dos anos 2000, quando o primeiro trabalho de anotação automática foi escrito. Principalmente analisadas para o inglês, muitos trabalhos avaliam cada constituinte da frase separadamente, e não se beneficiam da natureza sequencial de palavras em que a tarefa está incluída. Os últimos trabalho de APS tendem a descentralizar o enfoque inicial e reaproveitam metodologias utilizadas para a língua inglesa em suas próprias línguas, como o espanhol, chinês, francês, sueco e português. Alguns trabalhos já foram realizados para o português, porém, nenhum conseguiu atingir o nível de qualidade obtido para a língua inglesa, e não obstante, somente um trabalho capaz de anotar papéis semânticos a partir de textos puros foi encontrado. Desta forma, esta dissertação visa disponibilizar uma alternativa para anotar papéis semânticos em textos de português sem nenhuma informação agregada, utilizando o modelo de classificação sequencial, denominado *Conditional Random Fields*.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

SEMANTIC ROLE LABELING FOR PORTUGUESE USING CONDITIONAL RANDOM FIELDS

Luan Barbosa Garrido

March/2017

Advisor: Geraldo Bonorino Xexéo

Department: Systems Engineering and Computer Science

Semantic Role Labeling (SRL) can be described as a mean to achieve different purposes. Several subfields inside Natural Language Processing (NLP) benefit from semantic tags for their own goals. Reported in the literature over several centuries, the SRL task regained its popularity since 2000, when the first automatic annotated system was written. Large part of the literature is about SRL for the English language. Moreover, many papers evaluate each constituent of the sentence separately, and do not benefit from the sequential nature of words in which the task is included. The latest SRL works tend to decentralize the initial approach and reuse methodologies applied for the English language in their own languages, such as Spanish, Chinese, French, Swedish and Portuguese. Some methods were proposed for Portuguese, however, they failed to reach the level of quality obtained for the English language, and nonetheless, only one work was capable of annotating semantic roles from raw text. Thus, this work proposes an alternative system for semantically annotate portuguese text without embedded information, using a sequential model called Conditional Random Fields.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Contextualização

O processamento de linguagem natural (PLN) é o campo de estudo que compreende um grande conjunto de técnicas que, por sua vez, provêm recursos para o entendimento das características de um dado texto. São exemplos de subcampos da PLN: sumarização, tradução automática, reconhecimento de entidade nomeada e outros. A **anotação de papéis semânticos (APS)**, do inglês *Semantic Role Labeling*, uma das áreas do PLN, atenta para o completo entendimento do sentido de uma sentença a partir de suas características semânticas. Esse entendimento pode ser encarado de forma análoga à resposta da seguinte pergunta: "*Who did What to Whom, How, When and Where?* (5W1H)". Logo, **a APS tem como objetivo reconhecer e classificar as relações existentes entre os constituintes de uma dada sentença.**

Simploriamente falando, tem-se na APS um meio para um fim, e, desta forma, diferentes usos já foram empregados na literatura: Extração da Informação [5, 6], Tradução Automática [7, 8], Sumarização de Textos [9, 10], Sistemas de Perguntas e Respostas [11, 12], Detecção de Plágio [13], e ainda outros campos distintos como Auxílio à Criação de Patente [14] ou Marcação D'Água [15].

A tarefa de APS segue algumas etapas para sua realização e podem ser resumidas em encontrar os predicados da sentença - constituintes, normalmente verbos, que descrevem eventos -, delimitar seus argumentos - constituintes diretamente associadas à predicados que provêm valor semântico ao evento-, e posteriormente classificá-los de acordo com uma das etiquetas válidas (papéis semânticos). Exemplos:

A) $\{João\}_{AGENTE}$ **quebrou** $\{a\,janela\}_{PACIENTE}$.
B) $\{João\}_{EXPERIENCIADOR}$ **viu** $\{Maria\}_{TEMA}$.

No exemplo A), "João"possui o papel de *AGENTE*, ou seja, aquele que pratica a ação determinada pelo predicado quebrou, enquanto "a janela"possui o papel de

PACIENTE, ou seja, aquele que recebe a ação. Já no exemplo B), devido ao fato de *ver* corresponder a um estímulo sensorial e não uma ação, "João"é etiquetado como *EXPERIENCIADOR*, enquanto "Maria"é o *TEMA* do predicado, ou seja, diferentemente do exemplo anterior, não sofreu modificação pela ação.

Contudo, a determinação dos papéis semânticos nem sempre é direta. Por tratar do sentido de cada constituinte na sentença, algumas vezes mais de uma interpretação e relacionamentos são possíveis. Exemplo:

$$\text{C) } \{João\}_{AGENTE,\,EXPERIENCIADOR} \textbf{ achou } \{sua\,caneta\}_{TEMA}.$$

Neste caso, como *achar* não é um evento voluntário, não há unanimidade em dizer que "João"realizou a ação. Outros autores definem que "João"deveria ser classificado como agente, e desta forma, mais de uma conjectura de etiquetas pode ser atribuída ao constituinte.

A diferença entre os significados semânticos para as etiquetas foi um dos fatores que levaram a uma grande diversidade na literatura, pois não existe consenso entre os autores quanto o conjunto mínimo e explícito de papéis semânticos necessário para representar toda a informação de uma linguagem natural. Além de seus significados, a quantidade de papéis semânticos também é um ponto de discussão nos estudos. De acordo com [16], alguns dos papéis semânticos mais comuns são: *Agente, Paciente, Tema, Experienciador, Estímulo, Instrumento, Posição, Fonte, Meta, Receptor e Beneficiário*.

Apesar da primeira proposta de anotação automática de papéis semânticos ter sido realizada por [17], foi a partir da criação da base de dados *PropBank* [18–20] que grande parte dos trabalhos foi proposta. Esta base foi criada adicionando uma camada de anotação semântica sobre a informação sintática fornecida pelo formato *Penn TreeBank*, que define, entre diversas informações, a estruturação da árvore sintática de constituintes. Sua criação também proporcionou um crescimento da visibilidade da área, culminando em diversas competições que tiveram como tarefa principal a APS, dentre elas: *CoNLL-2004 Shared Tasks* [21], *SemEval Tasks 9* [22], *CoNLL-2009 Shared Tasks* [23] e diversas outras, vide seção 2.4.

Trabalhos baseados nas informações contidas no *PropBank* concentraram-se, principalmente, em aprendizado de máquinas por treinamentos supervisionados como [24–26] (onde os papéis semânticos são conhecidos para cada sentença), embora existam estudos sobre técnicas não-supervisionadas [27–29] (onde os papéis semânticos não estão previamente anotados) e técnicas semissupervisionadas [30–32] (abordagem híbridas entre as duas anteriores). A anotação de papéis semânticos é realizada por diversas formas entre autores, porém, três etapas aparecem em uma boa parte dos trabalhos da literatura: 1) identificação de predicado, 2) identificação dos argumentos e 3) classificação dos argumentos delimitados. Outras tarefas também são utilizadas

como *pruning* (ou poda), quando ramos da árvore sintática são removidas da lista de possíveis argumentos, ou inferência, quando após a finalização da etiquetação, são avaliadas algumas regras de anotação que impedem entre outras coisas argumentos sobrepostos.

Grande parte, senão a maioria, dos trabalhos já realizados dentro do estudo de anotação de papéis semânticos se beneficia da utilização de informações sintáticas para sua resolução. Tais informações são advindas de diversas estruturas auxiliares: identificação de orações da sentença, *shallow parsing* (ou *chunking*)[1] ou *full parsing*. Diferentes trabalhos [33, 34] já identificaram que essa categoria de informação traz benefícios à anotação semântica.

Porém, apesar de contribuir para uma melhor qualidade final do sistema, quando são utilizadas informações sintáticas, inseparavelmente existe a necessidade de ferramentas auxiliares para suas definições, uma vez que utilizar somente as informações fornecidas pelas base de dados impede a reprodução do sistema desenvolvido em novos textos sem marcação sintática ou semântica (doravante denominados **textos puros**).

Quando a língua tratada é rica em sistemas e ferramentas auxiliares, como o inglês, onde muitos estudos e trabalhos já foram realizados em diferentes áreas e a disponibilidade por ferramentas é grande, o custo pela utilização desta informação auxiliar agregada é amortizado. Porém, quando a língua tratada não possui tantos recursos, como o português, existe decréscimo da qualidade final do modelo, uma vez que é necessária uma procura/desenvolvimento paralela a fim de preencher tais lacunas, que por vezes propagam grandes diferenças (erros) nos resultados finais desejados.

Diferentemente, outros sistemas abordam a APS sem a utilização de qualquer informação sintática ou conhecimento explícito linguístico. O mais icônico desses sistemas, e que possui bons/equiparáveis resultados perante àqueles que utilizam denomina-se *SENNA* [35, 36]. Em seu trabalho, os autores utilizaram uma rede neural artificial de convolução, treinada a partir da vetorização de palavras (transformação de palavras em vetores de números) e criaram uma arquitetura capaz de responder diversas tarefas dentro do processamento de linguagem natural, dentre elas a APS.

Originalmente proposta para o inglês, o *PropBank* serviu como diretriz para criação de seus derivados para diversas línguas, tal qual o português. O *PropBank-BR* [37, 38], serviu como insumo para alguns trabalhos, cujos quais ainda apresentam resultados piores quando comparados ao inglês. Uma das possíveis causas para a piora dos resultados para o português pode ser atribuída ao tamanho da base de dados

[1] *Shallow Parsing* ou *Chunking* é o processo em que os sintagmas nominais, verbais, entre outros, são especificados para a sentença. Diferentemente do *full parsing*, o *shallow parsing* não fornece a estrutura de árvore da sentença e pode ser entendido como um "achatamento"da árvore sintática.

para treino. Enquanto o *PropBank* possui aproximadamente 1 milhão de palavras, o *PropBank-BR* possui apenas 1/7 do tamanho da base original. Vale ressaltar que abordagens propostas para outras línguas, que não dispõem de tantos recursos quanto inglês, também possuem valores inferiores [30].

Dentre os trabalhos mais notórios para o português, [39–41] a partir de um conjunto de regras criado manualmente, se propõe a realizar a APS sem o auxilio de aprendizado de máquina, embora apresente todas as limitações oriundas pela não generalização da abordagem. Apesar de possuir bons resultados e ser capaz de anotar textos puros, a proposta deste autor tem licença privada e não pode ser reproduzida sem permissão.

Já [1], realizou um amplo estudo sobre diversos modos de aprendizagem automática para a resolução da APS, obtendo os melhores resultados quando o sistema foi analisado no *PropBank-BR*. Em contrapartida, todas as abordagens efetivadas pelo autor foram realizadas utilizando dados *gold*[2], logo, não pode ser reproduzido em textos puros.

Em contrapartida [4], inspirado no *SENNA*, criou o primeiro e único sistema de APS cujo qual é possível realizar anotação para textos puros, denominado *NLPNET*. Através de vetorização das palavras, utilizando bases textuais escritas em português como o *Wikipedia* e o *corpus* PLN-BR [42], o autor avaliou os resultados sem a inserção de informações sintáticas. O autor também avaliou a qualidade do modelo utilizando tais informações e comprovou que os melhores resultados eram oriundos dos modelos que se valiam dos dados sintáticos.

Desta forma, mesmo que outros sistemas possuam qualidade superior ao relatado por [4], toda a comunidade da área de processamento de linguagem natural está limitada a utilização do *NLPNET*.

1.2 Objetivos

A partir da análise dos sistemas já desenvolvidos tantos para o português quanto para o inglês, **o principal objetivo deste trabalho é apresentar um sistema alternativo para a resolução do problema de anotação de papéis semânticos a partir de um texto puro para o português**, e assim, fornecer uma maior gama de ferramentas para outros pesquisadores se beneficiarem da análise semântica em seus trabalhos, que hoje, é obtida somente pela utilização do *NLPNET*.

Após uma análise das principais etapas da APS, identificou-se algumas características de natureza sequencial para anotação do texto. Para a etapa de identificação de predicados, dificilmente predicados são encontrados contiguamente, assim como,

[2]Dados *gold* são aqueles retirados diretamente da base de dados e não apresentam erros por sua extração dinâmica.

normalmente, após a ocorrência de um verbo auxiliar, um predicado é encontrado. Da mesma forma, para a identificação de argumentos, um constituinte é tido como continuação de um argumento prévio de acordo com suas características. Sendo assim, **esta dissertação também visa avaliar o impacto da utilização de modelos de aprendizado com abordagem sequencial para realização da APS**, como o *Conditional Random Fields* (CRF).

O CRF é uma abordagem discriminativa que ainda não foi amplamente utilizada dentro da área de APS. Para o português, apenas um trabalho foi encontrado na literatura [43], porém este não foi utilizado para a anotação completa de papéis semânticos, somente para três etiquetas, e apresentou resultado muito abaixo do esperado.

Como citado anteriormente, os trabalhos [33, 34] identificaram a forte relação entre informação sintática e semântica. Além disso, descreveram que, embora a análise sintática completa (*full parsing*) produza resultados melhores na tarefa de APS completa, a análise simplificada (*shallow parsing*) produz resultados tão bons, e até melhores, em cada uma das etapas individualmente. Aliada a tal questão, e devido ao fato da pouca expressividade de ferramentas auxiliares para anotação de papéis semânticos, apresenta-se neste trabalho uma proposta para anotação de papéis semânticos de textos puros, utilizando somente informações sintáticas advindas de *shallow parsing* para treinamento de modelos CRF.

1.3 Organização da Dissertação

Os capítulos que seguem são divididos da seguinte maneira: O capítulo 2 trata do conhecimento teórico a respeito dos papéis semânticos, além disso discorre sobre as bases de dados que serviram como insumo para os trabalhos baseados em aprendizado de máquina e descreve algumas das várias competições que objetivaram a APS. O capítulo 3 aborda os diversos trabalhos já realizados tanto para o inglês, quanto para o português. O capítulo 4 explana o CRF e elucida os motivos para a escolha deste modelo como solução para o problema de APS. Já o capítulo 5 apresenta a implementação proposta culminando no capítulo 6, que apresenta os resultados obtidos comparando-os à trabalho de terceiros. Finalmente, o último capítulo conclui o trabalho proposto.

Capítulo 2

Papéis Semânticos

Para consolidação dos papéis semânticos, predicados e argumentos, este capítulo tem por finalidade apresentar tais conceitos linguísticos, cujos quais serviram de base para o estudo e realização desta dissertação, bem como para a construção de ambas as bases *PropBank* e *PropBank-BR*.

Concomitantemente, este capítulo aborda algumas das diferentes representações apresentadas da literatura quanto à noção de papéis semânticos e finaliza tratando a respeito da visibilidade que a APS vem ganhando desde sua automatização em 2002, representada aqui pela apresentação de diversas competições em que a tarefa foi tida como área de estudo principal.

2.1 Conceituação

Incluso na área de processamento de linguagem natural (PNL) encontra-se o campo de estudo dos papéis semânticos, por vezes denominado papéis temáticos. Tal campo preocupa-se com a identificação das relações semânticas entre predicados e seus argumentos, ou seja, refere-se às funções semânticas dos sintagmas[1] na sentença em relação ao seu predicado.

Um predicado, representado por nomes, adjetivos, advérbios ou mais comumente verbos, estabelece uma relação de sentido com seu sujeito e seus complementos. Esse sentido pode derivar de ação, estado, eventos mentais e relacionais, entre outros. De acordo com [44], o predicado tem seu sentido incompleto, e somente atinge sua plenitude quando é especificado conjuntamente com seus argumentos. Assim, define-se papéis semânticos/temáticos as funções exercidas pelos argumentos (formas como a entidade está envolvida na sentença) quanto a um predicado.

[1]Sintagma deve ser entendido como uma sequência hierarquizada de elementos linguísticos, que compõem uma unidade na sentença. São eles: sintagmas nominais (NP), verbais (VP), preposicionais (PP), adverbiais (ADVP) ou adjetivais (ADJP)

Aos papéis semânticos atribuem-se diferentes significados de acordo com o predicado e a sentença em que estão empregados. Exemplo:

$$\{João\}_{AGENTE} \text{ quebrou } \{a\, janela\}_{PACIENTE}.$$

O predicado **quebrou** possui dois distintos argumentos de diferentes significados: *João* exemplifica o papel *AGENTE*, ou seja, aquele que realiza uma ação, enquanto *a janela* é o *PACIENTE* da sentença, ou seja, aquele que sofre a ação.

A introdução da noção de papéis semânticos se deu por [45], [46] e [47], pela alegação de que as funções sintáticas de sujeito e objeto são insuficientes para certas construções existentes:

1. $\{João\}_{AGENTE}$ abriu $\{a\, porta\}_{PACIENTE}$ $\{com\, a\, chave\}_{INSTRUMENTO}$.
2. $\{A\, porta\}_{PACIENTE}$ abriu $\{com\, a\, chave\}_{INSTRUMENTO}$.
3. $\{A\, chave\}_{INSTRUMENTO}$ abriu $\{a\, porta\}_{PACIENTE}$.

Nas três sentenças acima, define-se como sujeito respectivamente: (1) João, (2) A porta e (3) A chave. Os três distintos sujeitos desempenham diferentes papéis semânticos: (1) *AGENTE*, (2) *PACIENTE* e (3) *INSTRUMENTO*. Logo, verifica-se que somente a função sintática do constituinte é incapaz de determinar sua significância semântica [48]. Como os predicados são polissêmicos, diferentes quantidades e significados de argumentos são possíveis, de acordo com a sentença em questão.

A gama de tipos e significados de papéis semânticos conhecidos pela literatura é vasta e por vezes possui algumas divergências de acordo com diferentes autores, ou seja, ainda hoje não existe um consenso acerca de todos os papéis semânticos [16, 49, 50].

A principal delas se dá quanto a quantidade de papéis semânticos necessários para representação de uma linguagem natural. Todavia, outra diferença clara entre autores é a capacidade de um mesmo sintagma nominal possuir somente um papel semântico defendida por [46] ou pode ser preenchido por mais de um papel semântico, defendido por [47].

A seguir, elucida-se algumas das diversas teorias propostas ao longo dos anos.

2.2 Representações de Papéis Semânticos

2.2.1 Gramática de Casos de Fillmore: *The case for case*

Inicialmente descrita na gramática sânscrita desenvolvida por Pãnini, um gramático indiano, em 4 A.C., os papéis semânticos tornaram-se novamente foco de discusão por Fillmore em 1967 [46]. Neste trabalho, Fillmore parte da hipótese que línguas

humanas são restritas, e desta forma, as relações entre constituintes de um sentença (verbos e demais constituintes) se enquadram em um pequeno número de tipos, chamados **casos conceituais**, ou **relações de caso**.

A partir de tal afirmação, lança a primeira lista de casos conceituais, ligeiramente interpretados como papéis semânticos: *AGENTIVO, INSTRUMENTAL, DATIVO, FACTITIVO, LOCATIVO E OBJETIVO.*

- AGENTIVO (A): Instigador da ação identificado pelo verbo (tipicamente animado).
- INSTRUMENTAL (I): força ou objeto inanimado, causalmente implicado à ação ou estado identificados pelo verbo.
- DATIVO (D): ser animado afetado pelo estado ou ação identificados pelo verbo
- FACTIVO (F): objeto ou ser resultante da ação ou estado identificados pelo verbo ou compreendido como parte do significado do verbo.
- LOCATIVO (L): localização ou orientação espacial do estado ou ação identificados pelo verbo.
- OBJETIVO (O): qualquer coisa representada por um substantivo cujo papel na ação ou estado identificados pelo verbo é determinado pela interpretação semântica do próprio verbo.

Exemplo:

$$\{João\}_{AGENTIVO} \text{ abriu } \{a\,porta\}_{OBJETIVO}\ \{com\,a\,chave\}_{INSTRUMENTAL}.$$

Fillmore explicita que "uma sentença, em sua estrutura básica, é formada por um verbo e um ou mais sintagmas nominais, cada um associado ao verbo por meio de uma relação específica de caso e que cada relação de caso pode ocorrer apenas uma vez numa sentença simples". Também define que todo sintagma nominal associado a um verbo possui um caso conceitual e que o número e os tipos de papéis semânticos associados ao verbo são determinados pela semântica do verbo propriamente dita.

Com o passar dos anos, a lista de casos conceituais de Fillmore foi extendida por diversos autores, abrangendo mais casos de relacionamento semânticos não previamente identificados. Devido ao crescimento da quantidade de papéis semânticos, em [51] Fillmore identificou pelo menos 2.000 papéis semânticos para diferentes verbos.

2.2.2 Proto-Papéis de Dowty

Embasado na teoria de protótipos, Dowty [52] em vez de definir um único conceito subjacente que capturasse a essência do agente, procurou associar características definidoras suficientes com agentes e pacientes para que fosse sempre possível distinguilas. Tais características são tidas como acarretamentos lexicais despreendidos da relação estabelecida entre um predicado (verbo) e seus argumentos.

Desta forma, identifica duas categorias distintas que seriam capazas de substituir a série de papéis semânticos necessários para análise das relações conceituais: o papel de *Proto-Agente* e o papel de *Proto-Paciente*. O primeiro é responsável por carregar as noções de causas, movimentos, entre outros. Já o segundo responsável por mudanças de estado, de ser afetado por outro participante e de não existir independentemente do evento. As características de cada "proto-papel"estão arroladas abaixo:

Tabela 2.1: Definições dos acarretamentos lexiais dos "proto-papéis"de Dowty

Proto-Agente	**Proto-Paciente**
Envolvimento volitivo	Sofre mudança de estado
Consciência	Tema incremental
Causa um evento ou mudança de estado em outro participante	Afetado causalmente por outro participante
Movimento	Estacionário
Existe independentemente do evento nomeado pelo verbo	Não existe independentemente do evento

Concomitantemente em seu trabalho, Dowty busca relacionar argumentos verbais no léxico para a ocupação das posições sintáticas de sujeito e objeto. Tal mapeamento é realizado pela análise dos argumentos do verbo alvo. O argumento que possui mais características de proto-agente será declarado sujeito, enquanto o argumento que possuir mais características de proto-paciente será tido como objeto.

Dowty em seu trabalho reduz a granularidade e de definição de fronteiras entre os papéis semânticos, porém, aos mesmo tempo, os torna vagos quanto a seus significados, uma vez que apenas duas categorias não possuem alto teor explicativo.

2.2.3 Classes de Levin

Já para Levin, o conjunto de quadros sintáticos associado a um verbo específico reflete os componentes semânticos subjacentes, que restringem os argumentos permitidos. Em seu trabalho [53], define classes e subclasses para verbos do inglês conjuntamente com suas alternâncias sintáticas (diáteses), partindo da análise para cada verbo onde este é possível de ocorrer ou não em pares de quadros sintáticos, preservando suas alternâncias de diátese.

De acordo com [54]: "A hipótese para o agrupamento dos verbos é que a base da habilidade dos falantes de uma a língua em determinar o comportamento de um verbo está relacionada com o significado deste verbo. Assim, os verbos de um mesmo grupo compartilham tanto características sintáticas quanto aspectos semânticos ".

Ao final do estudo para mais de 3 mil verbos, foram identificados 47 classes principais, totalizando 193 classes e subclasses distintas.

2.2.4 Semântica de Frames de Fillmore

A partir do reconhecimento das limitações dos casos conceituais e da noção de Frames, utilizado por [55] para Representação do Conhecimento na Inteligência Artificial, Fillmore definiu a **semântica de frames**.

Este trabalho defende que o significado das palavras (unidades lexicais) devem ser analisados pela total análise dos **frames semânticos** que tais unidades evocam, ou seja, os significados das palavras devem ser descritos em relação aos frames semânticos, os quais são definidos como uma representação esquemática de situações onde existe participação de vários componentes.

Tais frames, de acordo com [56], "são esquematizações de estruturas conceptuais, de crenças, de práticas institucionais que emergem da experiência do dia a dia. Trata-se da representação de uma situação, um objeto ou evento inserida em um background (pano de fundo). Assim configurado, a compreensão do sentido de um item lexical implica conhecer o frame no qual determinado sentido está relacionado".

Exemplo: o frame *transação comercial* incluí elementos como *comprador, vendedor, mercadorias e dinheiro*. A este frames, referem-se diferentes verbos, que identificam e se relacionam a diferentes aspectos, como *comprar*, que se relaciona diretamente a comprador e mercadoria, afastando-se dos elementos de vendedor e dinheiro.

Elementos tais que não devem ser diretamente relacionados à papéis semânticos, e sim como papéis situacionais, diferentemente da gramática de casos.

2.2.5 Catálogo de Cançado

Cançado objetiva ser análogo à Levin em seu trabalho focado o português [57]. São descritos 861 verbos de mudanças do português brasileiro, divididos em 4 classes (mudança de estado, mudança de estado locativo, mudança de local e mudança de posse) e organizados de acordo com a teoria da decomposição de predicados, significados sintáticos e semânticos. Para cada classe, uma lista de verbos é apresentada, conjuntamente com suas propriedades sintáticas e semânticas. Exemplo:

O João quebrou o vaso. (Alternância causativa com um agente como sujeito)
A queda quebrou o vaso. (Alternância causativa com uma causa como sujeito)

A autora acredita que, uma vez que existem aproximadamente 6000 verbos em utilização no português brasileiro, seu catálogo possui uma boa representação do total de dados.

2.3 Bases Lexicais

Neste trabalho a principal base lexical utilizada é o *PropBank-BR*, derivado do *PropBank*. Porém, vale o comentário a respeito do FrameNet e do VerbNet.

A *FrameNet* [58], desenvolvida em Berkley, possui como base a semântica de Frames de Fillmore [55] e suas principais unidade de análise são o *frame* e a *unidade lexical*, uma vez que o segundo evoca o primeiro. De acordo com [59] um dos seus principais objetivos é: "identificar e descrever frames semânticos, analisar as relações presentes entre os evocados e identificar os padrões valenciais das palavras, considerando-se três níveis de anotação: Tipos Sintagmáticos (TS), Funções Gramaticais (FG) Elementos de Frames (EF), sendo os primeiros níveis sintáticos e o último micropapéis temáticos de natureza. semântica".

A partir do trabalho de Levin [53], [16] desenvolveu-se o *VerbNet*. Este repositório contém as classes de Levin, com os possíveis papéis semânticos associados a cada uma delas. O *VerbNet* além de identificar os papéis semânticos, apresenta restrições semânticas, de acordo com a classe, bem como predicados lógico-semânticos, que são responsáveis por definir os componentes de significados de cada classe.

A maior diferença entre as três bases distintas, se dá pela granularidade dos papéis semânticos. Em nível de abstração, do menor para o maior, tem-se:

$$FrameNet \rightarrow VerbNet \rightarrow PropBank$$

Onde, o *FrameNet* define papéis semânticos específicos para cada frame em questão (+ 2.500 *frames elements*), o VerbNet apresenta papéis que se enquadram

a uma determinada classe/subclasse (24 tipos de papéis semânticos) e o PropBank apresenta a solução mais abstrata, aplicando-se a qualquer contexto (16 tipos).

De acordo com os valores retirados de [26]:

Tabela 2.2: Comparativo entre bases lexicais. Da esquerda para direita, maior nível de abstração. Da direita para esquerda, maior nível de especificação.

Atributo	**FrameNet**	**VerbNet**	**PropBank**
Unidades Lexicais	11.600	5.733	6.204
Lemmas	6.000	3.965 (Verbos)	5.213 (verbos)
Papéis Semânticos	+2.500 *frame elements*	24 tipos de papéis	16 Args, 6.000+ (verbos com papeis específicos)
Dados Anotados	150.00 Sentenças abrangendo 6.800 unidades lexicais	1 Milhão de palavras (90% de cobertura de *token*	1,75 Milhões de palavras (todos os verbos)

Todas as bases supracitadas possuem análogas ao Português-Brasileiro.

2.3.1 *PropBank*

Criado originalmente para a língua inglesa [18–20], a base de dados de proposições, *PropBank* refere-se à um *corpus* anotado quanto papéis semânticos, tanto de argumentos quanto de adjuntos. Uma proposição é entendida como um verbo e seus sintagmas associados, em uma particular relação semântica.

PropBank pode ser entendido como uma camada superior ao *Penn TreeBank*, mantendo suas anotações sintáticas inerentes e adicionando a camada de anotação da estrutura predicado-argumento. Por ser baseado no *Penn TreeBank*, os textos que compõe o *PropBank* são de caracter jornalística (oriundos do *Wall Street Jornal* (WSJ) e, em menor parcela, literários, oriundos do *corpus Brown*.

Por se valer da estruturação em árvores fornecida pelo *Penn TreeBank*, a camada de anotação do *PropBank* age da seguinte forma: (1) Identifica o predicado alvo, (2) Anota o nó de árvore não-terminal que corresponda a completude do argumento e (3) anota tal nó com sua respectiva etiqueta.

Para cada predicado, os argumentos são numerados de 0 a 5 (*Arg0 - Arg5*), onde o *Arg0* geralmente apresenta características de Proto-Agente [52] e o *Arg1* apresenta característica de Proto-Paciente ou Tema.

Não existe generalização consistente para argumentos de maior numeração (*Arg2* - *Arg5*), apesar do esforço realizado para definição dos papéis através dos membros das classes do *VerbNet*. O o projeto *SemLink* [60] visa relacionar os *rolesets* do *PropBank* com as classes do *VerbNet* e os *frames* do *FrameNet*. Tal associação aumenta o esclarecimento obtido para cada um dos argumentos etiquetados de acordo com o verbo em questão, já que os 2 últimos *datasets* são mais específicos quanto suas anotações.

Vale a ressalva que não é necessário um predicado apresentar todas as etiquetas.

O *PropBank* define 3 construtores :

1. *Rolesets*: conjunto de papéis semânticos que correspondem à um distinto sentido de um verbo,
2. *Framesets*: *roleset* mais o *frame* semântico que o verbo participa (verbos polissêmicos podem ter mais de um *frameset*,
3. *Frame files* coleção de *framsets* de um verbo.

No exemplo abaixo, retirado de [19], temos uma anotação do *PropBank* para o *frame file* do verbo *leave*, onde é apresentado 2 de seus *framsets* com seus respectivos *rolesets*:

> **leave.01** sense: move away from
> roles: *Arg0*: Entity Leaving, *Arg1*: Place Left, *Arg2*: Attribute
> Example: The move left the companies as outsite bidders
> *Arg0*: The move
> *Rel*: left
> *Arg1*: the companies
> *Arg2*: as outside bidders

> **leave.02** sense: give
> roles: *Arg0*: Giver, *Arg1*: Thing given, *Arg2*: Receiver
> Example: An ambitious expansion has left Magna with excess capacity
> *Arg0*: An ambitious expansion
> *Rel*: left
> *Arg1*: Magna
> *Arg2*: with excess capacity

Além dos argumentos numerados de 0 à 5, o *PropBank* apresenta um conjunto de argumentos modificadores que podem ser aplicados à qualquer verbo em uma

sentença válida. Tais argumentos, chamados **ARGMs** são referentes a adjuntos, negação (**NEG)** e para verbos modais **(MOD)**. A seguir, a lista completa desses argumentos:

Tabela 2.3: Etiquetas do tipo ARGMs e suas descrições

Etiqueta	**Descrição**
ADV	Advérbio
CAU	Causa
DIR	Direção
DIS	Conectivos Discursivos
EXT	Extensão
LOC	Local
MNR	Maneira
MOD	Verbo Modal
NEG	Marcador de Negação
PNC	Propósito
TMP	Tempo

Apesar da maioria dos predicados do *PropBank* serem verbos, também existem predicados nominais e adjetivais. A distribuição completa da anotação do *PropBank* se encontra na lista abaixo [2]:

Tabela 2.4: Distribuição completa da anotação do *PropBank*

Classe Gramatical	***Frame Files***	**Predicados**	***Framesets***
Verbos	5941	6743	8123
Substantivos	2546	2666	3140
Adjetivos	1906	1923	2239

O principal objetivo da base de dados é servir como um recurso linguístico-computacional e permitir a criação de sistemas de aprendizado supervisionado para reconhecimento e etiquetação das classes dos argumentos. Por ser o *corpus* anotado mais abstratamente dentre o citados, é um dos *corpus* com maior utilização nos trabalhos internacionais, e possui versões em inglês, chinês, espanhol, português brasileiro e outros.

2.3.2 *PropBank-BR*

Com cerca de um sétimo de tamanho do seu análago para o inglês, o projeto *PropBank-BR* [37, 38] segue a linha de anotação de forma similar ao *corpus* original.

[2] http://verbs.colorado.edu/propbank/propbank-status-en.html, acessado em 02/2017.

A partir da anotação *TreeBank* do *corpus* Bosque da *Floresta Sintá(c)tica*, obtido pelo *parser* PALAVRAS [39, 41], e de uma revisão manual por linguístas, o *PropBank-BR*[3] consiste em uma base de dados de proposições anotadas para o português brasileiro. Da mesma forma que seu original, a versão em português apresenta argumentos numerados de 0 à 5 além de seus argumentos modificadores, listados abaixo:

Tabela 2.5: Etiquetas do tipo ARGMs, suas descrições e exemplos

Etiqueta	**Descrição**	**Exemplo**
ADV	Advérbio	Aqui **só** joga quem está bem.
CAU	Causa	Ele só não jogava porque **não estava bem**.
DIR	Direção	Eles viajam **em sentido contrário**.
DIS	Conectivos Discursivos	**Taí**, gostei da idéia!
EXT	Extensão	Tenho que estudar **muito**.
LOC	Local	**Aqui** só joga quem está bem.
MNR	Maneira	Façam **como eu**!
NEG	Marcador de Negação	Ele só **não** jogava porque não estava bem.
PNC	Propósito	Color sai **para jantar** e se assuta com *flashes*.
PRD	Predição Secundária	IBGE mostra inflação menor **no Rio e SP**.
REC	Recíproco	*Sommeliers* paulistas **se** superam em evento.
TMP	Tempo	**Nunca** tive muito interesse, apesar de já ter experimentado.

Exemplo de sentença anotada no *PropBank-BR*:

[3]http://143.107.183.175:21380/portlex/index.php/en/projects/propbankbringl

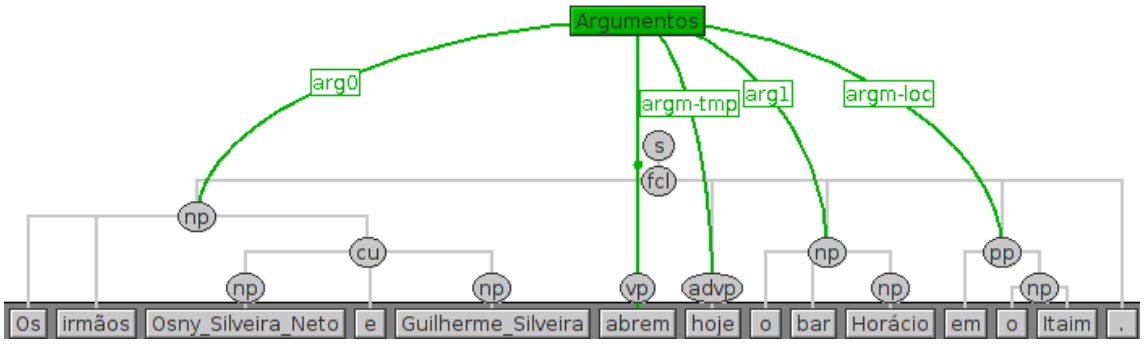

Figura 2.1: Exemplo de sentença retirada da versão 1.0 da base de dados e apresentada pelo programa SALTO.

O *PropBank-BR* está disponível em 4 versões distintas. Para manter o caráter comparativo à outros trabalhos, optou-se utilizar a versão 1.0 da base, porém abaixo estão listadas todas as versões:

1. **1.0**: Mesma utilizada por outras implementações de anotação de papel semântico para o português. Dividido em treino e teste com as mesmas partições utilizadas por outros autores [1, 4]. Esta versão apresenta as correções realizadas e descritas por [4] em seu trabalho.

2. **1.1**: Versão que apresenta algumas correções realizadas e identificação do sentido do verbo em português.

3. **2.0**: Originado por uma extração do *corpus* PLN-BR [42] sobre árvores não revisadas, usando dois anotadores para cada sentenção (anotação duplo-cego)

4. **3.0**: Novamente geradas pelo *parser* PALAVRAS, o *corpus* utilizados foi retirado pelo particionamento do Buscapé [61], um *corpus* de opiniões de usuários sobre produtos. As árvores sintáticas da amostra não foram revisadas por humanos.

A seguir algumas estatísticas da versão 1.0 da base de dados:

Tabela 2.6: Distribuição quanto predicados

Estatística	**Treino**	**Teste**
Sentenças Distintas	3.164	144
Contagem de Predicados	5.537	239
Contagem de Predicados Distintos	2.869	212
Média de Sentença Anotada por Predicado	1,9	1,12

Tabela 2.7: Distribuição quanto etiquetação de argumentos

Etiqueta	**Quantidade**		**Percentual**	
	Treino	**Teste**	**Treino**	**Teste**
A0	2934.0	124.0	22,63%	23,13%
A1	4937.0	211.0	38,07%	39,37%
A2	1063.0	38.0	8,20%	7,09%
A3	111.0	2.0	0,86%	0,37%
A4	74.0	1.0	0,57%	0,19%
AM-ADV	349.0	20.0	2,69%	3,73%
AM-CAU	155.0	1.0	1,20%	0,19%
AM-DIR	13.0	2.0	0,10%	0,37%
AM-DIS	283.0	11.0	2,18%	2,05%
AM-EXT	80.0	1.0	0,62%	0,19%
AM-LOC	751.0	27.0	5,79%	5,04%
AM-MNR	392.0	18.0	3,02%	3,36%
AM-NEG	316.0	19.0	2,44%	3,54%
AM-PNC	166.0	5.0	1,28%	0,93%
AM-PRD	186.0	6.0	1,43%	1,12%
AM-REC	60.0	5.0	0,46%	0,93%
AM-TMP	1097.0	45.0	8,46%	8,40%

Obs.: Por somente um predicado apresentar Arg5 ele não foi considerado na base, pela dificuldade de generalização. Além disso, complementos de argumentos (C-Arg*) também foram desconsiderados, pela prerrogativa que argumentos devem ser construídos de forma contígua.

2.4 Competições Internacionais de APS

A anotação de papéis semânticos já esteve em foco em diversas competições internacionais ao longo dos anos. Apesar de novos trabalhos se fazerem presentes até hoje, seu momento inicial se deu no início dos anos 2000, culminando ao longo da década à diversos sistemas propostos e abordagens inovadoras. Abaixo, segue uma lista de competições realizadas e suas características:

2.4.1 *Senseval-3 (2004) Taks*: Anotação Automática de Papéis Semânticos

Utilizando dados do *FrameNet* para o inglês, a competição [62] consistia em: dada uma sentença, o predicado e seu *frame, identificar e classificar os frame elements* da sentença.

A precisão média obtida pelas oito equipes participantes foi de 80,3%. Já a cobertura média foi de 75,7%.

2.4.2 *CoNLL 2004 - Shared Task*: Anotação de Papéis Semânticos)

Nesta competição [21], os dados utilizados eram originários do *PropBank*, sem as informações de árvores sintáticas completas e bases externas, embora informações como reconhecimento de entidade nomeada e identificação de sentenças (início- fim) fossem fornecidas por meio de sistemas auxiliares.

Dez participantes entraram na competição. O melhor sistema [63] obteve 69,49 na medida F_1 para a tarefa completa. Vale ressaltar que nesta competição os predicados da sentença já estavam identificados como informação de entrada.

2.4.3 *CoNLL 2005 - Shared Taks*: Anotação de Papéis Semânticos)

Utilizando uma versão mais robusta do *PropBank*, esta competição [64] diferentemente da sua antecessora, forneceu informaões de árvores sintáticas completas geradas por *parsers* auxiliares. Os desafios foram separados por fechado (onde somente os dados de treinamento poderiam ser utilizados) e aberto (onde informação externa poderia ser inserida).

Ao total de 19 participantes no desafio fechado (e nenhum para o desafio aberto), o melhor sistema implementado [24] atingiu 79,4 de medida F_1 no conjunto de teste do *Wall Street Journal (WJS)*, 67,8 para o *corpus Brown* e 77,9 para o teste combinado.

2.4.4 *SemEval-2007 Task 9*: Anotação Semântica Multi-Nível de Catalão e Espanhol

A competição [22] visou anotação em 3 níveis semânticos para o catalão e o espanhol:

- Classificação de papéis semânticos e desambiguação verbal,
- Desambiguação de todos os substantivos,

- Reconhecimento de entidades nomeadas.

Para a APS, o *PropBank* foi utilizado. O melhor sistema [65] obteve os resultados 83,4 e 84,1 na medida F_1 para catalão e espanhol respectivamente.

2.4.5 *CoNLL 2008 - Shared Taks*: Anotação Conjunta de Dependências Sintáticas e Semânticas

A *corpora* desta competição utilizou uma combinação de vários *corpus*: *Peen TreeBank, PropBank e NomBank* [66]. Foi dividida, também, em três etapas:

- Análise sintática de depêndencias,
- Identificação e desambiguação de predicados semânticos,
- Identificação de argumentos e atribuição de papéis semânticos para cada predicado.

O melhor sistema [67], dos 19 participantes, obteve 81,75, 69,06 e 80,37 para medida F_1 para os dados de testes do *WSJ*, *Brown* e conjunto respectivamente.

2.4.6 *CoNLL 2009 - Shared Task*: Dependência Sintáticas e Semânticas em Múltiplas Línguas

Análago a competição ocorrida 1 ano antes, esta competição [23] tinha uma abrangencia para 6 línguas além do inglês (catalão, chinês, tcheco, alemão, japonês e espanhol). As equipe participantes podiam escolher entre a participação para a tarefa conjunta (análise sintática de dependência e APS) ou somente a APS.

O melhor sistema [68], que considerou somente a tarefa de APS, dos 7 sistemas participantes, obteve a média da medida F_1 de 80,47 para todas as línguas da competição.

Ao longo dos anos outras competições também propuseram o estudo de APS, dentre elas: SemEval-2010[4], SemEval-2013[5], NAACL-13[6], entre outras.

2.5 Considerações Finais

Após uma explanação sobre a anotação de papéis semânticos, as diferenças e similaridades entre as diversas visões da literatura, o capítulo abordou sobre as principais competições que apontaram a APS como tarefa principal. Também foram abordadas

[4]http://semeval2.fbk.eu/
[5]https://www.cs.york.ac.uk/semeval-2013/
[6]http://naacl2013.naacl.org/

as duas principais base de dados utilizadas pela grande maioria dos trabalhos de anotação automática, o *PropBank* e seu análogo para o português *PropBank-BR*.

Capítulo 3

Trabalhos Relacionados

Diversas abordagens já foram realizadas para a anotação automática, principalmente após a criação do *PropBank*. É notório que no começo dos anos 2000 a grande quantidade de estudos publicados concentraram-se para o inglês, uma vez que era a língua mais propícia para tal: criação do *PropBank*, grande disponibilidade de ferramentas auxiliares, competições próprias, entre outros. Após a criação de bases derivadas do *PropBank* para outras línguas, uma tendencia verificada em buscas por trabalhos apontam um maior difusão dos trabalhos. Porém, ainda hoje, existe a busca pelo aperfeiçoamento da APS para o inglês.

Neste capítulo serão apresentadas diversas propostas realizadas ao longo dos anos, divididas pelas línguas se propuseram abordar.

3.1 A tarefa de APS

A tarefa de anotação de papél semântico pode ser dividida em diversas etapas, embora, as seguintes etapas serem tidas como obrigatórias para todos os sistemas (desconsiderando-se a etapa de pré-processamento):

1. **Reconhecimento de predicado alvo**: Apesar de diversos algoritmos e competições tomarem a anotação do predicado alvo como um dado de entrada, esta etapa é primordial para o início da APS. Diversos artigos exemplificam métodos para o reconhecimento do predicado, tais como: [4, 69–72]

2. **Poda da Árvore Sintática**: Para algoritmos que utilizam como dado de entrada a árvore sintática da sentença em questão, [73] desenvolveram uma heurística onde se é possível desconsiderar ramos com pouca probabilidade de identificação de argumentos. Tal heurística reduz o tempo de treinamento total do algoritmo e aprimora os resultados, um vez que impede que constituintes improváveis sejam erroneamente tratados como argumentos. [33] em seu traba-

lho afirma que bons resultados na etapa de poda é essencial para a qualidade final do sistema de APS. Exemplo de árvore sintática na figura 3.1

3. **Identificação de Argumentos**: Nesta etapa todos os constituintes da sentença são analisados a fim de se estabelecer inícios e fins para todos os argumentos do predicado alvo. Neste etapa diferentes formas de delimitação são utilizadas, tal como etiquetação IOB, IOBES [74], início-fim, entre outras.

4. **Classificação de Argumentos (Etiquetação)**: Na etapa de classificação, todos os argumentos identificados na etapa anterior são etiquetados por uma das etiquetas válidas.

5. **Inferência**: A etapa de inferência consiste em atribuir regras para a etiquetação dos argumentos, tais como: Somente aceitar argumentos contíguos, não permitir repetição de argumentos, remoção de sobreposição de argumentos, entre outros. Alguns algoritmos utilizam outros pós-processamentos baseados em lógica [75] ou outras abordagens nesta etapa.

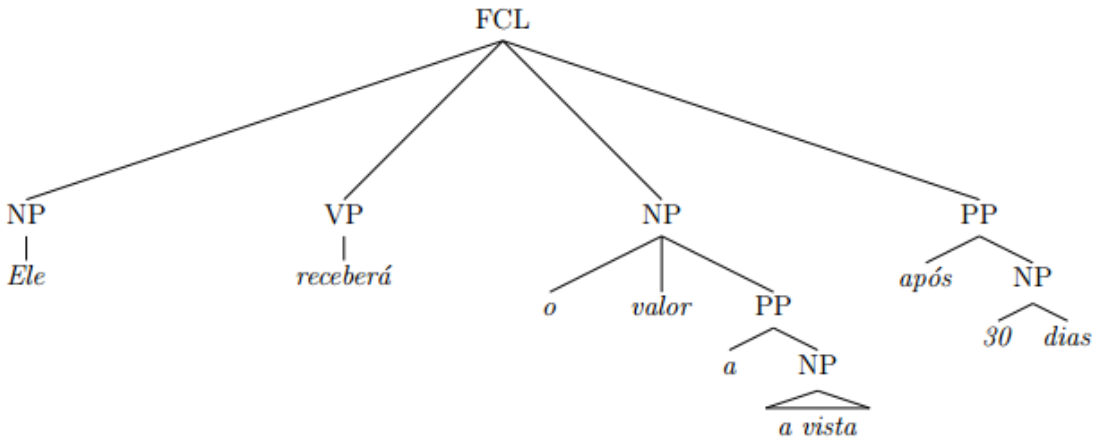

Figura 3.1: Exemplo de árvore sintática: **Fonte:** [1]

.

A respeito das etapas de identificação e classificação de argumentos, o estudo de [25] esclarece que para a primeira tarefa é requerido um maior conhecimento sintático sobre a estruturação dos constituintes na da árvore sintática de dependências, enquanto para a classificação, um é necessário um maior conhecimento no âmbito semântico/lexical.

A seguir, alguns dos principais sistemas desenvolvidos serão apresentados de acordo com sua língua de estudo.

3.2 Sistemas para o Inglês

3.2.1 Gildea e Jurafsky - 2002

Dentro do PNL a primeira tentativa automatizada para a APS se deu pelo aprendizado supervsivionado de [17], embasados pela *Framenet* como *corpus* de treinamento (neste momento o *PropBank* ainda não tinha sido criado). Este trabalho considerou a árvore sintática das sentenças e as *features* utilizadas no algoritmo de aprendizagem foram derivadas desta.

Os resultados obtidos indicam uma acurácia de 82% na classificação de papéis semânticos (com os limites dos argumentos delimitados como entrada) e 64,6% de precisão e 64% de cobertura na tarefa combinada (identificação + classificação).

Vale a ressalva que para o sistema desenvolvido, o número médio de sentenças disponíveis para cada palavra alvo, identificado como predicado, é de 34 instâncias.

Quando analisado no *PropBank*, o sistema obteve acurácia para classificação de 79,9% e 82% quando os argumentos foram manualmente delimitados e utilizaram-se informação de árvore sintática automática e manual respecitvamente. Já para a tarefa conjunta os valores foram 68,6% e 74,3% para precisão e 57,8% e 66,4% para cobertura.

3.2.2 Hacioglu - 2004

Melhor resultado da competição CoNLL-2004, este algoritmo obteve tal colocação utilizando como algoritmo de aprendizagem o *Support Vector Machines* (SVM). Sua proposta consistiu em utilizar *chunkers*[1] que carregam tanto informação sintática quanto semântica.

Os seguintes resultados, para a base de testes da mesma competição, foram obtidos:

Tabela 3.1: Resultados obtidos por Hacioglu - 2004

d **Tarefa**	**Precisão**	**Cobertura**	F_1
Combinada	72,43%	66,77%	69,49

Embora não tenha utilizado as informações de árvore completa, este proposta utilizou as informações de posição de orações.

[1] *Chunk* Em linguística, chunks são um grupo de palavras ou frases que podem ser compreendidos como uma unidade individual. São sintagmas que contém palavras relacionadas sintaticamente.

3.2.3 Punyakanokk et al. - 2005a

Neste artigo os autores analisam o impacto da utilização de árvores sintáticas e de *shallow parsing* conjuntamente com a identificação de orações.

Para tal, os autores executam diversos algoritmos já propostos utilizando informações completas da árvore sintática ou da *falsa árvore* gerada pela combinação de *chunkers* e orações identificadas.

Ao final dos testes, os autores identificam, que em cada uma das etapas, separadamente, os resultados obtidos pelas duas abordagens são parecidos. Porém, na etapa de poda, as informações completas da árvore sintática são importantes e geram menor propagação de erro ao final de todas as etapas.

Quando todas as etapas são analisadas, as informações completas possuem um valor médio de 14 e 5 pontos abaixo na medida F_1 quando são utilizados dados *gold* e dados extraídos automaticamente respectivamente. Os valores obtidos estão apresentados na tabela abaixo:

Tabela 3.2: Resultados obtidos por Punyakanokk et al. - 2005a

Dados	***Full***			***Shallow***		
	Precisão	**Cobertura**	F_1	**Precisão**	**Cobertura**	F_1
Gold	88,81%	89,35%	89,08	75,34%	75,28%	75,31
Auto	77,09%	75,51%	76,29	75,48%	67,13%	71,06

Além desta análise os autores utilizam uma combinação de dois *full parsers* [76, 77] para geração de um resultado combinado com 78,69 pontos de medida F_1 no conjunto de teste da competição CoNLL-2004.

3.2.4 Punyakanokk et al. - 2005b

O sistema de melhor resultado da CoNLL-2005 utilizou as etapas: 1) poda, 2) identificação de argumentos, 3) classificação de argumentos e 4) inferência para obter seus valores. Sua etapa de inferência consiste em avaliar a saída de vários classificadores conjuntamente com os *parsers* [76, 77].

Mediante ao crescrimento que a base lexical do *PropBank* teve em relação a sua versão prévia, os resultados obtidos, para tarefa combinada, pelos autores foram:

Tabela 3.3: Resultados obtidos por Punyakanokk et al. - 2005b

Base de Teste	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Test WSJ	82,28%	76,78%	79,44%
Test Brown	73,38%	62,93%	67,75%
Test WSJ+Brown	81,18%	74,92%	77,92

3.2.5 Pradhan - 2008

Este trabalho [25], utilizando SVM e *features* extraídas de árvores sintáticas obteve os seguintes resultados, quando utilizados os conjuntos de teste da CoNLL-2005:

Tabela 3.4: Resultados obtidos por Pradhan - 2008

Tarefa	**WJS**		**Brown**	
	$\boldsymbol{F_1}$	**Acurácia**	$\boldsymbol{F_1}$	**Acurácia**
Identificação	85,9	–	81,2	—
Classificação	—	92%	—	—
Tarefa Combinada	80	—	69,9	—

Os valores apresentados foram obtidos quando o autor utilizou um *parser* automático para estruturação sintática. Quando utilizado dados *gold* retirados da base de dados, os valores para identificação de argumentos e tarefa combinada obtiveram resultados superiores de aproximadamente 10 pontos na medida F_1 para ambos os *corpus*.

Diversos outros resultados são apresentados pelos autores.

3.2.6 Toutanova et al. - 2008

No mesmo ano que [25], Toutanova [78] propôs uma nova abordagem que obteve resultados semelhantes ao estado-da-arte naquele momento. O algoritmo de atribuição conjunta modela dependências entre as etiquetas dos constituintes e entre cada etiqueta e os atributos de entrada dos outros constituintes.

Os resultados obtidos, para tarefa conjunta e para o conjunto de treinamento do WSJ para a CoNLL-2005 foram de:

Tabela 3.5: Resultados obtidos por Punyakanokk et al. - 2005b

Base de Teste	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Test WSJ	81,90%	78,81%	80,32%

Para obtenção de seus resultados, foram utilizados os cinco melhores *parsers* de Charniak [77] e os autores defenderam que o maior gargalo para o desempenho obtido se dá pela não excelência dos *parsers* sintáticos disponíveis.

3.2.7 Palmer et al. - 2010

Apesar de não apresentar os melhores resultado à época o trabalho de [26] merece destaque por servir como uma ampla fonte de informação a respeito de toda tarefa de APS.

As autoras fazem um apanhado geral a respeito a tarefa de APS, recursos lexicais, diferentes propostas de aprendizado de máquina para APS e sobre a perspectiva multilínguas.

3.2.8 Collobert e Weston - 2011

Diferentemente do que se era mais utilizado á época (aprendizado sobre a árvore sintática), os autores propuseram uma arquitetura, denominado *SENNA* [35, 79], que se baseia em classificadores de palavras.

O SENNA não deve ser entendido como um sistema para APS, e sim como uma arquitetura, pois ele possui diversas finalidades, tais quais:

- **POS Tagging**,
- **Chunking**,
- **Reconhecimento de Entidade Nomeada**,
- **Anotação de Papéis Semânticos**: considerada pelos autores a tarefa mais difícil de ser realizada,
- **Modelos Línguais**: predição da próxima palavra ser w em uma sentença de entrada.
- **Reconhecimento de palavras semanticamente correlatas**: sinônimos, antônimos, hiperonímias, entre outros.

O funcionamento do SENNA consiste em vetorizar palavras através de uma rede neural de convolução, onde nenhum conhecimento sintático é requerido para esta abordagem. Para a vetorização, foram utilizadas 130.000 palavras originárias da *corpora* da *Wikipedia* e *Reuters*.

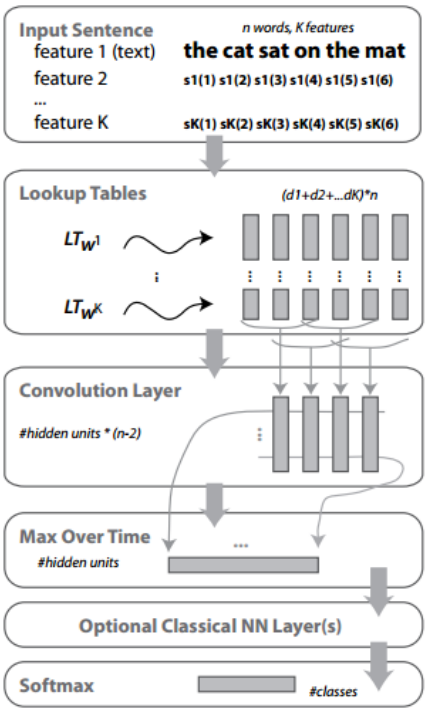

Figura 3.2: Arquitetura do SENNA

.

Diferentemente dos outros sistema, onde o classificador binário tinha como uma entrada um constituinte da árvore sintática, o SENNA precisa delimitar os argumentos e tratá-los como *tokens* (palavras) separadamente um-a-um.

Quando a tarefa de anotação de papéis semântico foi executada conjuntamente com a tarefa de modelos línguais a taxa de erro foi de 14,30%, a menor obtida até então para as seções de testes do *PropBank*.

O valor obtido para APS para a CoNLL-2005 foi de 75,49 pontos, e apesar de não possuir o resultado para o estado-da-arte, seu tempo de execução e memória utilizada são reduzidos quando comparado à outros sistemas de melhores resultados.

Tabela 3.6: Comparação sobre recursos utilizadas entre sistemas de APS.

Sistema	**RAM(MB)**	**Tempo(s)**
Punyakanokk et al. - 2005b	3400	6253
SENNA	124	51

3.2.9 Li e Chang - 2015

Os autores propõe uma rede neural recursiva, utilizando informações sintáticas da sentença, para a tarefa de APS. A vetorização de palavras também é realizada neste trabalho, porém, a escala da quantidade de *tokens* distintos é bem menor que a utilizada por [79]. Apenas 18,551 palavras foram utilizadas do *corpus Giga Word*.

O melhor resultado obtidos pelos autores foi de 72,27 pontos na medida F_1. Ao comparar os resultados com outros sistemas, mesmo não obtendo o resultado estado-da-arte, o sistema proposto obteve tempo de execução mínimo, vide a tabela abaixo:

Tabela 3.7: Comparação sobre recursos utilizadas entre sistemas de APS.

Sistema	**RAM(MB)**
Punyakanokk et al. - 2005b	6253
SENNA	51
Sistema Proposto	3

Diversos outros sistemas foram implementados para outras línguas senão o inglês, tais como: Alemão [80], Chinês [81, 82], Coreano [83, 84], Catalão [85], Francês [86, 87], entre outras línguas, como Sueco, Holandês e Árabe.

3.3 Sistemas para o Português

Comparado ao inglês, a abrangência de trabalho para o português é mínima, e este é um dos fatores que determinaram a escolha do tema dessa dissertação. Para o português, os sistemas já propostos baseiam-se em: sistemas baseados em regras, sistemas de classificação automática por aprendizado de máquina, por constituintes, por palavras, entre outros.

Neste seção, estão descritos os principais trabalhos para o português.

3.3.1 Bick - 2000 , 2006 e 2007

Em seu trabalho [39–41], Bick, baseado em seu sistema PALAVRAS, manualmente construiu um conjunto de 500 (quinhentas) regras gramaticais de mapeamento, além de outras regras para desambiguação. Também explorou as relações de dependências e funções sintáticas e de protótipos de classes semânticas.

Considerou-se um total de 38 categorias semânticas distintas, entre elas: agente, paciente, localização, papel, origem temporal, entre outros, e analisou-se os dados para português europeu.

A figura 3.3.1 apresenta um exemplo de regra desenvolvido por Bick. Define-se a regra de mapeamento do papel de agente (§AG) em sujeitos que apontam para direita se eles são humanos e seguidos – sem interferência de uma cláusula de borda - por um verbo principal. Este verbo principal é um verbo de movimento ou pertence a uma classe de valência transitiva.

```
MAP (§AG)
    TARGET @SUBJ>
    (0 HUM)
    (*1 @MV BARRIER CLB
    LINK 0 VT-ALL OR V-MOVE/TR
    LINK NOT 0 PAS) ;
```

Figura 3.3: Exemplo de regra de anotação de Bick

.

Ao final de seus estudos, analisando 2.500 palavras da base Floresta Sintá(c)tica1 treebank [88], obteve uma cobertura média de 86,6% e uma precisão média de 90,5%. Tais resultados são referentes a uma versão do *corpus* que não foi disponibilizada com anotação de papéis semânticos revisada. Também é importante ressaltar que o conjunto próprio de papeís semânticos atribuídos e a quantidade de dados para análise podem ter influenciado nos resultados.

As desvantagens deste tipo de abordagem consistem na necessidade de prever as diferentes formas de categorizar o papel semântico e criar manualmente regras cada uma das regras heuristicamente identificadas. Desta forma, a escalabilidade do sistema é comprometida, além da dificuldade em transportar as regras definidas utilizando língua alvo, no caso português, para outras línguas.

Outra desvantagem, agora inerente ao *parser* PALAVRAS propriamente dito é sua licença de uso. Mesmo obtendo um bom desempenho, sua licença tem custo bastante elevado.

O *parser* PALAVRAS está disponível para utilização[2], porém não pode ser reproduzido ou distribuido em parte ou todo sem consentimento da Universidade do Sul da Dinamarca:

3.3.2 Amancio - 2009

O autor, em sua dissertação de mestrado [89], pretendia avançar a área de elaboração textual para o português, que consiste em um conjunto de técnicas para acrescentar material redudantes em textos a fim de auxiliar a compreensão deste, principalmente por analfabetos funcionais.

Para tal o autor se comprometeu ao estudo de dois módulos distintos: 1) *REMET* responsável por trazer definições sobre entidades mencionadas na sentença e 2) *PET* responsável por prevê o uso de perguntas elaboradas, tais como: Quem?, Como? Onde? Com o quê?. Tal tarefa aproxima-se da anotação de papel semântico pois identifica os argumentos dos predicados a fim de possibilitar tais perguntas.

[2]http://visl.sdu.dk/

Utilizando um *corpus* de apenas 104 artigos de notícias (2.184 sentenças) em português brasileiro, retirados do jornal Zero Hora, executou uma simplificação textual atravéz do projeto PorSimples[3], a fim de facilitar sua compreensão por um maior número de leitores. Posteriormente, executa uma anotação manual quanto às perguntas elaboradas da seguinte forma:

Delimitação		**Categorização**	**Elaboração**
Verbo	**Segmentos de Resposta**	**Categorias**	**Pergunta Elaborada**
Avisou	"Nelson Hubner"	Quem?	Quem avisou?
	"Que o governo pensa em elevar a taxa para 3% de forma autorizada."	O que?	Avisou o que?
	"Ontem"	Quando?	Avisou quando?
Pensa	"Em elevar a taxa para 3 % de forma autorizada"	Em que?	Pensa em que?
Elevar	"A taxa"	O que?	Elevar o que?
	"Para 3 %."	Para quanto?	Elevar para quanto?
	"Ainda em 2008."	Quando?	Elevar quando?
	"De forma autorizada, isto é, não obrigatória."	Como?	Elevar como?

Figura 3.4: Exemplo de anotação de sentença de Amancio. **Sentença**: Ontem, Nelson Hubner avisou que o governo pensa em elevar a taxa para 3% de forma autorizada.

.

Tal *corpus*, após o *parsing* realizado pelo PALAVRAS, foi utilizado para o treinamento de um classificador embasado em SVM, que obteve resultado de 79 pontos na medida F_1.

3.3.3 Sequeira et al. - 2012

Os autores em seu trabalho [43] propuseram uma abordagem para anotação de papéis semânticos utilizando sentenças extraídas do *corpus Bosque 8.0* (Português - Portugal).

Os autores, utilizando as sentenças e o *parser* PALAVRAS, criaram o *BosqueUE*, um *corpus* semanticamente anotado quanto três categorias: predicado e argumentos do tipo Arg0 e Arg1.

[3]http://www.nilc.icmc.usp.br/simplifica/sobre.php

Utilizando dois algoritmos de aprendizado diferentes, *Contidional Markov Models, trained with Support Vector Machines (SVMCMM)* e *Conditional Random Fields (CRF)*, avaliaram os resultados tanto no *BosqueUE* quanto no conjunto de treinamento da competição CoNLL-2004.

Os melhores resultados, obtidos pelo CRF, estão apresentados abaixo:

Tabela 3.8: Resultados obtidos por Sequeira et al. - 2012 para o BosqueUE

Tag	**BosqueUE**	**Brown**
P	52,7	83,6
Arg0	31,1	52,3
Arg1	19,0	23,4

Um dos motivos que os autores questionam que pode ter influenciado nos resultados para o BosqueUE é o fator desse ter apenas um terço do tamanho do conjunto de teste da CoNLL-2004.

3.3.4 Manchego - 2013

Alva Manchego em sua dissertação de mestrado [1] apresentou 3 propostas diferentes, baseadas em aprendizado supervisionado, não-supervisionado e semissupervisionado. Como o próprio autor esclarece que os resultados obtidos no aprendizado não-supervisionado não podem ser utilizados para comparação aos outros sistemas, o método não será abordado nesta dissertação.

Proposta Supervisionada

O próprio autor afirma que desenvolveu a primeira abordagem de aprendizado de máquina disponível para o português do Brasil. Utilizando o corpus PropBank-BR (v1.0), Alva Manchego criou um sistema supervisionado que extrai atributos dos constituintes das sentenças.

Esse sistema dividiu-se em 4 (quatro) etapas:

1. **Identificação do verbo**: cujo qual foi retirado automaticamente da base de dados.
2. **Poda**: Seguindo as diretrizes de [73]
3. **Identificação de argumentos**: Classificador binário para cada constituinte da árvore sintática.
4. **Classificação de argumentos**: Classificador com *features* próprias, diferente do anterior.

Os classificadores utilizados pelo autor utilizam o algoritmo de regressão logísica (RL), em um subconjunto de atributos definidos por: [17, 25, 78] e outros. São exemplos de *features* utilizados:

- **Caminho**: Caminho através da árvore sintática do verbo até o constituinte alvo analisado. Diversos autores indicam que este atributo possui grande relevância na etapa de identificação de argumentos.
- **Distância em Constituintes na Árvore**: número de constituintes que separam o constituinte alvo analisado até o verbo da sentença.
- **Núcleo, Lema do Núcleo, PoS do Núcleo**: Análise do núcleo do constituinte alvo analisado.
- **Primeira e Última Palavras / PoS do Constituintes**: Primeira e última palavra do constituinte acrescido de sua etiqueta de part-of-speech.
- **Tipo de Sintagma**: Análise do constituinte alvo, quanto sua categoria que pode ser: sintagma nominal, sintagma preposicional, etc.
- Outras features como: Contexto do Predicado, Palavras do Constituinte, Parentes do Constituinte, Pontuação, Voz, etc.

Como tratou-se do primeiro sistema, também desenvolveu um sistema baseline baseado em 6 regras adaptadas das Shared Task do CoNLL, para servir como comparação com o sistema supervisionado treinado. O baseline foi criado especificamente para a identificação das etiquetas A0, A1 e AM-NEG, por serem as etiquetas com mais ocorrência na base de dados. Desta forma, os resultados obtidos pelo sistema baseline e a abordagem supervisionadas são apresentados na tabela abaixo.

Tabela 3.9: Resultados obtidos por Manchego - 2013 para o *baseline* proposto e para o classificador supervisionado.

Etiqueta	***Baseline***			**RL**		
	Precisão	**Cobertura**	F_1	**Precisão**	**Cobertura**	F_1
Todas	64,6%	40,9%	50,1	80,0%	79,3%	79,7
A0	49,7%	7,9%	58,5	90,8%	79,8%	85,0
A1	79,4%	53,1%	63,6	87,6%	90,1%	88,8
AM-NEG	90,5%	100,%	95,0	95,0%	100,0%	97,4

O sistema de aprendizado supervisionado superou o baseline em todas as categorias, porém, só se aproximou ao estado-da-arte dos sistemas previamente desenvolvidos para o inglês por outros autores na tarefa de identificação dos argumentos, fato que

não acontece para a tarefa de classificação dos argumentos. Alva Manchego explicita três razões para tal ocorrência: (1) Um possível mal conjunto de features utilizado e (2) A insuficiente quantidade de dados anotados para generalização do aprendizado e (3) O identificador anota somente classes binárias (ARG - NULL), enquanto o classificador precisa lidar com uma classificação multiclasse.

Quando utilizou umas das *features* criadas para sua abordagem não-supervisionada, seus resultados melhoraram. A *feature* **Função Sintátia** extrai a relação de dependência do núcleo de um constituinte com o seu regente. Desta forma, para o algoritmo supervisionado, seus melhores valores foram:

Tabela 3.10: Melhores resultados obtidos por Manchego - 2013 para o classificador supervisionado

Tarefa	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$	Acurária
Identificação	94,9%	93,7%	94,3	-
Classificação	-	-	-	85,5%
Conjunta	83,0%	81,7%	82,3	-

Vale ressaltar que o autor **não identifica predicados automaticamente para novos textos**, nem apresentou resultados obtidos por árvores geradas automaticamente, ao invés de extraídas em formato *gold* da base de dados.

Proposta Não-Supervisionada

Abaixo, dois motivos principais para a proposta se valer à critério de comparação:

- O método supervisionado considerou anotação sintática por constituintes, enquanto os métodos de IPS consideraram anotação sintática por dependente. Este fato implica que o método supervisionado anota papéis semânticos por conjunto de palavras (ou argumento), enquanto o IPS atribui papel semântico somente ao núcleo do argumento.

- Os métodos supervisionados atribuem uma etiqueta de papel semântico a cada argumento, enquanto os métodos de IPS só indicam o *cluster* ao qual cada argumento pertence, não atribuindo nenhuma etiqueta presente na base de dados.

Proposta Semissupervisionada

Para verificar sua principal hipótese, de "que é possível empregar técnicas de aprendizado de máquina semi-supervisionado para anotar automaticamente com papéis semânticos sentenças escritas em português do brasil com um desempenho comparável

ao de um anotador supervisionado para a mesma língua.", o autor empregou o *self-training* com modificações para minimizar o problema sofrido com o desbalanceamento dos dados disponíveis para o português, já comentados anteriormente.

Selecionando arbitrariamente as 1.000 primeiras sentenças dos corpus de treinamento, para servir como dados anotados, e selecionando o restante como dados não anotados, Manchego verifica que a proporcionalidade de frequência dos papéis semânticos é mantida, enquanto a diversidade dos verbos presentes é significativamente menor.

Manchego constrói uma abordagem semi-supervisionada pelo o algoritmo self-training. Este algoritmo utiliza os próprios resultados para aprendizado em um processo iterativo. O funcionamento principal deste algoritmo consiste em:

- Treinar um conjunto de dados anotados por uma abordagem supervisionada.
- Aplicar o classificador obtido para categorizar dados não anotados.
- Selecionar um novo conjunto de informações, que tenham obtido um certo grau de confiança quanto sua classificação, e inseri-lo no conjunto de dados anotados.
- Repetir o processo até uma condição de parada ser satisfeita.

Após a execução mais simples do algoritmo, Manchego realizou algumas modificações para tentar obter melhores resultados. Dentre elas:

- **Condição de parada simplificada**: Removeu-se o número máximo de iteração consecutivas sem modificação do classificador.
- **Balanceamento dos dados**: Implementa uma função auxiliar que não permite que um alto número de instâncias anotadas para um mesmo papel semântico seja adicionada de uma vez no corpus de treinamento.
- **Balanceamento auxiliado por similaridade**: Considera a similaridade que existe entre o candidato avaliado e os registros pertencentes ao conjunto de retreinamento disponível.

Os resultados finais estão apresentados na tabela abaixo:

Tabela 3.11: Melhores resultados obtidos por Manchego - 2013 para o classificador semissupervisionado

	Tarefa		
	Identificação	**Classificação**	**Conjunta**
Abordagem	$\boldsymbol{F_1}$	Acurácia	Conjunta
Classificação Supervisionada	94,3	85,5%	82,3
Self-Training simples	94,2	83,0%	79,6
Self-Training alterado	94,2	83,2%	80,5

A abordagem semi-supervisionada não foi capaz de superar significativamente a abordagem supervisionada, porém, seus resultados são semelhantes. Uma das conclusões do autor é o fato que estimar os parâmetros de entrada para o sistema semi-supervisionado é difícil.

3.3.5 Fonseca - 2013

Partindo da hipótese de que era possível desenvolver um sistema de anotação de papéis semânticos para portuguêsm utilizando aprendizado de máquina e sem a utilização de atributos originários de ferramentas de NLP externas, como um parser sintático, Fonseca [4] utilizou a representação de palavras em espaço vetorial como insumo de dados para o treinamento de uma rede neural com uma camada de convolução, inspirada na arquitetura desenvolvida nos trabalhos de [35, 36], conhecido como SENNA.

Para a vetorização de palavras, iniciou seus estudos realizando um teste comparativo entre os resultados obtidos pelas abordagens: Hyperspace Analogue to Language (HAL) [90, 91], Random Indexing (RI) [92] e um modelo neural similar aos dos estudos de [35]. Os dados utilizados para o treinamento foram obtidos do Wikipedia[4] e do corpus PLN-BR [42], fonte de dados para textos jornalísticos extraídos da Folha de São Paulo[5].

Selecionando aleatoriamente 10 palavras de sua base de dados, Fonseca classificou manualmente o conjunto de palavras semelhantes fornecido como resposta de cada uma das abordagens. Os modelos HAL e RI foram os que obtiveram os melhores resultados. Como a dimensionalidade do modelo HAL era de 40 (quarenta) vezes menor que o do RI, o HAL foi selecionado como abordagem a ser utilizada no trabalho.

A análise do sistema completo foi realizada utilizando o *PropBank-BR*. Para todas

[4]Disponível em: https://pt.wikipedia.org/
[5]Disponível em: http://www.folha.uol.com.br/

as etapas foram analisados os resultados com a inserção ou não de *chunks* sintáticos. A utilização dos *chunks* impede que o sistema seja executado sem a necessidade de um parser sintático utilizado conjuntamente.

Ao final de diversas parametrizações dos dados de entrada, os resultados obtidos estão descritos na Tabela 3.12.

Tabela 3.12: Resultado do sistema supervisionado sem identificação automática de predicados.

Tarefa	**Precisão**	**Cobertura**	F_1	**Acurácia**
Identificação (sem *chunks*)	76,42%	72,44%	74,38*	-
Identificação (com *chunks*	80,15%	78,96%	79,17	–
Classificação (sem *chunks*)	–	–	–	88,14%*
Conjunta (sem *chunks*)	67,06%	63,31%	65,13*	–
Conjunta (com *chunks*	68,97%	67,04%	67,99	–

O autor decide não utilizar *chunks* sintáticos na classificação dos argumentos pois acredita que a estrutura sintática da sentença não é tão importante para esta tarefa quanto é para a identificação de argumentos.

O resultado apresentado acima desconsidera a identificação automática dos predicados. Quando este reconhecimento é realizado pelo próprio sistema, os resultados obtidos são piores, como descritos na Tabela 3.13.

Tabela 3.13: Resultado do sistema supervisionado com identificação automática de predicados.

Tarefa	**Precisão**	**Cobertura**	F_1
Identificação de predicados	88,62%	91,21%	89,90*
APS Completa (sem *chunks*)	66,95%	58,10%	62,21*
APS Completa (com *chunks*)	69,01%	62,21*%	65,43

O sistema desenvolvido por Fonseca é o único capaz de, para o português, analisar integralmente um texto puro em português fornecido pelo usuário, porém neste

caso, a utilização dos chunks sintáticos fica comprometida. Seu sistema NLPNET[6] encontra-se disponível para download.

Obs.: Todos os valores anotados com asterico (*) serão novamente discutidos na seção 5

3.3.6 Hartmann - 2015

Em seu trabalho de mestrado, conveniado com a empresa SAMSUNG, Hartmann [2] diferentemente de outros autores não propôs um novo sistema puramente baseado em aprendizado de máquina para a tarefa de anotação dos papéis semânticos. Seu trabalho propôs um sistema híbrido de anotação de papel semântico utilizando tanto regras quanto aprendizado de máquina.

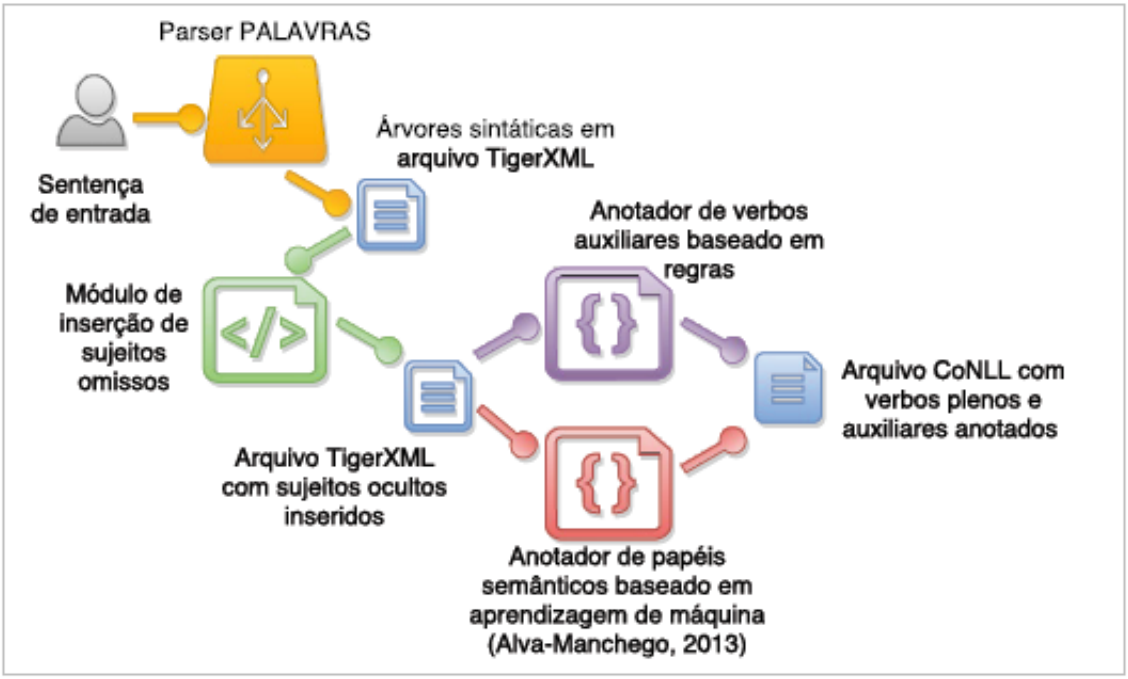

Figura 3.5: Sistema híbrido desenvolvido por Hartmann, utilizando o parse PALAVRAS e o sistema de Manchego - 2013. **Fonte:** [2]

.

Analisando os resultados de seu projeto tanto pelo sistema supervisionado de [1], quanto o trabalho realizado por [4], Hartmann propôs as seguintes hipóteses:

- Explicitar sujeitos ocultos melhora a tarefa de APS.
- A anotação de verbos auxiliares com papéis semânticos modificadores por meio

[6]Disponível em: https://github.com/erickrf/nlpnet

de regras possui precisão similar à obtida por um sistema de aprendizado de máquina.

- O uso do sentido do verbo alvo, melhora a medida F1 de um sistema de anotação de papel semântico.
- É melhor treinar um sistema de APS utilizando árvores sintáticas não revisadas ao invés das revisadas, uma vez que o cenário real consiste na primeira opção.
- Existe uma degradação de 10% na medida F1, quando é realizada uma troca de gênero dos textos utilizados nas etapas de treinamento e teste.

Além do PropBank-BR como fonte de dados para aprendizado de máquina, Hartmann utilizou o corpus PNL-BR [42] – textos de origem jornalística - e o corpus Buscapé [61] – textos de opinião sobre produtos coletados da web - anotados manualmente por uma equipe de 7 anotadores e 1 adjudicador[7], e 3 anotadores e 1 adjudicador respectivamente para cada um dos corpus.

Cada um desses dois corpus foi separado em dois conjuntos: Concordância (quando todos os anotadores concordaram com a identificação e classificação dos argumentos) e Adjudicados (quando foi necessário o voto de minerva).

Dentre diversos testes de hipótese realizados, utilizando o teste de Wilcoxon [93], com confiança de 95% (noventa e cinco) por Hartmann, houve as seguintes conclusões:

✓ Há evidências de que há diferença estatística significativa entre os resultados obtidos pela inserção ou não dos sujeitos, somente nas primeiras pessoas, no corpus do Buscapé.

✓ Não há evidências para que se possa afirmar que há diferença entre os desempenhos dos sistemas treinados incorporando o sentido do verbo alvo e aquele que não o faz. Os testes foram realizados utilizando o corpus PNL-BR, PropBank-BR e a mescla deles.

✓ Sistemas treinados sobre árvores sintáticas perfeitas possuem maiores dificuldades ao anotar árvores sintáticas não revisadas, tanto quando são analisados textos de origem jornalística (corpus PNL-BR e PropBank-BR), quanto quando foi analisado o corpus Buscapé.

× Não há evidências para rejeitar a hipótese nula de que não há diferença significativa entres os desempenhos dos sistemas treinados sobre o PLN-BR (corpus jornalístico), utilizando ou não a inserção do sujeito.

[7]Retirado do campo jurídico, adjudicar neste contexto refere-se em resolver o processo de ambiguidade formada pelas diferentes perspectivas de dois ou mais anotados quanto ao processo de anotação de papel semântico.

× Não foi possível afirmar que a queda de desempenho ao efetuar a troca de gêneros dos textos das bases de dados nas etapas de treinamento e teste é de aproximadamente 10%, uma vez que o autor somente obteve uma perda de aproximadamente 3,8% na medida F1.

Ao final do seu trabalho, Hartmann concluiu que dentre suas principais limitações estava o desempenho obtido pelo parser PALAVRAS, pois:

× Impactou na inserção dos sujeitos das terceiras pessoas. Errar a pessoa gramatical do verbo, rotular o sujeito como verbo e construir indevidamente as árvores sintáticas foram alguns dos erros mais comuns.

× Inviabiliza a desambiguação do sentido do verbo onde não se possua uma alta frequência do verbo para cada um dos seus sentidos.

× Os erros obtidos pelo parser são carregados na etapa de anotação de papel semântico.

3.4 Resumo dos Trabalhos para Português

Após a revisão dos principais trabalhos realizados para o português, a tabela 3.14 visa apresentar uma visão geral dos trabalhos propostos. Foram considerados os trabalhos que mais se aproximam do objetivo desta dissertação:

Tabela 3.14: Resumo dos trabalhos realizadas para o português

	Trabalho			
Quesito	**Bick**	**Manchego**	**Sequeira**	**Fonseca**
Embasamento Teórico	Regras Lingísticas	Árvores sintáticas e regressão logística	SVMCMM CRF	Vetorização de palavras e RNA de convolução (SENNA)
Método para Resolução	Construção de regras manuais	Aprendizado Supervisionado	Aprendizado Supervisionado	Aprendizado Supervisionado
Base de dados	Privado e não revisado	*PropBank-BR v1.0*	Bosque UE (PT-PT) CoNll 2004	*PropBank-BR v1.0*
Melhor Resultado	Pr: 90,5% Cob: 86,6% F1: 88,50	Pr: 93,0% Cob: 81,7% F1: 82,3	Pr: 41,5% Cob: 29,4% F1: 34,3	Pr: 66,95% Cob: 58,10% F1: 62,21
Implementação Disponível	licença Privada	Não encontrado	Não encontrada	Sim

Além de todos os trabalhos apresentados para o portuguès, vale comentar sobre o trabalho de [94] que investiga, entre outras coisas, anotação de papéis semânticos no contexto do aprendizado semissupervisionado baseado em grafos.

Seus resultados são promissores, porém como foram avaliados em um conjunto muito pequeno de dados, ainda não é possível definir uma qualidade geral do seu trabalho para, especificamente, a tarefa de APS.

3.5 Considerações Finais

Até a presente data a anotação de papéis semânticos está em voga no cenário mundial. Em todo mundo empresas de grande renome patrocinam estudos e pesquisar a fim de aprimorar os resultados das propostas dos estudiosos, empresas como *Samgsung*[8], *Google*[9], *Intel*[10], *Microsoft*[11], entre outras.

A visão que se tem em buscas sobre artigos da área apresentam que a área está se difundindo para diversas línguas com a fomento da criação de ferramentas,

[8]http://www.samsung.com/
[9]http://www.google.com
[10]http://www.intel.com
[11]https://www.microsoft.com

arquiteturas e bases auxiliares para o aprendizado. E assim, neste capítulo foram apresentados alguns dos principais trabalhos em APS, focando em trabalhos para o inglês e para o português.

Capítulo 4

Conditional Random Fields

Este capítulo apresenta o *Conditional Random Fields*, principalmente o *Linear-Chain CRF* e sua motivação a partir do *Hidden Markov Model* (HMM), como um modelo discriminativo desenvolvido a partir de modelos de grafo que se adéqua à problemas onde a natureza sequencial está envolvida.

Além desta apresentação, o capítulo descreve os principais motivos para a seleção do CRF em detrimentos à outros modelos/algoritmos para a tarefa de APS,

4.1 Por que CRF para APS?

A escolha pelo modelo de aprendizado seguiu uma linha de raciocínio criteriosa. A busca consistiu em procurar por abstrações que se adequassem ao problema enfrentado e que possuíssem um nível de custo/complexidade razoável e factível. Ao fim, a escolha pelo modelo *Conditional Random Fields* (CRF) se deu por diversas razoes.

4.1.1 Caráter Sequencial

A primeira e principal razão pela escolha do CRF foi seu caráter de classificação sequencial. Uma vez que diversas etapas da tarefa de APS possuem natureza sequencial (identificação de *chunks*, identificação de argumentos, etc.), é interessante que o modelo de aprendizagem também apresente, inerentemente, tal característica, pois ao ignorar este aspecto, muita informação pode ser perdida. Devido a este fato, propostas de solução como a realizada por [1] (Regressão Logística), que não apresentam inerentemente o reconhecimento de classes por sequência, foram desconsideradas.

4.1.2 Modelo Discriminativo

Modelos de aprendizado podem ser agrupados em dois grandes grupos: 1) Generativos e 2) Discriminativos. Esses grupos diferem quanto sua intenção e formas onde devem ser empregados. Considere uma tarefa onde se deseja determinar a língua que um indivíduo está falando. Desta forma:

- **Abordagem Generativa**: Tenta aprender todas as línguas do domínio onde a tarefa se encontra e determina em qual delas o indivíduo está falando.
- **Abordagem Discriminativa**: Tenta determinar a língua falada de acordo com as caraterísticas da língua falada.

Modelos Generativos são tais que sua função de otimização atuam sobre a probabilidade conjunta $P(X, Y)$, ou seja, são modelos que procuram encontrar o conjunto de parâmetros que maximizam a probabilidade de ocorrência da base de dados de treinamento (Equação 4.1). Para problemas de classificação, etapas auxiliares são necessárias, como estimador de máxima verossimilhança (*maximum likelihood estimation*) ou a aplicação da regra de Bayes.

$$f(x) = argmax(y)P(X|Y)P(Y) = argmax(y)P(X, Y) \tag{4.1}$$

Além de outras características, modelos generativos também podem ser utilizados para a geração de dados sintéticos devido às suas características. Exemplo de modelos: Naive Bayes, Hidden Markov Models (HMM), etc.

Já modelos Discriminativos se propõem a obter diretamente a distribuição $P(Y|X)$ (Equação 4.2). Como o problema de classificação, em sua maioria das vezes, deriva da necessidade pela obtenção desta distribuição, modelos discriminativos não necessitam de etapas auxiliares, como a aplicação da regra de Bayes, para sua resolução. Exemplo de modelos: Regressão Logística, *Conditional Random Fields* (CRF).

$$f(x) = argmax(y) \prod_i P(Y|X) \tag{4.2}$$

Outra questão válida é que, para seu aprendizado, modelos generativos necessitam de grande quantidade de dados quando as tarefas se tornam mais complexas: estimação da $P(X)$, que mesmo que possa ser calculado marginalmente, pode se refletir difícil; ou para tentar reduzir o *overfitting*, uma vez que sua proposta é maximizar a ocorrência da base de treino.

Outras questões como o fato de métodos discriminativos se comportarem melhores perante a dependência entre *features* também foram analisadas. Desta forma, a escolha por um algoritmo discriminativo foi outro ponto importante para a seleção do CRF.

4.1.3 Modelo Estado-da-Arte

O CRF se apresenta como uma das mais novas abordagens para resolução de diversos problemas como: *POS tagging* [95], reconhecimento de entidade nomeada [96, 97], recuperação de documento [98], reconhecimento de imagem [99], entre outros.

Diferentemente de outras abordagens como Regressão Logística ou SVM, a tarefa de APS completa não apresenta tantos trabalhos que se propõem em resolvê-la pela utilização de CRF. Desta forma, sua escolha também foi realizada pelo caráter exploratório do estudo.

4.1.4 Caráter Multilingual

Por fim, a capacidade do modelo CRF ser aplicável para diversas línguas também foi um fator determinante. Uma vez que, para a resolução do problema, o CRF não depende da utilização de *features* que sejam inerentes somente ao português, sua utilização deixa margens para reutilização em outras línguas.

4.2 Definição

Conditional Random Field refere-se a distribuição condicional $P(Y|X)$ com uma estrutura de grafos associadas. Para um conjunto de varíaveis aleatórios $V = X \cup Y$, onde X é um conjunto de variáveis de entradas observadas e Y é um conjunto de varíaveis de saída, e um subconjunto $A \subset V$, define-se um modelo de grafo não-direcional da seguinte forma:

$$p(x, y) = \frac{1}{Z} \prod_{A} \Psi_a(x_a, y_a) \tag{4.3}$$

Que pode ser escrito para qualquer escolha de fatores $F = \{\Psi_A\}$, denominada função característica. E onde Z é um fator de normalização, utilizado para garantir que a distribuição some 1, e normalmente denominado de função de particionamento, definido como:

$$Z = \sum_{z,y} \prod_{A} \Psi_a(x_a, y_a) \tag{4.4}$$

A equação 4.3 pode ser entendida como um grafo de função, ou seja, um grafo bipartido $G = (V, F, E)$ cujos nós da variável $v_s \in V$ está conectado à um nó de função $\Psi_A \in F$ caso v_s seja argumento para Ψ_A. Na figura abaixo exemplifica-se

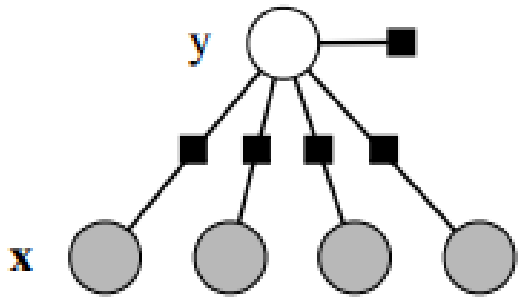

Figura 4.1: Exemplo de grafo de função. Círculos são nós de variável e quadrados são nós de função. Fonte: [3]

A questão principal para utilização desta abordagem é que com o modelo de grafos é possível representar a distribuição de uma grande quantidade de variáveis aleatórias pelo produto de funções locais, que dependem de um pequeno número de variáveis [3].

Para modelos de PNL, as variáveis X (entrada do modelo) e Y (saída do modelo) podem ser entendidas como a sequência de palavras de uma sentença e a sequência de classes (etiquetas) de cada uma destas palavras, respectivamente.

4.2.1 *Linear-Chain CRF*

O *Linear-Chain CRF* teve sua motivação no *Hidden Markov Model* (HMM). Seu conceito pode ser descrito como extrair a probabilidade condicional da probabilidade conjunta do HMM. Uma das suas características mais marcantes está no fato de utilizar o resultado da entrada imediatamente anterior para a classificação do registro em análise, ou seja: Dado que um $registro_{k-1}$ foi marcado com a etiqueta A, para a marcação do $registro_k$ a etiqueta A do registro anterior é um dado de entrada.

O HMM pode ser escrito da seguinte forma:

$$p(y,x) = \prod_{t=1}^{T} p(y_t|y_{t-1}P(x_t|y_t)) \tag{4.5}$$

Onde, $p(y_t|y_{t-1})$ reflete a probabilidade de transição dos estados e $P(x_t|y_t)$ consiste na probabilidade de emissão. A equação 4.5, pode ser reescrita na forma:

$$p(y,x) = \frac{1}{Z} exp\big(\sum_{t}\sum_{i,j\in S} \lambda ij 1_{y_t=i}1_{y_{t-1}=j} + \sum_{t}\sum_{i\in S}\sum_{o\in O} \mu_{oi}1_{y_t=i}1_{x_t=o}\big) \tag{4.6}$$

Onde:

- $\theta = \{\lambda_{ij}, \mu oi\}$

- $\lambda_{ij} = log\, p(y' = i|y = j)$
- $1_{x=x'}$ assume o valor 1 quando $x = x'$, 0 c.c.

Já a equação 4.6 pode ser escrita em função das funções características da seguinte forma:

$$p(y, x) = \frac{1}{Z} exp\big(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\big) \tag{4.7}$$

Assim, a equação 4.7, seguindo as regras de Bayes e da Marginalização define:

$$p(y|x) = \frac{p(y, x)}{\sum_{y'} p(y', x)} = \frac{exp\big(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\big)}{\sum_{y'} exp\big(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\big)} \tag{4.8}$$

A partir da motivação do HMM, o *Linear-Chain CRF* foi criado e definido da seguinte forma:

$$p(y|x) = \frac{1}{Z(x)} exp\big(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\big) \tag{4.9}$$

e

$$Z(x) = \sum_{y} exp\big(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\big) \tag{4.10}$$

A figura 4.2 reflete visualmente a interpretação do *Linear-Chain CRF* beaseada no HMM.

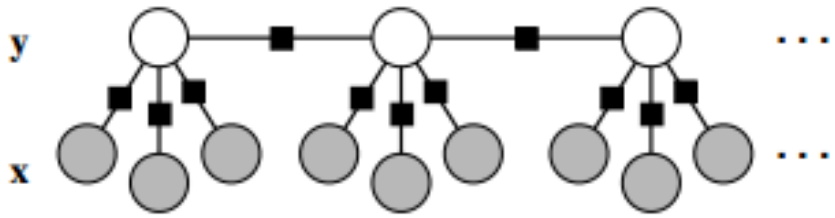

Figura 4.2: Modelo de grafo para o *Linear-Chain CRF* baseado em HMM. Fonte [3].

Podemos adicionar *features* ao CRF para que as observações atuais **x**, palavras da sentença para o caso de APS, sejam levadas em conta, desta forma, define-se a figura 4.3.

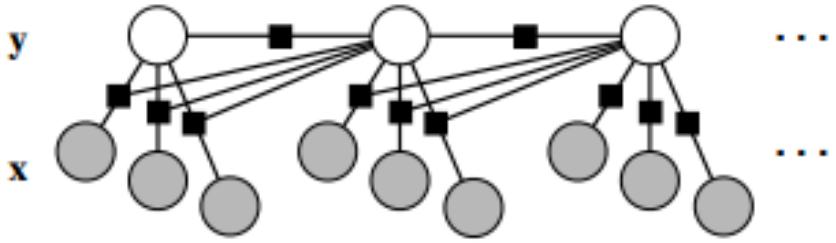

Figura 4.3: Modelo de grafo para o *Linear-Chain CRF*, com adição de *features* da observação atual. Fonte [3].

As funções características atuam de acordo com a construção da implementação do algoritmo do *CRF*. Já os pesos (λ) para cada uma das funções podem ser estimados de diversas formas. Os autores de [3] definem-na pela maximização da *contidional log likelihood*, da seguinte forma:

$$l(\theta) = \sum i = 1^{N} logp(y^i|x^i) \tag{4.11}$$

Outras formas de estimação são possíveis.

Diversas implementações, principalmente para o *Linear-Chain CRF* estão disponíveis na *internet*. Neste trabalho, após uma pesquisa de diversas formas, utilizou-se o CRF++[1] como solução, principalmente pelo fato de parecer ser uma das mais utilizadas, não só para o português, como para o inglês.

São exemplos de funções características, com *features* retiradas para o caso de *POS Tagging*:

```
func1 = if (output = B-NP and feature="U01:DT") return 1 else return 0
func2 = if (output = I-NP and feature="U01:DT") return 1 else return 0
func3 = if (output = O and feature="U01:DT") return 1  else return 0
....
funcXX = if (output = B-NP and feature="U01:NN") return 1  else return 0
funcXY = if (output = O and feature="U01:NN") return 1  else return 0
...
```

Figura 4.4: Exemplo de funções geradas por unigramas pelo CRF++.

4.3 Considerações Finais

Neste capítulo foram discutidas as questões que levaram à escolha do modelo para resolução do problema de anotação de papéis semânticos. Tais escolhas foram realizadas a partir de um conjunto bem definido de critérios e apesar do fato que possa

[1]Disponível em: https://taku910.github.io/crfpp/

haver modelos mais propícios, o CRF detém um conjunto de características válidas em sua natureza que o adequam para a tarefa.

Também foram apresentadas as relações entre modelos generativos e discriminativos e, por fim, o par HMM-CRF foi explicado por meio de detalhamento matemático. Maiores explicações do modelo apresentado podem ser encontrados em [3, 100].

Capítulo 5

Proposta de Anotador Semântico em Português por CRF

Neste capítulo estão descritos as etapas propostas para a realização da APS. Para este trabalho são consideradas 5 subtarefas para a completude da anotação semântica, tais tarefas são expandidas devido a necessidade de geração automática de informação (*POS Tag, Chunks*) para posterior utilização do sistema em textos puros.

Também são enumeradas todas as *features* utilizadas como insumo para as funções características a respeito dos modelos gerados para cada uma das tarefas realizadas.

5.1 Proposta

Como descrita anteriormente, a proposta dessa dissertação é apresentar um novo sistema para anotação de papéis semânticos de textos puros. O anotador semântico consiste na junção de 5 etapas, onde um texto puro é anotado sem a necessidade de qualquer informação prévia, senão o texto propriamente dito. As etapas consistem em:

1. POS Tagging,
2. Identificação de Predicados,
3. Identificação de *Chunkings* Sintáticos,
4. Identificação de Argumentos,
5. Classificação de Argumentos.

A figura 5.1 fornece uma visão geral do anotador semântico, sobre os aspectos de treinamento, avaliação e utilização.

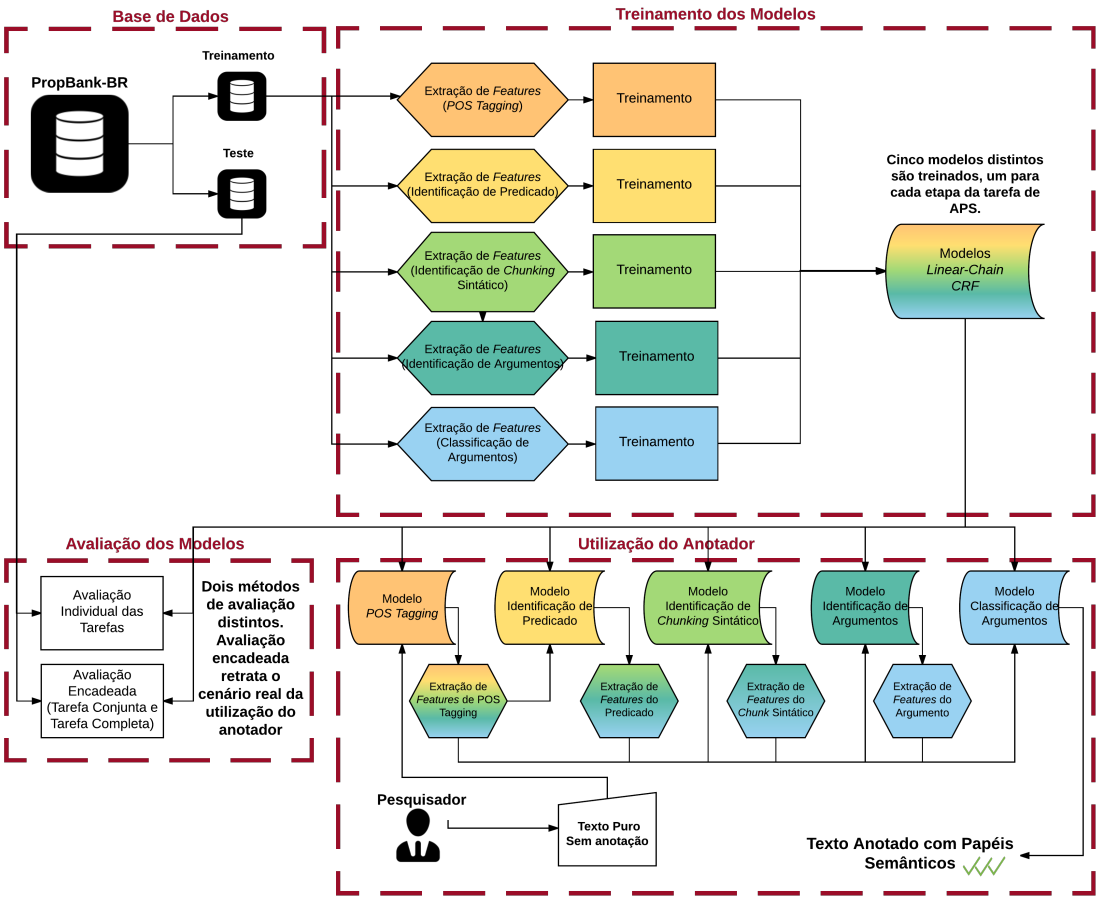

Figura 5.1: Visão geral da sequência de etapas realizadas para anotação de textos.

Como descrito acima, as subtarefas envolvidas, individualmente, são realizadas a partir do mesmo particionamento do *PropBank-BR* empregado por [4], que definiu os conjuntos de treinamento e teste. Desta forma, ao mensurar os conjuntos sem efetuar alterações, os resultados obtidos podem ser comparados com coerência ao *NLPNET*.

O treinamento para cada uma das subtarefas utiliza a porção de treinamento do *PropBank-BR*, e assim, são criados 5 modelos baseados no *Linear-Chain CRF* distintos e próprios. A figura 5.2 apresenta o processo comum de treinamento para cada uma das etapas[1].

[1] A coloração degradê das figuras indica que o processo descrito é comum a mais de uma etapa supracitada.

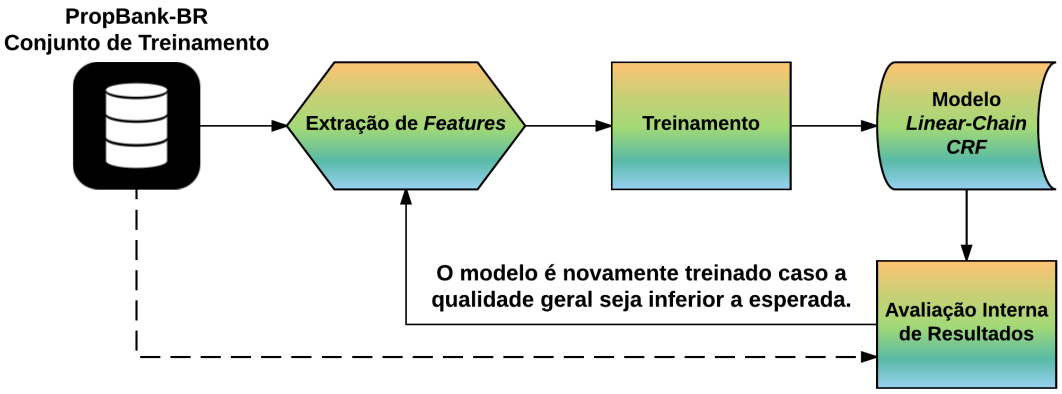

Figura 5.2: Fluxo realizado para o treinamento dos modelos CRF. O fluxo é análogo para todas as subtarefas.

É importante atentar que a avaliação dos resultados descrita na figura 5.2 é um análise dentro da amostra, ou seja, utiliza os próprios dados presentes no conjunto de treinamento para avaliação, e não os dados do conjunto de testes do *PropBank-BR*. A avaliação é realizada desta forma para o processo de aprendizagem não sofrer com o viés causado pela *data snooping*[2].

O treinamento individual de cada etapa não utiliza os dados oriundos da etapa anterior, e assim, os dados de entrada são retirados diretamente da base de dados (formato *gold*).

Já no processo de avaliação dos modelos, o conjunto de testes do *PropBank-BR* é utilizado. A partir desta seção é que serão descritos os resultados finais na seção 6. A avaliação é realizada de duas formas distintas:

- **Avaliação Individual**: A qualidade da subtarefa é avaliada individualmente, e não há encadeamento das etapas envolvidas. Neste caso, os dados de entrada são retirados diretamente da base de dados, vide figura 5.3.

- **Avaliação Encadeada**: A qualidade avaliada é obtida pelo sequenciamento das etapas envolvidas. Para comparar os resultados com [4], esta avaliação está presente em duas situações distintas:

  – Tarefa Conjunta: Identificação de Argumentos + Classificação de Argumentos. Neste caso os dados das etapas anteriores são extraídos em formato *gold*, e apenas as duas últimas etapas são avaliadas de forma encadeada, vide figura 5.4.

[2]Neste caso, *data snooping* deve ser entendido como o aumento do *overfiting* pela tentativa de maximizar o resultado pela análise indevida e constante da classificação dos dados de testes, que não devem ser utilizados neste contexto, e sim somente para avaliação final do modelo.

– APS Completa: Consiste na avaliação completa das cinco tarefas descritas, onde nenhum dado é retirado em formato *gold*, apenas o texto puro da sentença analisada. Este cenário representa o cenário real onde um pesquisador pode utilizar o anotador semântico em novos textos, vide figura 5.5.

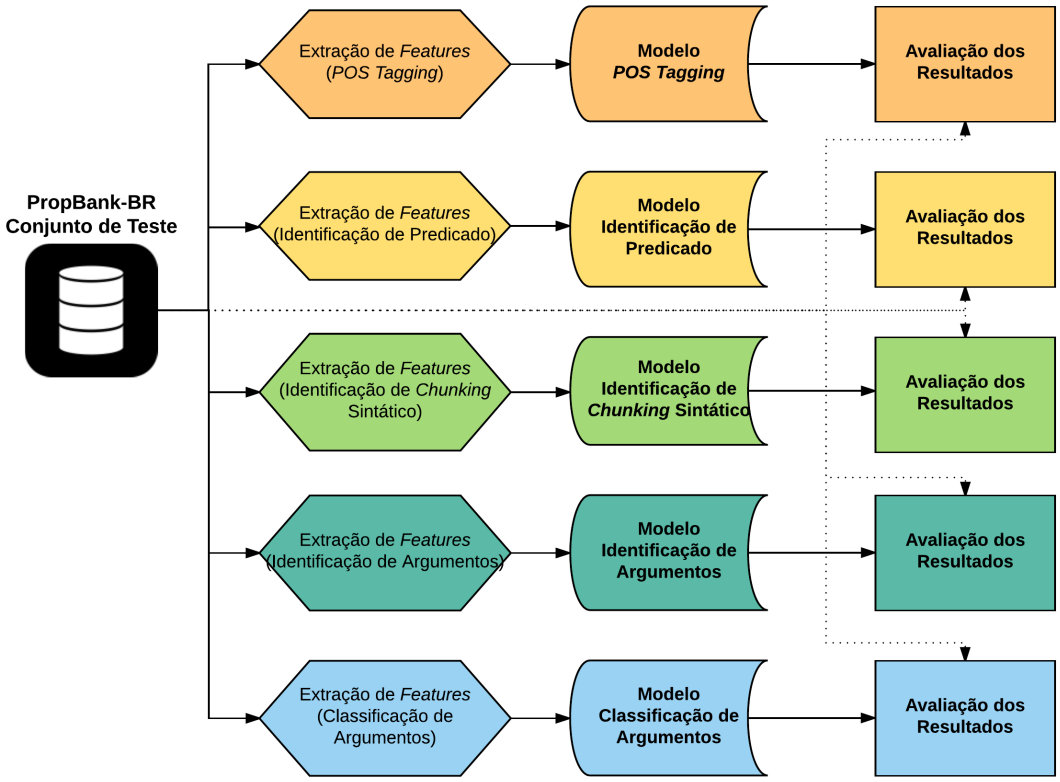

Figura 5.3: Avaliação individual das etapas.

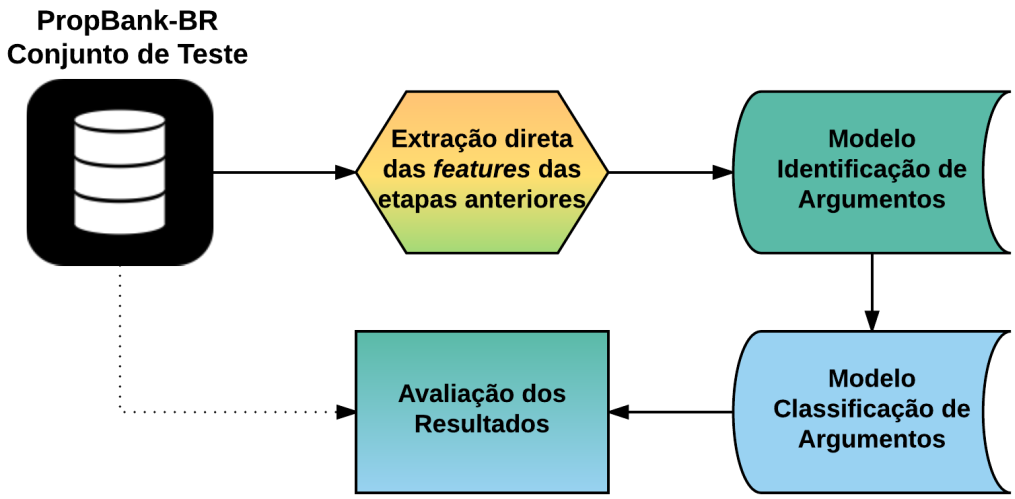

Figura 5.4: Avaliação da tarefa conjunta (Identificação + Classificação de argumentos).

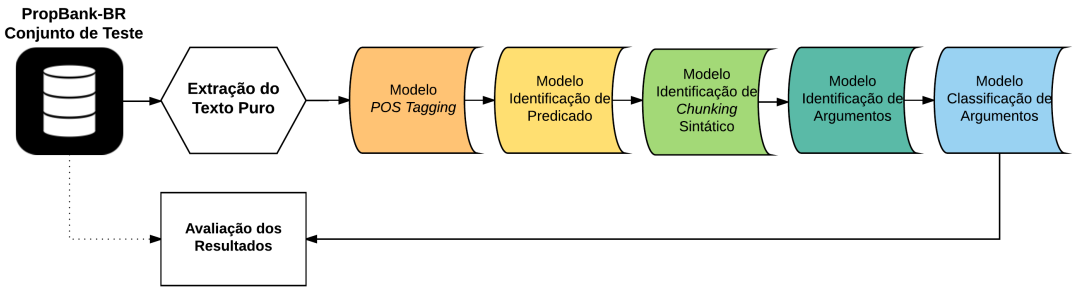

Figura 5.5: Avaliação da tarefa de APS completa. Somente o texto é extraído da base de dados.

5.1.1 *Features* Extraídas

Para cada uma das etapas, um conjunto próprio de *features* para treinamento do modelo foi utilizado. Essas *features* podem ser extraídas diretamente da base de dados ou serem dinamicamente obtidas pela execução de uma das subtarefas.

A lista de *features* utilizadas, em pelo menos alguma das etapas, estão descritas abaixo:

- ***Word*** : A própria palavra da sentença propriamente dita.
- ***Prefix***: Os primeiros caracteres da palavra são extraídos. São utilizados até os 5 primeiros caracteres.

- ***Suffix***: Os últimos caracteres da palavra são extraídos. São utilizados até os 5 últimos caracteres.
- ***Capitalization***: O modo de capitalização da palavra, que deriva do modo em que seus caracteres são escritos. São eles:
  - Sem capitalização,
  - Caracter inicial capitalizado,
  - Todos caraceteres capitalizados,
  - Capitalização interna,
  - Não alfabética, ou seja, caracteres numéricos e/ou especiais envolvidos,.
- ***Word POS Tag***: *POS Tag* da palavra propriamente dita.
- ***Chunk Tag***: Tag do *chunk* cujo qual a palavra alvo está inserida. São elas: PP, NP, ADVP, ADJP, VP e O (para *tokens* fora de *chunks*).
- ***Chunk Position***: Anotação de uma das formas de *tagging*: IOB ou IOBES.
- ***Chunk Size***: Tamanho do *chunk* da palavra alvo.
- ***Head Word***: A palavra inicial do *chunk* da palavra alvo.
- ***Head Word POS Tag***: *POS Tag* da palavra inicial do *chunk* da palavra alvo.
- ***Argument Word***: A palavra inicial do argumento cujo qual a palavra alvo está inserida.
- ***Argument Word POS Tag***: *POS Tag* da palavra inicial do argumento da palavra alvo.
- ***Argument Size***: Tamanho do argumento da palavra alvo. Até dois caracteres: *Small*, Até cinto: *Medium*, Maior que cinco: *Large*.
- ***Predicate***: Palavra-predicado do argumento analisado.
- ***Predicate Distance***: Distância em palavras para a palavra analisada ao predicado alvo. Somente foi utilizado valores absolutos.
- ***Position to Predicate***: Posição referente ao predicado na sentença: -1 para palavras à esquerda, 1 caso contrário.
- ***Word Function Role***: Função da palavra com relação a análise. São elas: ARG (argumento), PRED (predicado) ou sem função.

- **Bigram**: Sequência de duas palavras.

Considerando a seguinte relação:

- Etapa A → *POS Tagging*
- Etapa B → Identificação de Predicado
- Etapa C → Identificação de *Chunkig* Sintático
- Etapa D → Identificação de Argumentos
- Etapa E → Classificação de Argumentos

A utilização das *features* por cada uma das etapas fica descrita na tabela 5.1.

Tabela 5.1: Resumo das *features* utilizadas para cada uma das etapas da implementação proposta.

Feature	**A**	**B**	**C**	**D**	**E**
Word	✓	✓	✓	✓	✓
Prefix	✓				
Suffix	✓				
Capitalization	✓				
Word POS Tag		✓	✓	✓	✓
Chunk Tag				✓	✓
Chunk Position				✓	✓
Chunk Size				✓	✓
Head Word				✓	✓
Head Word POS Tag				✓	✓
Argument Word					✓
Argument Word POS Tag					✓
Argument Size					✓
Predicate				✓	
Predicate Distance				✓	
Position to Predicate					✓
Word Function Role					✓
Bigram	✓	✓	✓	✓	✓

As seções abaixo tratam individualmente sobre as tarefas propostas e exemplificam os resultados obtidos.

5.1.2 POS Tagging

Como tarefa mais básica dentre as citadas nessa dissertação, esta etapa utiliza apenas as características do próprio texto. Prefixos e sufixos são extraídos considerando janelas de até 5 caracteres. A capitação da palavra é obtida pela seguinte ordem de prioridade {Não Alfabética; Toda Capitalizada; Capitalĩzação Interna; Capitalização Inicial; Não Capitalizada}.

O algoritmo 1 apresenta os passos realizados para a extração dos *features* de entrada para esta etapa:

Algorithm 1 Extração para etapa de *POS Tagging*

1: **procedure** POSTaggingFeaturesExtraction(String token)
2: $List < String > prefixes \leftarrow getPrefixes(token)$;
3: $List < String > suffixes \leftarrow getSuffixes(token)$;
4: $String\, capitalization \leftarrow getCapitalization(token)$;
5: $String\, realPOSTag \leftarrow getRealPOSTag(token)$;
6: **return**$(token, prefixes, suffixes, capitalization, realPOSTag)$;

Ao final, a tag real da palavra é concatenada para avaliação da qualidade do modelo. O conjunto de *tag* possíveis foi retirado da própria base de treino do *PropBank-BR* e para a utilização do classificador para outras base de dados é necessário um novo treino utilizando o conjunto válido para esta nova base.

Exemplo de entrada: "Nem todos passam por as duas fases."

Exemplo de *features*:

Nem	N	Ne	Nem	Nem	Nem	m	em	Nem	Nem	Nem	InsideCapitalized	ADV
todos	t	to	tod	todo	todos	s	os	dos	odos	todos	NonCapitalized	PRON-DET
passam	p	pa	pas	pass	passa	m	am	sam	ssam	assam	NonCapitalized	V-FIN
por	p	po	por	por	por	r	or	por	por	por	NonCapitalized	PRP
as	a	as	as	as	as	s	as	as	as	as	NonCapitalized	ART
duas	d	du	dua	duas	duas	s	as	uas	duas	duas	NonCapitalized	NUM
fases	f	fa	fas	fase	fases	s	es	ses	ases	fases	NonCapitalized	N
.	.	.	.	.	.	.	.	.	.	.	NonAlphabetical	PU

Figura 5.6: Exemplo de *features* para etapa de *POS Tagging*.

Exemplo de resultado:

$Nem_{ADV}\ todos_{PRON-DET}\ passam_{V-FIN}\ por_{PRP}\ as_{ART}\ duas_{NUM}\ fases_{N}\ .{}_{PU}$

5.1.3 Identificação de Predicados

A identificação de predicados não seguiu nenhuma restrição mais complexa e foi realizada utilizando caraterísticas derivadas das informações inerentes do próprio sequenciamento de palavras.

O algoritmo 2 apresenta os passos realizados para a extração dos *features* de entrada para esta etapa:

Exemplo de entrada: "Nem todos passam por as duas fases."

Algorithm 2 Extração para etapa de Identificação de Predicados

1: **procedure** PREDICATEFEATURESEXTRACTION(STRING TOKEN, SENTENCE SENTENCE)
2: $String\, posTag \leftarrow getPOSTag(token, sentence)$;
3: $Boolean\, realPredicate \leftarrow isRealPredicate(sentence)$;
4: **return**$(token, posTag, realPredicate)$;

Exemplo de *features*:

Nem	ADV	false
todos	PRON-DET	false
passam	V-FIN	true
por	PRP	false
as	ART	false
duas	NUM	false
fases	N	false
.	PU	false

Figura 5.7: Exemplo de *features* para etapa de Identificação de Predicados.

Exemplo de resultado:

$$Nem_{false}\ todos_{false}\ passam_{true}\ por_{false}\ as_{false}\ duas_{false}\ fases_{false}\ ._{false}$$

Já para o exemplo: "$Adoraria_{true}\ ser_{false}\ candidata_{false}\ ._{false}$", o modelo de identificação de predicados não foi capaz de reconhecer o predicado "adoraria", errando assim a predição.

5.1.4 Identificação de *Chunking* Sintático

A definição dos *chunks* sintáticos seguiu a mesma heurística definida por [4]. Essa heurística segue uma priorização de tipos de *chunks* em detrimentos à outros, utilizando informações da árvore sintática da base de treino[3].

Desta forma, o modelo obtido foi treinado a fim de seguir as regras descritas abaixo:

- Para cada palavra da sentença é atribuída a categoria do nó não-terminal mais próximo na árvore sintática.
- Se a categoria não for reconhecida como válida para o *PropBank-BR*, é atribuída a *tag* "O". Exemplo: Interjeições e pontuação.

[3]Vale a ressalva que a árvore sintática não se torna necessário para a APS de um novo texto neste trabalho, uma vez que o modelo de *chunkig* já está treinado e é capaz de classificar um novo conjunto de dados sem informações obtidas pela análise de sua árvore.

- Se a categoria for válida, é realizada uma busca recursiva para os nós antecessores da árvore e atribuída uma nova etiqueta caso haja uma prioridade maior em detrimento à já atribuída. A ordem de priorização é estabelecida da seguinte forma: PP → NP → ADJP → ADVP → VP.

Segue um exemplo de anotação com priorização retirada de [4]. A identificação dos delimitadores dos *chunks* seguiu a marcação IOB.

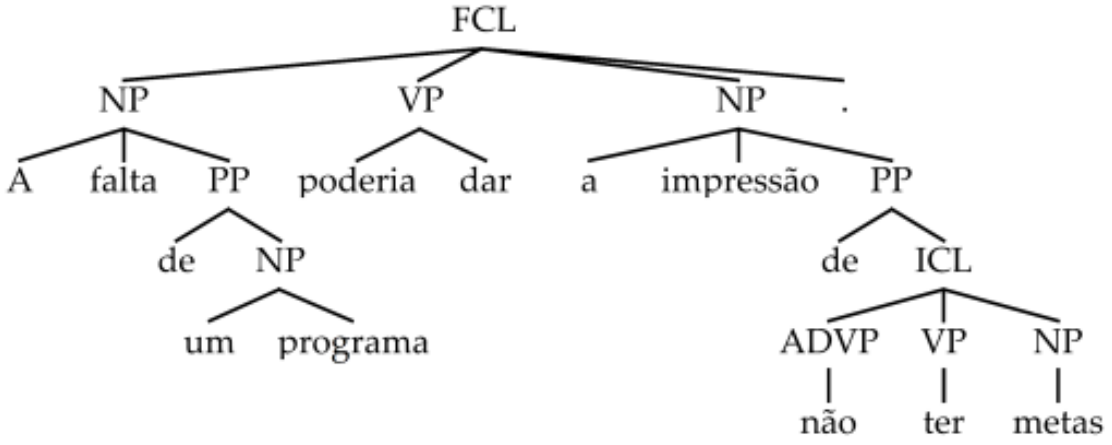

Figura 5.8: Exemplo de árvore sintática com seus descritores. FCL e ICL indicam marcações de tipo de oração e não são consideradas *tags* valídas para o processo de *chunking*. Fonte: [4]

$$A_{B-NP}\ falta_{I-NP}\ de_{B-PP}\ um_{I-PP}\ programa_{I-PP}\ poderia\ dar_{I-VP}\ a_{B-NP}\ impressão_{I-NP}\ de_{B-PP}\ não_{B-ADVP}\ ter_{B-VP}\ metas_{B-NP}\ \cdot_{O}$$

Novamente, somente informações da própria palavra e de *POS tagging* foram utilizadas. O algoritmo 3 descreve o processo.

Algorithm 3 Extração para etapa de Identificação de *Chunking* Sintático

```
1: procedure ChunkingFeaturesExtraction(String token, Sentence sentence)
2:     String posTag ← getPOSTag(token, sentence);
3:     String realChunkPosition ← getRealChunkPosition(token, sentence);
4:     return(token, posTag, realChunkPosition);
```

Exemplo de entrada: "Nem todos passam por as duas fases."[4]

Exemplo de *features*:

[4]O *PropBank-BR* não utiliza contrações, desta forma, seus constituintes são escritos separadamente.

Nem	ADV	NP-B
todos	PRON-DET	NP-I
passam	V-FIN	VP-B
por	PRP	PP-B
as	ART	PP-I
duas	NUM	PP-I
fases	N	PP-I
.	PU	O

Figura 5.9: Exemplo de *features* para etapa de Identificação de *Chunking* Sintático.

Exemplo de resultado:

$$Nem_{ADVP-B}\ todos_{NP-B}\ passam_{VP-B}\ por_{PP-B}\ as_{PP-I}\ duas_{PP-I}\ fases_{PP-I}\ .{}_{O}$$

Neste exemplo, pode-se identificar que a primeira palavra foi classificada erroneamente, o que propagou o erro para o segundo *token*. O resto da sentence foi corretamente classificado de acordo com seu *chunk*.

5.1.5 Identificação de Argumentos

A identificação de argumentos é responsável por delimitar os constituintes que participam de um argumento para o predicado alvo. Para tal, duas abordagens foram testadas:

- IOB → *Tokens* que marcam o início de argumentos são marcados com a *tag* "B". Seus consequentes são marcados com *tag* "I". Tokens que não são referentes a nenhum argumento de predicados são marcados com *tag* "O". Dois argumentos contíguos são identificação pela sequência: "...I B ...".
- IOBES → As descrições das *tags* anteriores se mantém, e são acrescidas pela delimitação de fim para argumentos "E"e para a *tag* de indicação de argumentos de *tokens* únicos "S". Diferentemente da abordagem anterior, dois argumentos contíguos são identificados pela sequência "... (E | S) B ...".

A seguir, dois exemplos de identificação de argumentos utilizando ambas as abordagens, respectivamente, para a sentença: "O treinador Levir Culpi não quis adiantar o substituto de Éder Aleixo."

- $\{O_B\ treinador_I\ Levir_I\ Culpi_I\}\ \{não_B\}\ quis\ \{adiantar_B\ o_I\ substituto_I\ de_I\ Éder_I\ Aleixo_I\}\ .{}_{O}$
- $\{O_B\ treinador_I\ Levir_I\ Culpi_E\}\ \{não_S\}\ quis\ \{adiantar_B\ o_I\ substituto_I\ de_I\ Éder_I\ Aleixo_E\}\ .{}_{O}$

As características utilizadas para esta etapa da APS são derivadas das informações inerentes do próprio sequenciamento de palavras, do *chunk* cujo qual a palavra está inserida e do predicado alvo. O algoritmo 4 descreve o processo de obtenção das *features*.

Algorithm 4 Extração para etapa de Identificação de Argumentos

procedure ARGUMENTIDENTIFICATIONFEATURESEXTRACTION(STRING TOKEN, SENTENCE SENTENCE)
    $String\, posTag \leftarrow getPOSTag(token, sentence)$;
    $String\, chunkType \leftarrow getChunkType(token, sentence)$;
    $String\, chunkPosition \leftarrow getChunkPosition(token, sentence)$;
    $String\, chunkSize \leftarrow getChunkSize(token, sentence)$;
    $String\, headWord \leftarrow getHeadWord(token, sentence)$;
    $String\, headWordPOSTag \leftarrow getHeadWordPOSTag(token, sentence)$;
    $Predicate\, targetPredicate \leftarrow getTargetPredicate(token, sentence)$;
    $Integer\, predicateDistance \leftarrow getPredicateDistance(token, targetPredicate)$;
    $String\, argumentPosition \leftarrow getRealArgumentPosition(token, sentence)$;
    return$(token, posTag, chucnkType, chunkPosition, chunkSize, headWord, headWordPOSTag, targetPredicate, predicateDistance, argumentPosition)$;

Exemplo de entrada: "Nem todos passam por as duas fases."

Exemplo de *features*:

Nem	ADV	NP	B	2	Nem	ADV	passam	-2	B
todos	PRON-DET	NP	I	2	Nem	ADV	passam	-1	E
passam	V-FIN	VP	B	1	passam	V-FIN	passam	0	O
por	PRP	PP	B	4	fases	N	passam	1	B
as	ART	PP	I	4	fases	N	passam	2	I
duas	NUM	PP	I	4	fases	N	passam	3	I
fases	N	PP	I	4	fases	N	passam	4	E
.	PU	O	O	NULL	NULL	NULL	passam	5	O

Figura 5.10: Exemplo de *features* para etapa de Identificação de Argumentos.

Exemplo de resultado:

$$Nem_B\ todos_E\ passam_O\ por_B\ as_I\ duas_I\ fases_I\ ._O$$

No exemplo acima, todos os argumentos do predicado "passam"foram identificados corretamente. Porém, há casos onde os argumentos são identificados indevidamente, como demonstra o exemplo abaixo:

Exemplo de entrada: "Após várias divergências com Fishel, Antinori o demitiu em 1990."

Exemplo de *features*:

Após	PRP	PP	B	3	divergências	N	demitiu	-8	B
várias	PRON-DET	PP	I	3	divergências	N	demitiu	-7	I
divergências	N	PP	I	3	divergências	N	demitiu	-6	I
com	PRP	PP	B	2	Fishel	PROP	demitiu	-5	I
Fishel	PROP	PP	I	2	Fishel	PROP	demitiu	-4	E
,	PU	O	O	NULL	NULL	NULL	demitiu	-3	O
Antinori	PROP	NP	B	1	Antinori	PROP	demitiu	-2	S
o	PRON-PERS	NP	B	1	o	PRON-PERS	demitiu	-1	S
demitiu	V-FIN	VP	B	1	demitiu	V-FIN	demitiu	0	O
em	PRP	PP	B	2	1990	NUM	demitiu	1	B
1990	NUM	PP	I	2	1990	NUM	demitiu	2	E
.	PU	O	O	NULL	NULL	NULL	demitiu	3	O

Figura 5.11: Segundo exemplo de *features* para etapa de Identificação de Argumentos. Um argumento foi perdido pelo modelo.

Exemplo de resultado:

$$Após_O\ várias_O\ divergências_O\ com_O\ Fishel_O\ ,_O\ Antinori_S\ o_S\ demitiu_O\ em_B\ 1990_E\ ._O$$

Já para este exemplo, pode-se averiguar que o primeiro argumento, para o predicado "demitir", não foi identificado corretamente e, desta forma, a informação foi perdida. A falta de identificação do primeiro argumento será contado como um erro no capítulo 6.

Para o exemplo: "A impressão de que o índice de alta de os preços cai se restringe a 27% de as respostas."

Exemplo de *features*:

Após	PRP	PP	B	3	divergências	N	demitiu	-8	B
várias	PRON-DET	PP	I	3	divergências	N	demitiu	-7	I
divergências	N	PP	I	3	divergências	N	demitiu	-6	I
com	PRP	PP	B	2	Fishel	PROP	demitiu	-5	I
Fishel	PROP	PP	I	2	Fishel	PROP	demitiu	-4	E
,	PU	O	O	NULL	NULL	NULL	demitiu	-3	O
Antinori	PROP	NP	B	1	Antinori	PROP	demitiu	-2	S
o	PRON-PERS	NP	B	1	o	PRON-PERS	demitiu	-1	S
demitiu	V-FIN	VP	B	1	demitiu	V-FIN	demitiu	0	O
em	PRP	PP	B	2	1990	NUM	demitiu	1	B
1990	NUM	PP	I	2	1990	NUM	demitiu	2	E
.	PU	O	O	NULL	NULL	NULL	demitiu	3	O

Figura 5.12: Segundo exemplo de *features* para etapa de Identificação de Argumentos. O argumento foi expandido pelo modelo.

Exemplo de resultado:

$$A_B\ impressão_I\ de_I\ que_I\ o_I\ índice_I\ de_I\ alta_I\ de_I\ os_I\ preços_E\ cai_O\ se_O\ restringe_O\ a_O\ 27\%_O\ de_O\ as_O\ respostas_O\ ._O$$

Neste exemplo, o argumento do predicado "cai"teve seus delimitadores expandidos em relação ao argumento original. Logo, este caso também é contato como um erro no processo de avaliação da qualidade do modelo.

Porém, dependendo da situação em que o anotador semântico é utilizado, a resposta fornecida pelo modelo pode satisfazer as necessidades do

pesquisador. Uma vez que o argumento expandido abrange seu núcleo real, sistemas como extratores de informação não terão perda de cobertura com o resultado gerado no exemplo acima. Situações onde o argumento predito é menor que o argumento original também podem ser considerados em outros trabalhos, caso seu núcleo de pesquisa seja mantido nos resultados finais.

Exemplo:

Real $\{O\,presidente\,da\,República\}sanciona\,\{o\,projeto\,de\,Lei\,da\,Terceirização\}$.
Predito $O\,\{presidente\}\,da\,República\,sanciona\,o\,projeto\,de\,\{Lei\,da\,Terceirização\}$.

Neste caso, se considerarmos **presidente** como o núcleo do argumento, nenhuma informação foi perdida.

5.1.6 Classificação de Argumentos

A classificação de argumentos é responsável por etiquetar os argumentos previamente delimitados com uma das etiquetas válidas para o *PropBank-BR*.

Também nesta etapa, para esta dissertação, foi definida qualquer abordagem de inferência final sobre os dados gerados. Exemplos de inferência consistem em: 1) Junção de argumentos contíguos com etiquetas similares, 2) Não repetição de etiquetas, entre outras.

As características utilizadas para esta etapa da APS são derivadas das informações inerentes do próprio sequenciamento de palavras, do *chunk* cujo qual a palavra está inserida, do predicado alvo e do argumento cujo qual a palavra pertence. O processo de extração é descrito no algoritmo 5.

Exemplo de entrada: "Nem todos passam por as duas fases."

Exemplo de *features*:

Nem	ADV	ARG	NP	B	2	Nem	ADV	passam	-1	Nem	ADV	Small	A1
todos	PRON-DET	ARG	NP	I	2	Nem	ADV	passam	-1	Nem	ADV	Small	A1
passam	V-FIN	PRED	VP	B	1	passam	V-FIN	passam	0	NULL	NULL	NULL	V
por	PRP	ARG	PP	B	4	fases	N	passam	1	por	PRP	Medium	A2
as	ART	ARG	PP	I	4	fases	N	passam	1	por	PRP	Medium	A2
duas	NUM	ARG	PP	I	4	fases	N	passam	1	por	PRP	Medium	A2
fases	N	ARG	PP	I	4	fases	N	passam	1	por	PRP	Medium	A2
.	PU	O	O	O	NULL	NULL	NULL	passam	1	NULL	NULL	NULL	O

Figura 5.13: Exemplo de *features* para etapa de Classificação de Argumentos.

Exemplo de resultado:

$$Nem_{A0}\ todos_{A0}\ passam_{V}\ por_{A2}\ as_{A2}\ duas_{A2}\ fases_{A2}\ .{}_{O}$$

Neste exemplo, um dos argumentos foi classificado de forma indevida, enquanto o outro foi etiquetado corretamente. A troca das etiquetas $A0$ e $A1$ acontece com certa frequência nos resultados. Possivelmente, a identificação do tempo verbal da frase poderia reduzir este padrão de erro.

Algorithm 5 Extração para etapa de Classificação de Argumentos

```
1: procedure ArgumentIdentificationFeaturesExtraction(String token, Sentence sentence)
2:     String posTag ← getPOSTag(token, sentence);
3:     String tokenRole ← getTokenRole(token, sentence);
4:     String chunkType ← getChunkType(token, sentence);
5:     String chunkPosition ← getChunkPosition(token, sentence);
6:     String chunkSize ← getChunkSize(token, sentence);
7:     String headWord ← getHeadWord(token, sentence);
8:     String headWordPOSTag ← getHeadWordPOSTag(token, sentence);
9:     Predicate targetPredicate ← getTargetPredicate(token, sentence);
10:    Integer predicateDistance ← getPredicateDistance(token, targetPredicate);
11:    String argumentHeadWord ← getArgumentHeadWord(token, sentence);
12:    String argumentHeadWordPOSTag ← getArgumentHeadWordPOSTag(token, sentence)
13:    String argumentSize ← getArgumentSize(token, sentence);
14:    String realClass ← getRealClass(token, sentence);
15:    return(token, posTag, tokenRole, chunkType, chunkPosition,
    chunkSize, headWord, headWordPOSTag, targetPredicate, predicateDistance,
    argumentHeadWord, argumentHeadWordPOSTag, argumentSize, realClass);
```

5.2 Considerações Finais

Neste capítulo foram apresentadas todas as etapas que descrevem a proposta de anotador semântico para o português. Cada etapa gera um modelo próprio a partir do conjunto de treinamento do *PropBank-BR* que pode ser posteriormente utilizado para a anotação semântica de um texto puro. Também foram descritas todas as *features* utilizadas na proposta e a forma em que cada uma foi empregada. Foram abordados exemplos de resultados corretos e errados, assim como outros que, apesar de serem considerados incorretos, podem ser aproveitados dependendo da situação em que são empregados (identificação de argumentos)

Fica claro que o sistema é capaz de executar a anotação de textos puros inseridos diretamente pelo pesquisador que queira reutilizar o modelo gerado em seus trabalhos. Cada uma das etapas gera as informações utilizadas na etapa posterior, e assim, tarefa de APS completa é abarcada pela proposta de sistema.

Mais exemplos de resultados estão disponíveis no apêndice desta dissertação.

Capítulo 6

Resultados

Define-se como **método científico** como "um conjunto de regras básicas utilizadas no desenvolvimento de uma investigação a fim de produzir conhecimento dito científico, com a obtenção de resultados os mais confiáveis possíveis, seja na produção de novos conhecimentos, bem como na correção e integração de conhecimentos já existentes.". Dentro das conhecidas etapas do método científico, a **experimentação** verifica se uma hipótese é verdadeira a partir de experimentos controlados, cujos dados são medidos e seus resultados anotados.

A partir da obtenção do sistema *NLPNET* de Fonseca [4], único encontrado e disponível para testes, esta capítulo inicia-se com a investigação dos resultados citados por Fonseca para comparação com o proposto nessa dissertação.

Após a avaliação, o capítulo trata sobre a comparação entre os resultados.

6.0.1 Avaliação do *NLPNET*

Antes de mais nada, deixa-se claro que o objetivo desta análise não é rebater os resultados do autor do *NLPNET*, e sim elucidar os resultados obtidos a partir de instruções fornecidas em seu site[1], as quais indicam as etapas necessárias para sua configuração e para a anotação de papéis semânticos a partir de textos puros.[2].

Para garantir a imparcialidade dos dados, os experimentos realizados são avaliados, principalmente, de acordo com o *script* oficial utilizado na competição CoNLL-2004 [21]. Este *script*, escrito em *perl*, foi utilizado por todos os competidores a fim de testar seus resultados finais de anotação. O script está disponível para download[3] por qualquer pessoa.

[1] nilc.icmc.usp.br/nlpnet/

[2] Textos puros devem ser entendidos como sequências de palavras simples, sentenças, anotações auxiliares, como sintáticas ou semânticas.

[3] http://www.lsi.upc.edu/ srlconll/st04/st04.html

Outras formas de comparações também foram utilizadas, como o próprio *script* de avaliação disponibilizado por Fonseca.

Identificação de Predicados

Para a identificação de predicados as avaliações foram realizadas a partir do modelo pré-gerado e disponibilizado para *download*. Seus resultados foram cálculados utilizando o *script* disponibilizado por Fonseca e pelo *script* da *CoNll-2004*.

A partir das duas avaliações, esperava-se encontrar o valor declarado pelo autor de 89,9 pontos na medida F_1 para esta tarefa.

Nesta avaliação, sem re-treinos envolvidos, o resultado obtido apontaram 93,16 pontos na medida F_1, ou seja, o modelo apresentado por Fonseca possui pontuação maior que a declarada em sua dissertação para a identificação de predicados, quando avaliado por seu próprio algoritmo.

Já no *script* da competição seu resultado apresenta 93,10 na medida F_1. A tabela 6.1 resume os valores obtidos.

Tabela 6.1: Avaliações quanto a identificação de predicados por Fonseca - 2013

Valores	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Descritos pelo Autor	88,62%	91,21%	89,90
Script de Fonseca	92,21%	94,14%	93,16
Script da *CoNLL-2004*	96,00%	90,38%	93,10

A análise dos resultados permite observar que, mesmo que os valores obtidos pelos dois *scripts* de avaliação não sejam os mesmo, ambos apresentam melhores valores que os declarados por Fonseca em sua dissertação de mestrado [4].

Identificação de Argumentos

A segunda tarefa realizada pelo autor consiste na delimitação dos argumentos do predicado para posterior avaliação. Novamente, as avaliações dos resultados serão obtidas pela execução dos dois *scripts* supracitados.

Como o *NLPNET* não identifica *chunks* de palavras automaticamente, não é possível avaliar os resultados declarados pelo seu criador quanto essas informações foram inseridas para aprendizagem da rede neural. Desta forma, serão somente comparados os resultados sem *chunks* sintáticos.

Os resultados obtidos estão declarados abaixo na tabela 6.3.

Tabela 6.2: Avaliações quanto a identificação de argumentos por Fonseca - 2013

Valores	**Precisão**	**Cobertura**	F_1
Descritos pelo Autor	76,42%	72,44%	74,38
Script de Fonseca	69,07%	67,78%	68,42
Script da *CoNLL-2004*	69,08%	64,18%	66,54

A tabela mostra que os valores obtidos pelos dois *scripts* de avaliação se aproximaram, enquanto o declarado pelo autor está aproximadamente 6 à 8 pontos acima. Tal redução de resultado poderá impactar o resultado final da anotação de papéis semânticos, uma vez que o erro agregado pode se propagar entre as subsequentes tarefas.

Classificação de Argumentos

Novamente, somente os resultados sem o auxílio das informações sintáticas são analisados, desta vez, não só pela falta de reconhecimento automático do sistema, bem como pela inexistência desta abordagem por [4]. Espera-se que os valores obtidos aproximem-se de 88,14%.

Tabela 6.3: Avaliações quanto à identificação de argumentos por Fonseca - 2013

Valores	**Acurária**
Descritos pelo Autor	88,14%
Script de Fonseca	85,30/

Nesta avaliação, somente o *script* do autor foi executado, devido ao fato que, para poder avaliar esta etapa no *script* da competição seriam necessárias alterações dentro do sistema de Fonseca. Para não correr riscos de alterar os resultados, e desta forma prejudicar a pontuação do autor, optou-se por pela não execução do segundo *script*.

Tarefa Conjunta

A tarefa conjunta é dita pela execução das tarefas de identificação de argumentos e por sua classificação. Desconsidera-se nesta tarefa a identificação de predicados, assim, desta forma, esses são extraídos automaticamente da base de dados.

Para a tarefa conjunta, somente o *script* oficial da competição será utilizado, pois Fonseca [4] não disponibilizou um avaliador próprio para esta tarefa.

Esperava-se que, ao analisar os resultados sem a utilização da informação sintática dos constituintes, a pontuação de 65,13 pontos da medida F_1 fosse obtida. Todos os valores estão descritos na tabela abaixo.

Tabela 6.4: Avaliações quanto à tarefa conjunta por Fonseca - 2013

Valores	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Descritos pelo Autor	67,06%	63,31%	65,13
Script da *CoNLL-2004*	59,35%	58,02%	58,68

Novamente, os resultados apresentam uma baixa de aproximadamente 7 pontos nos resultados finais.

APS Completa

Por fim, a tarefa completa de anotação de papéis semânticos é avaliada. Neste cenário, a identificação de predicados precede as etapas da tarefa conjunta. Os resultados estão descritos na tabela 6.5.

Tabela 6.5: Avaliações quanto à tarefa conjunta por Fonseca - 2013

Valores	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Descritos pelo Autor	66,95%	58,10%	62,21
Script da *CoNLL-2004*	59,44%	55,22%	57,25

Os resultados apontam nova perda, agora de aproximadamente apenas 5 pontos, na medida F_1.

Ao final de todas as análises realizadas, os valores obtidos por Fonseca, sem nenhum conhecimento sintático como dados de entrada, estão descritos na seguinte tabela:

Tabela 6.6: Resultado de Fonseca após avaliação no *script* oficial da competição *CoNLL-2004*.

Tarefa	**Precisão**	**Cobertura**	F_1	**Acurácia**
Identificação de Predicados	96,00%	90,38%	93,10	-
Identificação de Argumentos	69,08%	64,18%	66,54	-
Classificação de Argumentos	-	-	-	85,30%
Conjunta	59,35%	58,02%	58,68	-
APS Completa	59,44%	55,22%	57,25	-

Quando comparados ao trabalho apresentado nesta dissertação, os valores serão retirados da tabela 6.8.

6.1 Resultados Obtidos

Esta seção é referente aos resultados obtidos quanto a proposta tratada nesta dissertação.

As descrições das métricas utilizadas para medição dos resultados, *precisão, cobertura, acurácia* e F-1, podem ser encontradas em [101].

É importante atentar que, para um mesmo conjunto de entrada, o algoritmo do *Linear-Chain CRF* produz o mesmo resultado. Como as divisões utilizadas por [4] já estão definidas, não realizou à análise de testes estatísticos. Além do fato de não se re-treinar o modelo do *NLPNET* e cometer erros de parametrização.

A seguir, os melhores resultados, para cada uma das etapas declaradas na seção 5.

6.1.1 *POS Tagging*

O modelo criado para *POS Tagging* foi avaliado utilizando dois conjuntos distintos de dados. Além do *PropBank-BR* utilizou-se a base denominada *Mac-Morpho* [102] na sua terceira versão. Utilizou-se essa base de dados para que os resultados obtidos pudessem ser comparado à outros trabalhos presentes na literatura para o português brasileiro.

Diferentes trabalhos já foram realizados para o *POS Tagging* de dados advindos do português. Dentre eles citam-se os resultados estado-da-arte obtidos por Fonseca [103] e por Santos [104]. Seus respesctivos valores, quando avaliados na base *Mac Morpho v3.0* são apresentados na tabela 6.7

Tabela 6.7: Comparacão dos resultados de *POS Tagging* na base *Mac-Morpho v3.0*

Trabalho	**Acurácia**
Fonseca	97,33%
Santos	97,47%
Modelo Proposto	96,84%

.

Apesar de se assemelhar aos resultados obtidos por outros autores, ambos os trabalhos apresentam valores maiores que o obtido neste trabalho. Porém, para manter seu caráter purista, optou-se por não utilizar outros *POS Taggers* senão o proposto nesse trabalho.

O modelo criado também foi avaliado para a base de dados *PropBank-BR*, porém, pela inexistência de outros trabalhos avaliados nesta base, não é possível compará-los à outros trabalhos. O resultado de acurácia final para o *PropBank-BR*, na base de testes, foi de 97,32%.

6.1.2 Identificação de Predicados

Por ser o único trabalho na literatura, para o português, que identifica predicados automaticamente no *PropBank-BR*, o trabalho de Fonseca [4] é o único apto à comparação.

Conforme comentado na seção 6.0.1, todos os resultados deste autor serão comparados utilizando os descritos na tabela 6.8.

Ao final das execuções, o melhor resultado encontrado obteve resultados melhores que o de Fonseca. A tabela abaixo resume os resultados obtidos. A avaliação dos resultados foi realizada utilizando o *script* oficial.

Tabela 6.8: Comparação dos resultados para a identificação de predicados.

Trabalho	$\boldsymbol{F_1}$
Fonseca	93,13
Modelo Proposto	95,12

Quando analisado em *script* próprio de avaliação, o identificador de predicado desenvolvido apontou 96,62 pontos na F_1. Essa incompatibilidade de valores pode ser atribuída aos dados originais conterem tags do tipo (C-*), as quais não foram consideradas neste trabalho como explicado na seção 2.3.2.

Quando comparados aos resultados obtidos para o inglês, os resultados obtidos apresentam-se competitivos, porém, vale a ressalva que o *PropBank-BR* somente apresenta predicados verbais, enquanto outras bases apresentam também nominais,

adverbiais e outros. Este fato pode influenciar positivamente os resultados alcançados para o português.

6.1.3 *Identificação de Chunking Sintático*

A identificação de *chunking* sintático seguindo a abordagem declarada não foi encontrada em nenhum outro trabalho na literatura, desta forma, a anotação de *chunking* não foi comparada.

Como esta tarefa propriamente dita não faz parte dos objetos mensuráveis pelo *script* da competição *CoNLL-2004*, um avaliador próprio foi criado. Um *chunk* foi considerado correto somente se todas as palavras que o compõe foram etiquetadas com o correto valor, ou seja, reconhecimentos incompleto ou sobressalente foram considerados erros.

Os resultados obtidos estão presentes na tabela 6.9.

Tabela 6.9: Resultados obtidos para a identificação de *chunking* sintáticos.

Trabalho	**Acurácia**	F_1
Modelo Proposto	80,00	82,50%

6.1.4 *Identificação de Argumentos*

Na etapa de identificação de argumentos foram obtidos resultados pelas duas variantes citadas: IOB e IOBES. Para descarte de erros agregados, todas as *features* de entrada foram retiradas de forma *gold* da base de dados, desta forma, a avaliação diz a respeito apenas para a identificação dos argumentos.

A avaliação dos resultados foi realizada utilizando o *script* oficial.

Tabela 6.10: Comparação dos resultados para a identificação de argumentos.

Trabalho	**Precisão**	**Cobertura**	F_1
Fonseca	69,27%	67,72%	68,49
Modelo Proposto (IOB)	75,92%	73,51%	74,69
Modelo Proposto (IOBES)	75,09%	74,81%	74,95

Ambas as marcações geraram modelos com resultados semelhantes. As duas apresentaram valores superiores ao de Fonseca.

Analogamente ao que foi medido para os *chunks*, um argumento é considerado correto apenas se for identificado completamente, sem adição ou remoção de palavras.

6.1.5 *Classificação de Argumentos*

Para a classificação de argumentos, novamente os argumentos de entrada foram assinalados de forma *gold*. A avaliação dos resultados foi realizada utilizando o *script* oficial.

Tabela 6.11: Comparação dos resultados para a classificação de argumentos.

Trabalho	**Acurácia**
Fonseca	85,30%
Modelo Proposto (Com repetição de etiquetas)	78,92%
Modelo Proposto (Sem repetição de etiquetas)	78,17%

Ainda outras propostas de inferência foram avaliadas, tais como utilizar vetorização de palavras para agrupar predicados semelhantes e extrair informações de seus argumentos no auxílio da classificação ou estender argumentos de mesma tag para formulação de um argumento único. Todos os resultados alcançados foram piores que Fonseca em seu trabalho.

6.1.6 Tarefa Conjunta

A tarefa conjunta para o nosso modelo uniu os melhores resultados obtidos: identificação de argumentos pela marcação IOBES e classificação dos argumentos com repetição válida de etiquetas.

Utilizando o algoritmo oficial da competição *CoNLL-2004*, os resultados alcançados estão descritos na tabela 6.15.

Tabela 6.12: Comparação dos resultados para a tarefa conjunta.

Trabalho	**Precisão**	**Cobertura**	F_1
Fonseca	59,35%	58,02%	58,68
Modelo Proposto (IOBES + Repetição)	61,61%	61,38%	61,50

Os resultados estão, em caráter mais descritivos estão apresentados na tabela abaixo.

Tabela 6.13: Comparação dos descritiva resultados para a tarefa conjunta.

	Fonseca				**Modelo Proposto**			
Etapa	**Prec.**	**Cob.**	F_1	**Acu.**	**Prec.**	**Cob.**	F_1	**Acu.**
Identificação	69,27%	67,72%	68,45	-	75,09%	74,81%	74,95	-
Classificação	-	-	-	85,67%	-	-	-	82,04%
Conjunta	59,35%	58,02%	56,68	-	61,61%	61,38%	61,50	-

Como descrito nas tabelas, a tarefa conjunto realizada pelo modelo proposto apresentou resultados melhores que o de Fonseca.

6.1.7 *APS Completa*

A tarefa completa de anotação de papéis semântica utiliza a configuração da tarefa conjunta e acresce a identificação de predicados (e, para o caso desta dissertação, as etapas de *POS Tagging* e reconhecimento de *chunks* sintáticos).

Esta tarefa é referente à uma anotação para um novo texto puro (sentença sem anotações sintática ou semânticas auxiliares). Os resultados, avaliados pelo *script* oficial, estão descritos na tabela 6.14.

Tabela 6.14: Comparação dos resultados para a tarefa de anotação de papéis semânticos completa.

Trabalho	**Precisão**	**Cobertura**	$\boldsymbol{F_1}$
Fonseca	59,44%	55,22%	57,25
Modelo Proposto (IOBES + Repetição)	59,53%	56,53%	58,00

Tabela 6.15: Comparação dos descritiva resultados para APS completa.

	Fonseca				**Modelo Proposto**			
Etapa	**Prec.**	**Cob.**	$\boldsymbol{F_1}$	**Acu.**	**Prec.**	**Cob.**	$\boldsymbol{F_1}$	**Acu.**
Identificação	68,08%	66,18%	66,54	-	71,12%	67,54%	69,28	-
Classificação	-	-	-	86,05%	-	-	-	83,70%
Conjunta	59,44%	55,22%	57,25	-	59,53%	56,53%	58,00	-

Novamente, as etapas só obtiveram valores menores na classificação (apesar da distância absoluta estar diminuindo), porém melhores nas outras etapas. Desta forma, apesar da diferença ser de menos de 1 ponto na medida F_1, o presente modelo apresenta os melhores resultados para anotação de papéis semânticos para o português, quando comparado ao modelo disponibilizado por Fonseca.

6.2 Considerações Finais

Após a avaliação do modelo *NLPNET*, utilizando principalmente o *script* fornecido oficialmente pela competição *CoNLL-2004 Shared Tasks*, o modelo apresentado nesta dissertação obteve melhores resultados em 2 das 3 principais tarefas da APS, além da tarefa conjunta.

Apesar da diferença para a anotação de papéis semânticos completa tenha obtido melhores resultados, a diferença não é significativamente maior. Mesmo assim, o objetivo deste trabalho foi alcançado, e desta forma, foi possível criar um modelo para APS alternativo para a anotação de textos puros.

Capítulo 7

Conclusão

Este projeto se prontificou a apresentar uma nova ferramenta para anotação de papéis semânticos para o português. A partir de textos anotados advindos da base *PropBank-BR*, seu desenvolvimento foi embasado pela utilização do *Conditional Random Fields* como modelo de resposta.

Como principal objetivo vislumbrado, o projeto em questão é capaz de anotar semanticamente textos puros, ou seja, sentenças sem prévia anotação sintática/semântica. Como benefício secundário, o projeto proposto é facilmente transladado para qualquer outra língua, necessitando somente que haja um recurso léxico referente à esta escolhida.

Após apresentações e comparações com outros sistemas de anotação para o português, principalmente o brasileiro, cada um com suas inerentes características, o trabalho apresentado obteve resultados próximos ao *NLPNET* (única ferramenta disponível para anotação de textos puros para o português), alcançando melhores resultados em 2 de 3 etapas para a tarefa anotação de papéis semânticos completo.

Outra característica da abordagem proposta é a não utilização de sistema externos, a não ser pela implementação do CRF, para a APS. Desta forma, podemos considerar o sistema purista, pois executa em seu espaço todas as etapas necessárias para a anotação.

Apesar do sistema apresentado ainda não ter alcançado os significativos resultados já alcançados para o inglês, os trabalhos mais recentes para o português estão se encaminhando para um estreitamente dessa diferença, mesmo considerando o *lack* de recursos quando são comparadas as duas línguas.

As principais contribuições desta dissertação são:

- Proveu um anotador semântico para o português alternativo ao *NLPNET*, que é capaz de anotar textos puros e ser reutilizado por pesquisadores em diversos trabalhos.

- Gerou um modelo que obteve resultados melhores em quase todas as etapas,

senão a classificação dos argumentos.

- Analisou a qualidade do *Conditional Random Fields* dentro da área de APS para o português, fato que não tinha sido realizado de forma completa anteriormente.

7.1 Trabalhos Futuros

Algumas considerações foram realizadas ao longo do trabalho para serem consideradas em investigações futuras sobre o tema. A seguir estão listadas as principais delas:

- Aumento dos dados anotados para a formulação de um *corpus* maior seguindo as diretrizes do *PropBank-BR*. Uma vez que o tamanho da base em português é muito menor para seu análogo em inglês, e a variação polissêmica é uma característica da língua portuguesa, uma base menor pode ser a explicação dos resultados alcançados para essa língua serem piores do que comparado a língua original.

- Enriquecimento das informações sintáticas para a APS. Diferentemente deste trabalho que utilizou apenas informações de *chunks* sintáticos para a anotação, informações advindas de análises de árvores ou identificação de orações podem auxiliar na tarefa como um todo, principalmente na etapa de identificação de argumentos, que uma vez que seus limites parecem estar relacionados aos limites dos constituintes, poderia ser bastante beneficiada.

- Utilização do algoritmo *Tree Conditional Random Fields*. Por ser pouco utilizado no contexto, optou-se pela utilização do modelo linear do algoritmo (*Linear-Chain CRF*), porém, uma das alternativas como modelo de solução é a utilização do *TCRF*. que pode promover melhores resultados para a **identificação de relações de longa distância**.

- Atribuir uma maior capacidade de questionamentos na etapa de inferência, uma vez que todas testadas reduziram a capacidade de predição do algoritmo.

- Optar por uma vertente semissupervisionada para reduzir o problema de escassez de dados.

Referências Bibliográficas

[1] MANCHEGO, F. E. A. *Anotação automática semissupervisionada de papéis semânticos para o português do Brasil*. Tese de Doutorado, Universidade de São Paulo, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-14032013-150816/en.php>.

[2] HARTMANN, N. S. *Anotação automática de papéis semânticos de textos jornalísticos e de opinião sobre árvores sintáticas não revisadas*. text, Universidade de São Paulo, jun. 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-27112015-140053/>.

[3] SUTTON, C., MCCALLUM, A. *An introduction to conditional random fields for relational learning*, v. 2. Introduction to statistical relational learning. MIT Press, 2006. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=lSkIewOw2WoC&oi=fnd&pg=PA93&dq=introduction+to+conditional+random+fields+for+relational+learning.+2006&ots=T0ENM2ffo2&sig=bz7cj0D6rf3lKgTjtCISuwbzEOI>.

[4] FONSECA, E. R. *Uma abordagem conexionista para anotação de papéis semânticos*. Tese de Doutorado, Universidade de São Paulo, 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-26062013-143120/en.php>.

[5] CHAMBERS, N., JURAFSKY, D. "Template-based information extraction without the templates". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 976–986. Association for Computational Linguistics, 2011.

[6] EMANUELE, B., CASTELLUCCI, G., CROCE, D., et al. "Textual inference and meaning representation in human robot interaction". In: *Joint Symposium on Semantic Processing.*, p. 65, 2013.

[7] AZIZ, W., RIOS, M., SPECIA, L. "Shallow semantic trees for SMT". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 316–322. Association for Computational Linguistics, 2011.

[8] BAZRAFSHAN, M., GILDEA, D. "Semantic Roles for String to Tree Machine Translation." In: *ACL (2)*, pp. 419–423, 2013.

[9] YAN, S., WAN, X. "SRRank: leveraging semantic roles for extractive multi-document summarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 22, n. 12, pp. 2048–2058, 2014.

[10] KHAN, A., SALIM, N., KUMAR, Y. J. "A framework for multi-document abstractive summarization based on semantic role labelling", *Applied Soft Computing*, v. 30, pp. 737–747, 2015.

[11] SHEN, D., LAPATA, M. "Using Semantic Roles to Improve Question Answering." In: *EMNLP-CoNLL*, pp. 12–21, 2007.

[12] KAISSER, M., WEBBER, B. "Question answering based on semantic roles". In: *Proceedings of the Workshop on Deep Linguistic Processing*, pp. 41–48. Association for Computational Linguistics, 2007.

[13] PAUL, M., JAMAL, S. "An improved SRL based plagiarism detection technique using sentence ranking", *Procedia Computer Science*, v. 46, pp. 223–230, 2015.

[14] HE, Y., LI, Y., MENG, L. "A New Method of Creating Patent Technology-Effect Matrix Based on Semantic Role Labeling". In: *Identification, Information, and Knowledge in the Internet of Things (IIKI), 2015 International Conference on*, pp. 58–61. IEEE, 2015.

[15] CHEN, J., YANG, F., MA, H., et al. "Text watermarking algorithm based on semantic role labeling". In: *Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016 Third International Conference on*, pp. 117–120. IEEE, 2016.

[16] SCHULER, K. K. "VerbNet: A broad-coverage, comprehensive verb lexicon", 2005. Disponível em: <`http://repository.upenn.edu/dissertations/AAI3179808/`>.

[17] GILDEA, D., JURAFSKY, D. "Automatic labeling of semantic roles", v. 28, n. 3, pp. 245–288. Disponível em: <`http://www.mitpressjournals.org/doi/abs/10.1162/089120102760275983`>.

[18] KINGSBURY, P., PALMER, M. "From TreeBank to PropBank." In: *LREC*. Citeseer, 2002. Disponível em: <`http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.180.5642&rep=rep1&type=pdf`>.

[19] KINGSBURY, P., PALMER, M. "Propbank: the next level of treebank". In: *Proceedings of Treebanks and lexical Theories*, v. 3. Citeseer, 2003. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.2882&rep=rep1&type=pdf>.

[20] PALMER, M., GILDEA, D., KINGSBURY, P. "The proposition bank: An annotated corpus of semantic roles", *Computational linguistics*, v. 31, n. 1, pp. 71–106, 2005. Disponível em: <http://dl.acm.org/citation.cfm?id=1122628>.

[21] CARRERAS, X., MÀRQUEZ, L. *Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling*. 2004.

[22] MÀRQUEZ, L., VILLAREJO, L., MARTÍ, M., et al. "Semeval-2007 task 09: Multilevel semantic annotation of Catalan and Spanish". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 42–47. Association for Computational Linguistics, 2007.

[23] HAJIČ, J., CIARAMITA, M., JOHANSSON, R., et al. "The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–18. Association for Computational Linguistics, 2009.

[24] PUNYAKANOK, V., KOOMEN, P., ROTH, D., et al. "Generalized inference with multiple semantic role labeling systems". In: *Proceedings of CoNLL*, v. 5, 2005. Disponível em: <http://cogcomp.cs.illinois.edu/papers/PunyakanokRoYi05a.pdf>.

[25] PRADHAN, S. S., WARD, W., MARTIN, J. H. "Towards robust semantic role labeling", *Computational Linguistics*, v. 34, n. 2, pp. 289–310, 2008. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.289>.

[26] PALMER, M., GILDEA, D., XUE, N. "Semantic role labeling", *Synthesis Lectures on Human Language Technologies*, v. 3, n. 1, pp. 1–103, 2010. Disponível em: <http://www.morganclaypool.com/doi/abs/10.2200/S00239ED1V01Y200912HLT006>.

[27] LANG, J., LAPATA, M. "Unsupervised induction of semantic roles". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 939–947. Association for Computational Linguistics, 2010.

[28] LANG, J., LAPATA, M. "Unsupervised semantic role induction via split-merge clustering". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1117–1126. Association for Computational Linguistics, 2011.

[29] LANG, J., LAPATA, M. "Unsupervised semantic role induction with graph partitioning". In: *Proceedings of the conference on empirical methods in natural language processing*, pp. 1320–1331. Association for Computational Linguistics, 2011.

[30] FÜRSTENAU, H., LAPATA, M. "Semi-supervised semantic role labeling". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 220–228. Association for Computational Linguistics, 2009.

[31] FÜRSTENAU, H., LAPATA, M. "Semi-supervised semantic role labeling via structural alignment", *Computational Linguistics*, v. 38, n. 1, pp. 135–171, 2012.

[32] KALJAHI, Z., SAMAD, R. "Adapting self-training for semantic role labeling". In: *Proceedings of the ACL 2010 Student Research Workshop*, pp. 91–96. Association for Computational Linguistics, 2010.

[33] PUNYAKANOK, V., ROTH, D., YIH, W.-T. "The necessity of syntactic parsing for semantic role labeling". In: *IJCAI*, v. 5, pp. 1117–1123, 2005. Disponível em: <http://ijcai.org/Past%20Proceedings/IJCAI-05/PDF/1672.pdf>.

[34] PUNYAKANOK, V., ROTH, D., YIH, W.-T. "The importance of syntactic parsing and inference in semantic role labeling", *Computational Linguistics*, v. 34, n. 2, pp. 257–287, 2008. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.257>.

[35] COLLOBERT, R., WESTON, J. "A unified architecture for natural language processing: Deep neural networks with multitask learning". In: *Proceedings of the 25th international conference on Machine learning*, pp. 160–167. ACM, 2008. Disponível em: <http://dl.acm.org/citation.cfm?id=1390177>.

[36] COLLOBERT, R., WESTON, J., BOTTOU, L., et al. "Natural language processing (almost) from scratch", *The Journal of Machine Learning Research*, v. 12, pp. 2493–2537, 2011. Disponível em: <http://dl.acm.org/citation.cfm?id=2078186>.

[37] DURAN, M. S., ALUÍSIO, S. M. "Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels". In: *8th Brazilian symposium in information and human language technology*, pp. 164–168, 2011. Disponível em: <http://www.aclweb.org/website/old_anthology/W/W11/W11-4519.pdf>.

[38] DURAN, M. S., ALUÍSIO, S. M. "Propbank-Br: a Brazilian Treebank annotated with semantic role labels." In: *LREC*, pp. 1862–1867, 2012. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/272_Paper.pdf>.

[39] BICK, E. "The parsing system Palavras", *Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, 2000. Disponível em: <http://beta.visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf>.

[40] BICK, E. "Noun sense tagging: Semantic prototype annotation of a portuguese treebank". In: *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories. Praga*, 2006. Disponível em: <http://ufal.mff.cuni.cz/tlt2006/pdf/126.pdf>.

[41] BICK, E. "Automatic semantic role annotation for Portuguese". In: *Proceedings of TIL 2007-5th Workshop on Information and Human Language Technology*, pp. 1713–1716, 2007. Disponível em: <http://beta.visl.sdu.dk/pdf/TIL2007_roles.pdf>.

[42] BRUCKSCHEN, M., MUNIZ, F., SOUZA, J., et al. "Anotação linguística em xml do corpus PLN-Br. Nilc-TR-09-08", *Série de relatórios do NILC*, 2008.

[43] SEQUEIRA, J., GONÇALVES, T., QUARESMA, P. "Semantic Role Labeling for Portuguese–A Preliminary Approach–". In: *Computational Processing of the Portuguese Language*, Springer, pp. 193–203, 2012. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-28885-2_22>.

[44] RAPOSO, E. P. *Teoría da gramática: a facultade da linguagem*. Ed. Caminho, 1992.

[45] GRUBER, J. S. "Studies in lexical relations." Disponível em: <https://dspace.mit.edu/handle/1721.1/13010>.

[46] FILLMORE, C. J. "The case for case." Disponível em: <http://eric.ed.gov/?id=ED019631>.

[47] JACKENDOFF, R. S. "Semantic interpretation in generative grammar." Disponível em: <http://eric.ed.gov/?id=ED082548>.

[48] ZILIO, L. "Verblexpor : um recurso léxico com anotação de papéis semânticos para o português", 2015. Disponível em: <http://www.lume.ufrgs.br/handle/10183/131590>.

[49] WAGNER, A. *Learning thematic role relations for lexical semantic nets*. Dissertation, Universität Tübingen, 2004. Disponível em: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/46255>.

[50] CANÇADO, M. "Propriedades semânticas e posições argumentais", *DELTA*, v. 21, n. 1, pp. 23–56, 2005.

[51] FILLMORE, C. "Frames and the semantics of understanding", *Quaderni di Semantica*, v. 6, pp. 222–254, 1985.

[52] DOWTY, D. "Thematic proto-roles and argument selection", *language*, pp. 547–619, 1991. Disponível em: <http://www.jstor.org/stable/415037>.

[53] LEVIN, B. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=6wIZWOrcBf8C&oi=fnd&pg=PA1&dq=levin+1993&ots=TA8HQZFXWJ&sig=6_ynoCMIqFiZgodiGONkLeClww4>.

[54] SCARTON, C. E. *VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil*. text, Universidade de São Paulo, jan. 2013. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-19042013-160640/>.

[55] MINKSY, M. "A framework for representing knowledge", *The psychology of computer vision*, v. 73, pp. 211–277, 1975.

[56] FONSECA, C. A., MIRANDA, N. S. "A semântica de frames como instrumento para a análise do discurso discente-marcadores de sucesso em um projeto escolar de dramaturgia", *Signo*, v. 39, n. 67, pp. 79–88, 2014. Disponível em: <https://online.unisc.br/seer/index.php/signo/article/view/5029>.

[57] CANÇADO, M., GODOY, L., AMARAL, L. "The construction of a catalog of Brazilian Portuguese verbs." In: *KONVENS*, pp. 438–445, 2012. Disponível em: <http://www.academia.edu/download/30881859/proceedings.pdf#page=438>.

[58] BAKER, C. F., FILLMORE, C. J., LOWE, J. B. "The berkeley framenet project". In: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pp. 86–90. Association for Computational Linguistics, 1998. Disponível em: <http://dl.acm.org/citation.cfm?id=980860>.

[59] FILLMORE, C. J., JOHNSON, C. R., PETRUCK, M. R. "Background to framenet", *International journal of lexicography*, v. 16, n. 3, pp. 235–250, 2003. Disponível em: <http://ijl.oxfordjournals.org/content/16/3/235.short>.

[60] LOPER, E., YI, S.-T., PALMER, M. "Combining lexical resources: mapping between propbank and verbnet". In: *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*, 2007. Disponível em: <http://verbs.colorado.edu/~kipper/Papers/semlink_iwcs7.pdf>.

[61] HARTMANN, N. S., AVANÇO, L. V., BALAGE, P. P., et al. "A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words". In: *International Conference on Language Resources and Evaluation, 9th.* European Language Resources Association-ELRA, 2014. Disponível em: <https://www.icmc.usp.br/~taspardo/LREC2014-HartmannEtAl.pdf>.

[62] LITKOWSKI, K. "Senseval-3 task: Automatic labeling of semantic roles". In: *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, v. 1, pp. 141–146, 2004.

[63] HACIOGLU, K., PRADHAN, S., WARD, W. H., et al. "Semantic Role Labeling by Tagging Syntactic Chunks." In: *CoNLL*, pp. 110–113, 2004. Disponível em: <https://www.cs.colorado.edu/~martin/Papers/hacioglu-conll04.pdf>.

[64] CARRERAS, X., MÀRQUEZ, L. "Introduction to the CoNLL-2005 shared task: Semantic role labeling". In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pp. 152–164. Association for Computational Linguistics, 2005. Disponível em: <http://dl.acm.org/citation.cfm?id=1706571>.

[65] MORANTE, R., BUSSER, B. "ILK2: Semantic role labelling for Catalan and Spanish using TiMBL". In: *Proceedings of the 4th International Workshop*

on Semantic Evaluations, pp. 183–186. Association for Computational Linguistics, 2007.

[66] SURDEANU, M., JOHANSSON, R., MEYERS, A., et al. "The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies". In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pp. 159–177. Association for Computational Linguistics, 2008.

[67] JOHANSSON, R., NUGUES, P. "Dependency-based syntactic-semantic analysis with PropBank and NomBank". In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pp. 183–187. Association for Computational Linguistics, 2008. Disponível em: <`http://dl.acm.org/citation.cfm?id=1596355`>.

[68] ZHAO, H., CHEN, W., KIT, C., et al. "Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 55–60. Association for Computational Linguistics, 2009.

[69] LI, J., ZHOU, G., ZHAO, H., et al. "Improving nominal SRL in Chinese language with verbal SRL information and automatic predicate recognition". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pp. 1280–1288. Association for Computational Linguistics, 2009. Disponível em: <`http://dl.acm.org/citation.cfm?id=1699674`>.

[70] BJÖRKELUND, A., HAFDELL, L., NUGUES, P. "Multilingual semantic role labeling". In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 43–48. Association for Computational Linguistics, 2009. Disponível em: <`http://dl.acm.org/citation.cfm?id=1596416`>.

[71] BJÖRKELUND, A., BOHNET, B., HAFDELL, L., et al. "A high-performance syntactic and semantic dependency parser". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pp. 33–36. Association for Computational Linguistics, 2010. Disponível em: <`http://dl.acm.org/citation.cfm?id=1944293`>.

[72] ROTH, M., LAPATA, M. "Neural Semantic Role Labeling with Dependency Path Embeddings", *arXiv:1605.07515 [cs]*, maio 2016. Disponível em: <`http://arxiv.org/abs/1605.07515`>. arXiv: 1605.07515.

[73] XUE, N., PALMER, M. "Calibrating Features for Semantic Role Labeling." In: *EMNLP*, pp. 88–94, 2004. Disponível em: <http://verbs.colorado.edu/~xuen/publications/emnlp04.pdf>.

[74] KUDO, T., MATSUMOTO, Y. "Chunking with support vector machines". In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8. Association for Computational Linguistics, 2001.

[75] RIEDEL, S., MEZA-RUIZ, I. "Collective semantic role labelling with Markov logic". In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pp. 193–197. Association for Computational Linguistics, 2008. Disponível em: <http://dl.acm.org/citation.cfm?id=1596357>.

[76] COLLINS, M. "Head-driven statistical models for natural language parsing", *Computational linguistics*, v. 29, n. 4, pp. 589–637, 2003.

[77] CHARNIAK, E. "Immediate-head parsing for language models". In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 124–131. Association for Computational Linguistics, 2001.

[78] TOUTANOVA, K., HAGHIGHI, A., MANNING, C. D. "A global joint model for semantic role labeling", *Computational Linguistics*, v. 34, n. 2, pp. 161–191, 2008. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/coli.2008.34.2.161>.

[79] COLLOBERT, R. "Deep learning for efficient discriminative parsing". In: *International Conference on Artificial Intelligence and Statistics*, 2011. Disponível em: <http://infoscience.epfl.ch/record/192374/files/Collobert_AISTATS_2011.pdf>.

[80] GARG, N., HENDERSON, J. "A Bayesian Model of Multilingual Unsupervised Semantic Role Induction", *arXiv:1603.01514 [cs]*, mar. 2016. Disponível em: <http://arxiv.org/abs/1603.01514>. arXiv: 1603.01514.

[81] BAI, X., XUE, N. "Generalizing the semantic roles in the Chinese Proposition Bank", *Language Resources and Evaluation*, pp. 1–24, 2016. Disponível em: <http://link.springer.com/article/10.1007/s10579-016-9342-y>.

[82] LI, T., LI, Q., CHANG, B. "Improving Chinese Semantic Role Labeling with English Proposition Bank". In: *China National Conference on Chinese*

Computational Linguistics, pp. 3–11. Springer, 2016. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-319-47674-2_1>.

[83] LEE, C., LIM, S., KIM, H. "Korean Semantic Role Labeling Using Structured SVM", *Journal of KIISE*, v. 42, n. 2, pp. 220–226, 2015. Disponível em: <http://www.koreascience.or.kr/article/ArticleFullRecord.jsp?cn=JBGHKW_2015_v42n2_220>.

[84] NAM, K.-M., KIM, Y.-S. "A Word Embedding and a Josa Vector for Korean Unsupervised Semantic Role Induction". In: *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. Disponível em: <http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11829>.

[85] MORANTE, R., VAN DEN BOSCH, A. "Feature construction for memory-based semantic role labeling of Catalan and Spanish", *Recent Advances in Natural Language Processing V: Selected Papers from RANLP 2007*, v. 309, pp. 131, 2009. Disponível em: <https://books.google.com.br/books?hl=pt-BR&lr=&id=xUWRy8OUGZsC&oi=fnd&pg=PA131&dq=Feature+Construction+for+Memory-Based+Semantic+Role+Labeling+of+Catalan+and+Spanish&ots=46njyo-eV_&sig=UYguoHfzBeucPNqM7ZVf8HW5yaA>.

[86] VAN DER PLAS, L., HENDERSON, J., MERLO, P. "D2. 3: Semantic Parser for French, built using one for English", 2009. Disponível em: <http://www.researchgate.net/profile/Paola_Merlo/publication/265319516_D2.3_Semantic_Parser_for_French_built_using_one_for_English/links/5512dee70cf270fd7e33a8fc.pdf>.

[87] VAN DER PLAS, L., HENDERSON, J., MERLO, P. "D6. 2: Semantic Role Annotation of a French-English Corpus", 2010. Disponível em: <http://www.researchgate.net/profile/Paola_Merlo/publication/265196087_D6.2_Semantic_Role_Annotation_of_a_French-English_Corpus/links/5512dee80cf20bfdad523c7a.pdf>.

[88] AFONSO, S., BICK, E., HABER, R., et al. "Floresta Sintá (c) tica: A treebank for Portuguese." In: *LREC*, 2002. Disponível em: <http://beta.visl.sdu.dk/pdf/AfonsoetalLREC2002.ps.pdf>.

[89] AMANCIO, M. A. *Elaboração textual via definição de entidades mencionadas e de perguntas relacionadas aos verbos em textos simplificados do português*. Tese de Doutorado, Universidade de São Paulo, 2009. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-31082011-122100/en.php>.

[90] KEVIN LUND, C. B. "Semantic and associative priming in high-dimensional semantic space", *Proceedings of the 17th Annual Meeting of the Cognitive Science Society*, pp. 660–665, 1995.

[91] LUND, K., BURGESS, C. "Producing high-dimensional semantic spaces from lexical co-occurrence", *Behavior Research Methods, Instruments, & Computers*, v. 28, n. 2, pp. 203–208, 1996. Disponível em: <http://link.springer.com/article/10.3758/BF03204766>.

[92] SAHLGREN, M. "An introduction to random indexing". In: *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE*, v. 5, 2005. Disponível em: <http://drop-out.googlecode.com/svn/trunk/01.%20Document/Vunb/RandomIndexing_intro.pdf>.

[93] SØGAARD, A., JOHANNSEN, A., PLANK, B., et al. "What's in a p-value in NLP?" In: *CoNLL*, pp. 1–10, 2014. Disponível em: <http://www.anthology.aclweb.org/W/W14/W14-16.pdf#page=11>.

[94] CARNEIRO, M. G. *Redes complexas para classificação de dados via conformidade de padrão, caracterização de importância e otimização estrutural*. text, Universidade de São Paulo, nov. 2016. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-01022017-100223/>.

[95] GHOSH, S., GHOSH, S., DAS, D. "Part-of-speech Tagging of Code-Mixed Social Media Text", *EMNLP 2016*, p. 90, 2016.

[96] STRAUSS, B., TOMA, B. E., RITTER, A., et al. "Results of the wnut16 named entity recognition shared task". In: *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pp. 138–144, 2016.

[97] DA SILVA, T. S. *RECONHECIMENTO DE ENTIDADES NOMEADAS EM NOTÍCIAS DE GOVERNO*. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2012. Disponível em: <http://objdig.ufrj.br/60/teses/coppe_m/TiagoSantosDaSilva.pdf>.

[98] ENSAN, F., BAGHERI, E. "Document Retrieval Model Through Semantic Linking". In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 181–190. ACM, 2017.

[99] ZHAO, J., ZHONG, Y., SHU, H., et al. "High-resolution image classification integrating spectral-spatial-location cues by conditional random fields", *IEEE Transactions on Image Processing*, v. 25, n. 9, pp. 4033–4045, 2016.

[100] LAFFERTY, J., MCCALLUM, A., PEREIRA, F. C. N. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", 2001. Disponível em: <https://works.bepress.com/andrew_mccallum/4/>.

[101] POWERS, D. M. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", 2011.

[102] FONSECA, E. R., ROSA, J. L. G. "Mac-morpho revisited: Towards robust part-of-speech tagging". In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013. Disponível em: <http://anthology.aclweb.org/W/W13/W13-4811.pdf>.

[103] FONSECA, E. R., ROSA, J. L. G., ALUÍSIO, S. M. "Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese", *Journal of the Brazilian Computer Society*, v. 21, n. 1, pp. 1–14, 2015. Disponível em: <http://link.springer.com/article/10.1186/s13173-014-0020-x>.

[104] DOS SANTOS, C. N., ZADROZNY, B. "Learning Character-level Representations for Part-of-Speech Tagging." In: *ICML*, pp. 1818–1826, 2014. Disponível em: <http://www.jmlr.org/proceedings/papers/v32/santos14.pdf>.

Apêndice A

Exemplo de Resultados

No apêndice estão apresentados diferentes resultados para todas as cinco etapas envolvidas. Para todos os resultados a primeira coluna refere-se ao *token* da sentença, a segunda refere-se ao valor real do resultado e a última ao valor predito.

A.1 *POS Tagging*

Mania	N	PROP
altera	V-FIN	V-FIN
comportamento	N	N

A	ART	ART
PMD	PROP	PROP
(	PU	PU
psicose	N	N
maníaco-depressiva	ADJ	ADJ
)	PU	PU
é	V-FIN	V-FIN
uma	ART	ART
doença	N	N
psiquiátrica	ADJ	ADJ
que	PRON-INDP	PRON-INDP
leva	V-FIN	V-FIN
a	PRP	PRP
uma	ART	ART
alteração	N	N
abrupta	ADJ	ADJ
de	PRP	PRP
o	ART	ART
comportamento	N	N
.	PU	PU

José_Arthur_Giannotti	PROP	PROP
,	PU	PU
Marilena_Chaui	PROP	PROP
,	PU	PU
Gloria_Kalil	PROP	PROP
e	CONJ-C	CONJ-C
Jorge_da_Cunha_Lima	PROP	PROP
faziam	V-FIN	V-FIN
parte	N	N
de	PRP	PRP
platéia-cabeça	N	N
que	PRON-INDP	PRON-INDP
acompanhou	V-FIN	V-FIN
anteontem	ADV	ADV
em	PRP	PRP
o	ART	ART
Masp	PROP	PROP
a	ART	ART
palestra	N	N
de	PRP	PRP
Claude_Lefort	PROP	PROP
apresentado	V-PCP	V-PCP
por	PRP	PRP
Sérgio_Cardoso	PROP	PROP
.	PU	PU

Essa	PRON-DET	PRON-DET
época	N	N
marca	V-FIN	V-FIN
a	ART	ART
formação	N	N
de	PRP	PRP
a	ART	ART
base	N	N
de	PRP	PRP
a	ART	ART
moral	N	N
samurai	N	V-FIN
,	PU	PU
resultado	N	N
de	PRP	PRP
a	ART	ART
disciplina	N	N
física	ADJ	ADJ
e	CONJ-C	CONJ-C
mental	ADJ	N
de	PRP	PRP
o	ART	ART
zen-budismo	N	PROP
,	PU	PU
de	PRP	PRP
os	ART	ART
ditames	N	N
de	PRP	PRP
o	ART	ART
confuncionismo	N	N
e	CONJ-C	CONJ-C
de	PRP	PRP
o	ART	ART
espírito	N	N
militarista	ADJ	ADJ
reinante	ADJ	N
.	PU	PU

Em todos os exemplos, como os resultados demonstraram, a qualidade geral do classificador foi alta. Somente alguns *tokens* são classificados incorretamente.

A.2 Identificação de Predicados

Miranda	false	false
--	false	false
Em_relação_a	false	false
mim	false	false
isso	false	false
não	false	false
vai	false	false
dar	true	true
absolutamente	false	false
em	false	false
nada	false	false
.	false	false

Me	false	false
indiciaram	true	true
precipitadamente	false	false
,	false	false
açodadamente	false	false
.	false	false

Nos dois casos acima, os predicados foram corretamente identificados.

Ele	false	false
afirmou	true	true
que	false	false
pretende	true	false
reforçar	true	true
seu	false	false
meio-campo	false	false
,	false	false
mas	false	false
não	false	false
adiantou	true	true
a	false	false
escalação	false	false
de	false	false
a	false	false
equipe	false	false
.	false	false

Neste exemplo, o token *pretende* foi erroneamente desclassificado como predicado. O exemplo acima é um dos poucos casos da base de dados onde dois *tokens* sequenciais são predicados reais.

O	false	false
senhor	false	false
teme	true	false
ser	false	false
condenado	false	true
?	false	false

Já neste exemplo, o predicado foi atribuído ao *token* incorreto.

A.3 Identificação de *Chunkings* Sintáticos

Um	NP-B	NP-B
programa	NP-I	NP-I
é	VP-B	VP-B
um	NP-B	NP-B
planejamento	NP-I	NP-I
para	PP-B	PP-B
se	NP-B	NP-B
atingir	VP-B	VP-B
certas	NP-B	NP-B
metas	NP-I	NP-I
.	O	O

Mesmo	ADVP-B	ADVP-B
criticado	VP-B	VP-B
,	O	O
ele	NP-B	NP-B
defende	VP-B	VP-B
sua	NP-B	NP-B
posição	NP-I	NP-I
claramente	ADVP-B	ADVP-B
em_favor_de	PP-B	PP-B
os	PP-I	PP-I
evangélicos	PP-I	PP-I
e	O	O
de	PP-B	PP-B
seu	PP-I	PP-I
ponto	PP-I	PP-I
de	PP-B	PP-B
vista	PP-I	PP-I
religioso	PP-I	PP-I
.	O	O

Os exemplos acima apresentam anotações corretas.

Mesmo	PP-B	ADVP-B
entendendo	VP-B	VP-B
suas	NP-B	NP-B
explicações	NP-I	NP-I
sobre	PP-B	PP-B
o	PP-I	PP-I
programa	PP-I	PP-I
de	PP-B	PP-B
governo	PP-I	PP-I
,	O	O
entendo	VP-B	VP-B
que	O	O
ele	NP-B	NP-B
tinha	VP-B	VP-B
que	VP-I	VP-I
ter	VP-I	VP-I
um	NP-B	NP-B
programa	NP-I	NP-I
de	PP-B	PP-B
governo	PP-I	PP-I
,	O	O
que	NP-B	NP-B
vai	VP-B	VP-B
dar	VP-I	VP-I
um	NP-B	NP-B
meio	NP-I	NP-I
de	PP-B	PP-B
cobrar	VP-B	VP-B
depois	ADVP-B	ADVP-B
.	O	O

Já este exemplo apresenta uma das maiores causas de erros para a etapa, a troca entre *chunks* preposicionais e adverbiais.

A.4 Identificação de Argumentos

A identificação de argumentos sofreu principalmente com acréscimo ou perda de alguns *tokens*, o que foi considerado incorreto nos cálculos dos resultados.

Essa	B	B
época	E	E
marca	O	O
a	B	B
formação	I	I
de	I	I
a	I	I
base	I	I
de	I	I
a	I	I
moral	I	I
samurai	I	I
,	I	I
resultado	I	I
de	I	I
a	I	I
disciplina	I	I
física	I	I
e	I	I
mental	I	I
de	I	I
o	I	I
zen-budismo	I	I
,	I	I
de	I	I
os	I	I
ditames	I	I
de	I	I
o	I	I
confuncionismo	I	I
e	I	I
de	I	I
o	I	I
espírito	I	I
militarista	I	I
reinante	E	E
.	O	O

O exemplo acima está correto, já o exemplo abaixo é considerado incorreto pois o *token* ","não foi predito. Possivelmente este é um erro da base de dados, pois em muitos outros casos pontuação são desconsiderados como argumentos.

A	B	B
professora	I	I
Cláudia_Morgado	I	I
,	I	I
25	I	I
,	I	I
de	I	I
a	I	I
academia	I	I
Competition	I	E
,	E	O
recomenda	O	O
exercícios	B	B
leves	E	E
,	O	O
como	B	B
a	I	I
corrida	E	E
.	O	O

Já o exemplo abaixo, apesar de encontrar integralmente o argumento, está incorreto, pois não o encontrou de forma contígua.

José_Arthur_Giannotti	B	B
,	I	I
Marilena_Chaui	I	I
,	I	I
Gloria_Kalil	I	I
e	I	I
Jorge_da_Cunha_Lima	E	E
faziam	O	O
parte	O	B
de	B	I
platéia-cabeça	I	I
que	I	I
acompanhou	I	I
anteontem	I	I
em	I	I
o	I	I
Masp	I	E
a	I	B
palestra	I	I
de	I	I
Claude_Lefort	I	I
apresentado	I	I
por	I	I
Sérgio_Cardoso	E	E
.	O	O

A.5 Classificação de Argumentos

Os	A0	A0
irmãos	A0	A0
Osny_Silveira_Neto	A0	A0
e	A0	A0
Guilherme_Silveira	A0	A0
abrem	V	V
hoje	AM-TMP	AM-TMP
o	A1	A1
bar	A1	A1
Horácio	A1	A1
em	AM-LOC	AM-LOC
o	AM-LOC	AM-LOC
Itaim	AM-LOC	AM-LOC
.	O	O

A	O	O
hidroginástica	O	O
é	O	O
uma	O	O
alternativa	O	O
para	O	O
quem	A0	A0
não	AM-NEG	AM-NEG
tem	V	V
acompanhante	A1	A1
.	O	O

Os dois primeiros exemplos apresentam classificações corretas.

«	O	O
O	O	O
importante	O	O
é	O	O
levantar	O	O
bem	O	O
o	O	O
joelho	O	O
,	O	O
manter	O	O
o	O	O
ritmo	O	O
e	O	O
encostar	V	V
o	A1	A1
calcanhar	A1	A1
em	A1	A1
o	A1	A1
chão	A1	A1
a	AM-TMP	A2
cada	AM-TMP	A2
movimento	AM-TMP	A2
»	O	O
,	O	O
diz	O	O
.	O	O

Mesmo	AM-ADV	AM-ADV
criticado	AM-ADV	AM-ADV
,	O	O
ele	A0	A0
defende	V	V
sua	A1	A1
posição	A1	A1
claramente	A2	A2
em_favor_de	AM-PRD	A2
os	AM-PRD	A2
evangélicos	AM-PRD	A2
e	AM-PRD	A2
de	AM-PRD	A2
seu	AM-PRD	A2
ponto	AM-PRD	A2
de	AM-PRD	A2
vista	AM-PRD	A2
religioso	AM-PRD	A2
.	O	O

Já os dois exemplos acima, apresentam etiquetações incorretas.

A matriz de confusão abaixo apresenta os resultados totais para a etapa de classificação.

```
---- Confusion Matrix: (one row for each correct role, with the distribution of predictions)
              -1    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
-1: -NONE-     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 0: A0         0  106   17    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0
 1: A1         0   10  186    5    0    0    1    1    0    0    0    2    1    0    0    1    0    4    0
 2: A2         0    0    9   22    1    0    0    1    0    0    0    2    0    0    1    0    0    2    0
 3: A3         0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0
 4: A4         0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
 5: AM-ADV     0    0    4    0    0    0   11    0    0    1    0    0    1    0    0    0    0    3    0
 6: AM-CAU     0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0
 7: AM-DIR     0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    0    0    1    0
 8: AM-DIS     0    2    0    1    0    0    0    0    0    7    0    0    0    0    0    1    0    0    0
 9: AM-EXT     0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0
10: AM-LOC     0    0    0    1    0    0    0    0    0    0    0   25    0    0    0    0    0    1    0
11: AM-MNR     0    0    3    1    0    0    2    1    0    0    0    3    6    0    0    0    0    2    0
12: AM-NEG     0    0    0    0    0    0    0    0    0    0    0    0    0   19    0    0    0    0    0
13: AM-PNC     0    0    1    1    0    0    0    0    0    0    0    0    0    0    3    0    0    0    0
14: AM-PRD     0    0    1    2    0    0    0    0    0    0    0    1    0    0    0    1    0    1    0
15: AM-REC     0    0    2    0    0    0    0    0    0    0    0    0    0    0    0    0    3    0    0
16: AM-TMP     0    1    5    1    0    0    0    0    0    0    0    4    2    0    0    0    0   32    0
17: V          0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0  239
```