PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Extração de Relações do Domínio de Organizações para o Português

SANDRA COLLOVINI DE ABREU

Tese de Doutorado apresentada à Faculdade de Informática como parte dos requisitos para obtenção do título de Doutor em Ciência da Computação. Área de concentração: Ciência da Computação.

Orientadora: Renata Vieira

Porto Alegre
2014

Dados Internacionais de Catalogação na Publicação (CIP)

A162e Abreu, Sandra Collovini de
Extração de relações do domínio de organizações para o português / Sandra Collovini de Abreu. – Porto Alegre, 2014.
112 p.

Tese (Doutorado) – Fac. de Informática, PUCRS.
Orientadora: Profª. Drª. Renata Vieira.

1. Informática. 2. Processamento da Linguagem Natural. 3. Recuperação da Informação. 4. Ontologia. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE TESE DE DOUTORADO

Tese intitulada "Extração de Relações do Domínio de Organizações para o Português", apresentada por Sandra Collovini de Abreu, como parte dos requisitos para obtenção do grau de Doutora em Ciência da Computação, aprovada em 16/01/2014 pela Comissão Examinadora:

Profa. Dra. Renata Vieira — PPGCC/PUCRS
Orientadora

Profa. Dra. Vera Lúcia Strube de Lima — PPGCC/PUCRS

Dra. Lucelene Lopes — DOCFIX/FACIN

Prof. Dr. Fernando Santos Osório — USP – São Carlos

Homologada em 24/04/2014, conforme Ata No. 006 pela Comissão Coordenadora.

Prof. Dr. Luiz Gustavo Leão Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 32 – sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

Para minha família,
amores da minha vida.

AGRADECIMENTOS

Primeiramente, agradeço à minha orientadora, Renata Vieira, pelo incentivo constante e amizade durante todo o Doutorado.

Aos meus saudosos pais, por terem acreditado em mim sempre, e por proporcionarem o início dos meus estudos. Os seus ensinamentos seguem comigo, e com certeza eles estão participando dessa conquista.

Aos *amores* da minha vida, James e Beatriz, que nasceu e cresceu lindamente durante o Doutorado. Agradeço a eles pelo infinito amor, força e inspiração, fundamentais no percurso de todo o Doutorado.

A toda família, pelo amor e incentivo. Em especial, a minha amada sobrinha Carolina, pelo auxílio na etapa de qualificação do Doutorado; à minha sogra Clenir, e à Karina, por todo o apoio e por cuidarem da Beatriz quando eu estava na PUCRS me dedicando ao Doutorado.

As minhas amigas do coração: a minha irmã Ângela e a Cássia, que sempre acompanharam as minhas jornadas. Agradeço a elas pela compreenção e força durante todo o Dourorado.

Aos colegas do laboratório de PLN pelo companheirismo e auxílio no transcorrer deste estudo: Clarissa, Evandro, Tiago, Nicolas, Lucas Hilgert, Lucas Pugens, Tiago Bonamigo, Marcelo, Marco, Roger, Larissa, Daniela, Aline, Denise, Marlo, Rodrigo, Bernardo. Em especial, agradeço ao meu bolsista Lucas Pugens pelo comprometimento e pela dedicação em importantes tarefas relacionadas à tese; às linguistas Aline e Denise pelo apoio e pela realização da anotação das relações no corpus da tese; aos colegas Daniela, Larissa e Marlo pela amizade e por me auxiliarem em momentos importantes da tese.

À FAPERGS e à CAPES pelo apoio financeiro durante parte do doutorado.

Extração de Relações do Domínio de Organizações para o Português

RESUMO

A tarefa de Extração de Relações a partir de textos é um dos principais desafios da área de Extração de Informação, tendo em vista o conhecimento linguístico exigido e a sofisticação das técnicas de processamento da língua empregados. Essa tarefa visa identificar e classificar relações semânticas que ocorrem entre entidades reconhecidas em um determinado texto. Por exemplo, o trecho "No próximo Sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate (...)" expressa uma relação de "vínculo-institucional" que ocorre entre as entidades nomeadas "Ronaldo Lemos" e "Creative Commons". Esta tese propõe um processo para extração de descritores de relação, os quais descrevem relações explícitas entre entidades nomeadas do domínio de Organizações (Pessoa, Organização e Local) utilizando o modelo probabilístico *Conditional Random Fields* (CRF), e sua aplicação em textos da Língua Portuguesa. O modelo probabilístico CRF tem sido aplicado eficientemente em diversas tarefas de processamento de texto sequencial, incluindo recentemente a tarefa de Extração de Relações. A fim de aplicar o processo proposto, um corpus de referência para extração de relações, necessário para o aprendizado, foi anotado manualmente, tomando como base um corpus de referência para entidades nomeadas (HAREM). Com base em uma extensa revisão da literatura sobre a tarefa de extração automática de relações, *features* de diferentes naturezas foram definidas. Uma avaliação experimental foi realizada com o objetivo de avaliar o modelo aprendido utilizando as *features* definidas. Diferentes configurações de *features* de entrada para o CRF foram avaliadas. Dentre elas, destacou-se a inclusão da *feature* semântica baseada na categoria da entidade nomeada, já que essa *feature* conseguiu expressar melhor o tipo de relação que se deseja identificar entre o par de entidades nomeadas. Por fim, os melhores resultados obtidos correspondem à extração de relações entre as entidades nomeadas das categorias Organização e Pessoa, na qual as taxas de F-measure foram de 57% e 63%, considerando as extrações corretas e parcialmente corretas, respectivamente.

Palavras-chave: Extração de Informação; Extração de Relações; Entidades Nomeadas; Reconhecimento de Entidades Nomeadas; Processamento de Linguagem Natural, Conditional Random Fields

Extração de Relações do Domínio de Organizações para o Português

ABSTRACT

The task of Relation Extraction from texts is one of the main challenges in the area of Information Extraction, considering the required linguistic knowledge and the sophistication of the language processing techniques employed. This task aims at identifying and classifying semantic relations that occur between entities recognized in a given text. For example, the sentence “Next Saturday, Ronaldo Lemos, director of Creative Commons, will participate in a debate [...]" expresses a “institutional-bond" relation that occurs between the named entities “Ronaldo Lemos" and “Creative Commons". This thesis proposes a process for extraction of relation descriptors, which describes the explicit relations between named entities in the Organization domain (Person, Organization and Location) by applying, to texts in Portuguese, Conditional Random Fields (CRF), a probabilistic model that has been used in various tasks efficiently in processing sequential text, including the task of Relation Extraction. In order to implement the proposed process, a reference corpus for extracting relations, necessary for learning, was manually annotated based on a reference corpus for named entities (HAREM). Based on an extensive literature review on the automatic extraction of relations task, features of different types were defined. An experimental evaluation was performed to evaluate the learned model utilizing the defined features. Different input feature configurations for CRF were evaluated. Among them, the highlight was the inclusion of the semantic feature based on the named entity category, since this feature could express, in a better way, the kind of relationship between the pair of named entities we want to identify. Finally, the best results correspond to the extraction of relations between the named entities of Organization and Person categories, in which the F -measure rates were 57% and 63%, considering the correct and partially correct extractions, respectively.

Keywords: Information Extraction; Relation Extraction; Named Entity; Named Entity Recognition, Natural Language Processing; Conditional Random Fields

LISTA DE FIGURAS

Figura 2.1 Representação Gráfica das cadeias lineares HMMs, MEMMs e CRFs. (Fonte: [59]) 31
Figura 2.2 Exemplo de Representação Gráfica do CRF para ER. 33
Figura 4.1 Processo de anotação manual das relações. 53
Figura 5.1 Visão geral do processo proposto. 57
Figura 6.1 Melhores resultados de descritores de relação. 77
Figura 6.2 Comparativo de F-measure da base ORG-ORG entre as diferentes configurações de *Features* com validação cruzada de *5-folds*. 78
Figura 6.3 Comparativo de F-measure da base ORG-PES entre as diferentes configurações de *Features* com validação cruzada de *5-folds*. 79
Figura 6.4 Comparativo de F-measure da base ORG-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *5-folds*. 79
Figura 6.5 Comparativo de F-measure da base ORG-PES-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *5-folds*. 79
Figura 6.6 Comparativo de F-measure da base ORG-PES-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *10-folds*. 79

LISTA DE TABELAS

Tabela 2.1 Exemplos de relações propostas na literatura 29
Tabela 2.2 Exemplo de extração de relações de um trecho de texto em português. . . . 30
Tabela 3.1 Dados e métodos de avaliação para o Inglês. 49
Tabela 3.2 Dados e métodos de avaliação para o Português. 50
Tabela 4.1 Corpus de referência. 52
Tabela 4.2 Número de instâncias de relações dos conjuntos de dados. 53
Tabela 4.3 Classificação das relações dos conjuntos de dados. 53
Tabela 4.4 Exemplos de porções de texto de instâncias de relação positivas. 54
Tabela 4.5 Exemplos de porções de texto de instâncias de relação negativas. 55
Tabela 5.1 Número de etiquetas da anotação BIO no conjunto de dados. 61
Tabela 5.2 Conjunto de *Features* baseadas em POS, adaptado de [66, 73]. 63
Tabela 5.3 Conjunto de *Features* Baseadas no Item Lexical, adaptado de [22, 66, 67, 73]. 63
Tabela 5.4 Conjunto de *Features* Sintáticas, adaptado de [66, 67]. 64
Tabela 5.5 Conjunto de *Features* Baseadas em Padrões, adaptado de [4]. 64
Tabela 5.6 Conjunto de *Features* Baseadas na Sequência Frasal, adaptado de [66, 73]. . 65
Tabela 5.7 Conjunto de *Features* Semânticas, adaptado de [73]. 65
Tabela 5.8 Conjunto de *Features* baseadas em Dicionário. 65
Tabela 5.9 Exemplos de vetor de *features*. 66
Tabela 6.1 Exemplo de Critério de Avaliação dos descritores de relação. 69
Tabela 6.2 Classificação BIO de ORG-ORG por conjunto de *features*. 72
Tabela 6.3 Resultados de ORG-ORG por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 72
Tabela 6.4 Classificação BIO de ORG-PES por conjunto de *features*. 73
Tabela 6.5 Resultados de ORG-PES por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 73
Tabela 6.6 Classificação BIO de ORG-LOCAL por conjunto de *features*. 74
Tabela 6.7 Resultados de ORG-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 74
Tabela 6.8 Classificação BIO de ORG-PES-LOCAL por conjunto de *features*. 75
Tabela 6.9 Resultados de ORG-PES-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 75
Tabela 6.10 Classificação BIO de ORG-PES-LOCAL por conjunto de *features*. 76
Tabela 6.11 Resultados de ORG-PES-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 77
Tabela 6.12 Exemplos classificados como falso-positivos na base ORG-ORG. 81
Tabela 6.13 Exemplos classificados como falso-positivos na base ORG-PES. 82

Tabela 6.14 Exemplos classificados como falso-positivos na base ORG-LOCAL. 82
Tabela 6.15 Exemplos classificados como falso-positivos na base ORG-PES-LOCAL. . . . 83
Tabela 6.16 Exemplos classificados como falso-negativos na base ORG-ORG. 84
Tabela 6.17 Exemplos classificados como falso-negativos na base ORG-PES. 84
Tabela 6.18 Exemplos classificados como falso-negativos na base ORG-LOCAL. 85
Tabela 6.19 Exemplos classificados como falso-negativos na base ORG-PES-LOCAL. . . . 85
Tabela 6.20 Comparativo dos resultados da relação de Localização. 86
Tabela 7.1 Resultados de trabalhos de ER do Português. 88
Tabela 8.1 Exemplos positivos da base ORG-LOCAL. 101
Tabela 8.2 Exemplos positivos da base ORG-ORG. 102
Tabela 8.3 Exemplos positivos da base ORG-PES. 103
Tabela 9.1 Classificação BIO de ORG-ORG por conjunto de features. 105
Tabela 9.2 Resultados de ORG-ORG por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 106
Tabela 9.3 Classificação BIO de ORG-PES por conjunto de features. 106
Tabela 9.4 Resultados de ORG-PES por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 107
Tabela 9.5 Classificação BIO de ORG-LOCAL por conjunto de features. 107
Tabela 9.6 Resultados de ORG-LOCAL por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior. 108
Tabela 10.1 Exemplo de entrada para a anotação dos descritores de relação. 109
Tabela 10.2 Exemplo de saída para a anotação dos descritores de relação. 110

LISTA DE SIGLAS

ACE Automatic Content Extraction
BI Business Intelligence
CRF Conditional Random Fields
EN Entidade Nomeada
EI Extração de Informação
ER Extração de Relações
FSM Finite State Machines
HAREM Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas
HMM Hidden Markov Model
IC Inteligência Competitiva
KNN k-Nearest-Neighbors
MEM Maximum Entropy Moldel
MEMM Maximum Entropy Markov Moldel
MUC-6 Message Understanding Conference
Open IE Open Information Extraction
PLN Processamento da Linguagem Natural
POS Part of Speech
ReRelEM Reconhecimento de Relações entre Entidades Mencionadas
REN Reconhecimento de Entidades Nomeadas
SemEval Evaluation Exercises on Semantic Evaluation
SRL Semantic Role Labeling
SVM Support Vector Machines
TAC Text Analysis Conference

SUMÁRIO

1. Introdução 23
   1.1 Objetivos . . . 24
      1.1.1 Objetivo Geral . . . 24
      1.1.2 Objetivos Específicos . . . 25
   1.2 Organização do Trabalho . . . 25

2. Fundamentação Conceitual 27
   2.1 Reconhecimento de Entidades Nomeadas . . . 27
   2.2 Extração de Relações . . . 28
   2.3 *Conditional Random Fields* . . . 31
   2.4 Definição da Tarefa de Extração de Relações no Contexto desta Tese . . . 33

3. Trabalhos Relacionados 35
   3.1 Sistemas/ Abordagens Computacionais para Extração de Relações . . . 35
   3.2 Sistemas de Extração de Relações para o Português . . . 38
   3.3 Extração de Relações utilizando Conditional Random Fields . . . 41
      3.3.1 *Integrating Probabilistic Extraction Model and Data Mining to Discover Relations and Patterns in Text* . . . 41
      3.3.2 *The Tradeoffs Between Open and Traditional Relation Extraction* . . . 42
      3.3.3 *Extracting Relation Descriptors with Conditional Random Fields* . . . 43
   3.4 Avaliação da Tarefa de Extração de Relações . . . 45

4. Corpus da Pesquisa 51
   4.1 Corpus do HAREM . . . 51
   4.2 Anotação dos Dados . . . 52

5. Processo Proposto 57
   5.1 Etiquetagem dos Textos . . . 57
   5.2 Reconhecimento das Entidades Nomeadas . . . 58
   5.3 Representação das Instâncias de Relações . . . 59
   5.4 Definição das *Features* . . . 63
   5.5 Geração e Validação do Modelo Probabilístico CRF . . . 66
   5.6 Descritores de Relação Extraídos . . . 66

6. Avaliação Experimental 69
6.1 Configuração da Avaliação Experimental 70
6.2 Avaliação dos Resultados 70
6.3 Discussão dos Resultados 77
6.4 Análise de Erros 81
6.5 Comparação dos Resultados do Processo Proposto com Sistemas de ER do Português 86

7. Considerações Finais 87
7.1 Contribuições 89
7.2 Trabalhos Futuros 90

Bibliografia 91

8. Apêndice A 101

9. Apêndice B 105

10. Apêndice C 109

1. Introdução

A Extração de Relações (ER) a partir de textos é um dos principais desafios da área de Extração de Informação (EI), a qual busca a extração e a identificação de relações que ocorrem entre entidades. Nesse contexto, dentre as tarefas de EI temos a identificação/reconhecimento de entidades nomeadas (ENs), como nomes de pessoas e de organizações, e a extração das relações entre essas entidades em textos em linguagem natural.

O reconhecimento apropriado de ENs contidas em textos é importante para a EI, uma vez que o sentido de um texto está frequentemente ancorado nestas entidades. De acordo com [90], textos informativos como artigos de notícias geralmente referem-se mais a entidades e a acontecimentos específicos do que a conceitos genéricos.

No entanto, por vezes o Reconhecimento de Entidades Nomeadas (REN) não é suficiente para tarefas de EI, requerendo também a identificação das relações estabelecidas entre essas entidades. Por exemplo, considerando o domínio de negócios, apenas a identificação de nomes de empresas contidas em um artigo de notícias não é tão informativa quanto identificar, por exemplo, a relação de "aquisição" (do inglês, "is acquired by") ocorrida entre duas entidades nomeadas que representam empresas [18, 64, 92].

Tendo em vista o fato das relações semânticas conterem conhecimento mais rico/informativo, surgiu a necessidade de avanços em técnicas para identificação e extração automática de relações. ER é uma tarefa que contribui com diversas áreas além da EI, tais como sistemas de perguntas e respostas, sumarização de textos, recuperação de informação, anotação para Web Semântica, construção e enriquecimento de recursos lexicais e ontologias.

O problema da ER tem sido estudado extensivamente em textos de linguagem natural, incluindo artigos de notícias, publicações científicas, blogs, e-mails, e recursos como a Wikipedia, o Twitter e a Web, em geral [92]. Há um interesse crescente em ER, principalmente motivado pelo crescimento exponencial da informação disponibilizada através da World Wide Web, que torna a tarefa de pesquisar e utilizar esta enorme quantidade de dados impossível através de meios manuais. A popularização de redes sociais como Twitter e Facebook, que impulsionam os usuários a inserir novos dados frequentemente em uma base, resulta em uma geração diária de centenas de milhões de pequenos textos. Esse contexto torna a tarefa de EI ainda mais complexa e uma área de pesquisa relevante [38].

Várias abordagens têm sido propostas para ER a partir de dados não estruturados, tais como aprendizado supervisionado ou não supervisionado; técnicas baseadas em corpus, estratégias linguísticas; recursos como bases lexicais e ontologias; modelos baseados em regras, modelos baseados em aprendizado estatístico e sistemas híbridos. Para alguns idiomas, como o Inglês, há uma extensa pesquisa sobre ER [2, 4, 13, 24, 25, 29, 32, 36, 37, 39, 51, 54, 55, 66, 69, 73, 83, 96, 97, 103, 105, 110], enquanto para o Português, encontramos um número menor de trabalhos que lidam com ER [6, 14, 16, 20, 41, 48, 89, 93, 100, 104]. Além disso, como os idiomas são diferentes não é possível

reutilizar para o Português, recursos e bases de dados existentes para o Inglês, limitando assim mais avanços nas pesquisas em Português.

Dentre as técnicas disponíveis para ER, destacam-se no aprendizado estatístico os modelos *Conditional Random Fields* (CRF). A literatura tem apresentado o modelo probabilístico CRF como uma boa alternativa, uma vez que tem sido aplicado eficientemente em diversas tarefas de processamento de texto sequencial, incluindo recentemente a tarefa de ER [4, 29, 66]. O modelo probabilístico CRF será apresentado na Seção 2.3.

No contexto deste trabalho, um processo para extração de descritores de relação entre entidades nomeadas do domínio de Organizações para o Português é proposto, utilizando o aprendizado estatístico, especificamente o CRF. A escolha por aplicar o modelo probabilístico CRF em particular é devido ao fato deste modelo ter se mostrado competitivo em comparação com outras abordagens [3, 18, 30, 39, 51, 54, 64, 83, 106, 109, 110].

Uma outra motivação para o uso do CRF neste trabalho, é a aplicação de aprendizado de máquina em tarefas que envolvem PLN, uma vez que a maioria dos trabalhos de ER para o Português são baseados em regras [14, 16, 20]. De forma similar, na dissertação de mestrado aplicamos aprendizado de máquina (árvores de decisão) para a tarefa de resolução de correferência em textos do Português [28]. Destaca-se também que, no nosso grupo de pesquisa em Processsamento da Linguagem Natural[1] temos trabalhos que aplicam o modelo CRF na tarefa de REN [33], etapa necessária para ER. Dessa forma, o CRF pode ser aplicado tanto em REN como em ER, e assim podemos ter uma abordagem única para tratar o Português que necessita de recursos para tais tarefas. Não temos conhecimento de nenhum trabalho de ER baseado em CRF para esta língua.

O estudo do domínio de Organizações foi escolhido por apresentar um potencial de aplicabilidade bem claro para diferentes áreas. Aplicações na área de Processamento da Linguagem Natural (PLN) e de Negócios podem se beneficiar com a extração de relações desse domínio, como por exemplo, a área de Vendas e Marketing que ao relacionar diferentes empresas possibilita a identificação de possíveis clientes em potencial. Muitos dos sistemas de REN tratam Organizações, pois são informações importantes para Corporações em geral. As áreas de *Business Intelligence* (BI) e Inteligência Competitiva (IC) estão em crescimento e começam a reconhecer a importância das tarefas de REN e ER como parte integrante de suas aplicações.

1.1 Objetivos

1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral propôr um processo para extração de descritores de relação entre entidades nomeadas do domínio de Organizações (Organização, Pessoa e Local) em textos da Língua Portuguesa.

[1] http://www.inf.pucrs.br/ linatural/

1.1.2 Objetivos Específicos

Com a finalidade de atingir o objetivo geral, os seguintes objetivos específicos são definidos:

- Apresentar uma análise detalhada do estado da arte sobre a tarefa de extração automática de relações semânticas;
- Construir um corpus de referência para a tarefa de ER no domínio de Organizações;
- Definir um conjunto de *features* para o aprendizado da ER;
- Aplicar o modelo probabilístico CRF para o aprendizado de ER;
- Avaliar o processo proposto.

1.2 Organização do Trabalho

Este trabalho está organizado da seguinte forma. No Capítulo 2 é apresentada a fundamentação conceitual do trabalho proposto. Na sequência, o escopo específico deste trabalho é definido no final deste capítulo. Sistemas e abordagens computacionais que representam o estado da arte de ER, e os sistemas para o Português que situam este idioma na tarefa de ER são apresentados no Capítulo 3. Sistemas de ER baseados em CRF são descritos no detalhe neste capítulo, seguida de uma discussão sobre avaliação da tarefa de ER. O corpus de pesquisa utilizado neste trabalho, bem como a sua correspondente anotação para ER são apresentados no Capítulo 4. O processo de ER proposto é descrito no Capítulo 5. Na sequência, a avaliação experimental do processo proposto é apresentada no Capítulo 6. Por fim, no Capítulo 7 as considerações finais são apresentadas.

2. Fundamentação Conceitual

2.1 Reconhecimento de Entidades Nomeadas

REN busca identificar, desambiguar e classificar expressões linguísticas contidas em textos, na sua maioria nomes próprios, que remetem para um referente específico [76]. Um estudo que aborda a tarefa de REN é apresentado em [79].

A tarefa de REN foi formalmente introduzida em 1995 na sexta edição da conferência de avaliação conjunta *Message Understanding Conference* (MUC-6) [77]. Desde então, o interesse pelo REN cresceu significativamente e outras conferências de avaliação conjunta têm sido dedicadas a essa tarefa, incluindo o programa da *Automatic Content Extraction* (ACE) [34] com início em 1999, seguido da conferência *Text Analysis Conference*[1] (TAC) com sua primeira edição em 2008, e por fim, a conferência Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas (HAREM), dedicada exclusivamente à Língua Portuguesa com sua primeira edição em 2005 [91].

De uma forma geral, o REN é realizado em duas etapas: (1) Identificação (ou delimitação), em que as palavras que formam a EN; são selecionadas; (2) Classificação, em que é atribuída a categoria semântica da EN. A classificação mais comum das ENs, estabelecida no MUC-6, considera três tipos: as que referenciam Pessoas Singulares (antropônimos); Coletivas (empresas e organizações) e Lugares (topônimos) [76]. Para exemplificar tomemos a sentença: "Mafalda Reis e Nuno Cunha palestraram na Universidade de Coimbra". No exemplo temos três ENs: "Mafalda Reis", "Nuno Cunha", "Universidade de Coimbra", sendo as duas primeiras correspondentes à categoria semântica Pessoa e a última, à categoria semântica Organização. Entretanto, existem outras categorias de ENs, como as menções a Obras (por exemplo, "Código Da Vinci"); Acontecimentos (por exemplo, "Festa de Santo Antônio"), Tempo (por exemplo, "meio-dia"); Coisa (por exemplo, "barco"), entre outras.

Várias abordagens têm sido adotadas na tarefa de REN; a abordagem mais simples é a que se baseia na consulta de almanaques (*gazetteers*). Tal abordagem depende fortemente da existência de listas de nomes próprios como antropônimos, topônimos, designadores de nomes de empresas, organizações e outras palavras que podem servir para identificar/classificar um nome, como, por exemplo, abreviaturas (Ltda., Jr., etc.), conforme apresentado em [87]. Soluções iniciais se baseavam em regras intuitivas construídas manualmente, as quais exigiam conhecimento de especialistas, como, por exemplo, regras descritas por meio de autômatos de estados finitos [49]. Trabalhos mais recentes sobre REN utilizam métodos estatísticos de aprendizado de máquina, tais como Modelos Ocultos de Markov (HMM) [11], Modelos de Máxima Entropia (MEM) [23], Máquinas de Vetores Suporte (SVM) [57] e Campos Aleatórios Condicionais (CRF) [94]. Abordagens híbridas para REN que se baseiam numa combinação de técnicas também são utilizadas, como, por exemplo, o sistema descrito em [72], o qual combina gramáticas baseadas em regras, modelos de máxima entropia e

[1] http://www.nist.gov/tac/about/index.html

almanaques.

REN pode beneficiar diversas aplicações de PLN. Em sistemas de perguntas e respostas, por exemplo, as palavras candidatas a respostas frequentemente são ENs que necessitam ser previamente identificadas e classificadas por um módulo/sistema de REN [65]. De acordo com [60], o primeiro passo para a maioria das tarefas de EI é detectar e classificar todos os nomes próprios mencionados em um texto, uma tarefa geralmente referida como REN. Segundo [59], a extração de estruturas mais complexas, como eventos ou relações, depende do bom desempenho do REN como uma etapa de pré-processamento.

Na literatura, existem vários trabalhos que consideram REN como parte de sistemas de ER [2,51, 64,96], dado que a etapa de REN pode auxiliar na identificação dos argumentos/entidades que fazem parte de uma determinada relação. Este trabalho considera REN na etapa de pré-processamento do processo para extração de descritores de relação entre ENs no domínio de Organizações, descrito na Seção 5.2.

2.2 Extração de Relações

ER consiste em detectar e classificar relações semânticas que ocorrem entre (pares de) entidades reconhecidas em um determinado texto [60]. Relações semânticas são relações entre conceitos ou significados envolvendo diferentes unidades linguísticas e componentes, como a relação de meronímia ("parte-de") que ocorre entre uma mesma Organização Comercial constituída de uma Filial que possui a relação "parte-de" com a sua Matriz.

Na literatura, encontramos uma variedade de tipos de relações, entretanto uma relação é considerada relevante de acordo com vários fatores, principalmente pelo tipo de informação que se deseja extrair, bem como o objetivo da tarefa de extração. Uma forma de definir as relações relevantes para um determinado domínio, bem como identificar padrões que descrevam tais relações, é a partir da análise dos dados. Segundo Hearst, em [52], diferentes relações podem ser expressadas utilizando um pequeno número de padrões léxico-sintáticos.

De forma geral, a tarefa de ER não possui um conjunto padrão de relações-alvo. Na Tabela 2.1 são apresentadas algumas das relações tratadas por diferentes autores. Destaca-se que várias das relações apresentadas ocorrem entre ENs, como, por exemplo, *employee_of* (*funcionário_de*); *location_of* (*localização_de*); *member_of* (*membro_de*), entre outras.

Diferentes relações têm sido propostas em conferências de avaliação conjunta. Por exemplo, para o Inglês, o programa ACE propôs a detecção de vários tipos/subtipos de relações entre entidades. Conferências com foco em relações como ACE e MUC são referência para diversas pesquisas propostas para Inglês, Chinês e Árabe.

Seguindo esta linha de pesquisa, para o Português foi proposta a trilha de Reconhecimento de Relações entre Entidades Mencionadas (ReRelEM) no HAREM [46]. Mais detalhes das conferências MUC, ACE e HAREM são apresentados na Seção 3.4.

Relações	**Trabalhos para o Inglês**	**Trabalhos para o Português**
author-title (autor-título)/ work_of (trabalho_de) authorOf / author_of (autor_de)	[3, 13]	[16, 46]
location / locatedIn / location_of / located (localização)	[18, 51, 78, 80]	[14, 16, 104]
employment-organization (ocupação-organização) role (papel) / organization-role (organização-papel)	[51, 66, 80, 97, 98]	[27]
quotation-author (citação-autor)	[62, 84]	[40, 93]
CeoOf (diretor_de) person-organization (pessoa-organização)	[18]	[27]
birthPlace / bornIn (local_nascimento_de / natural_de)	[4, 64]	[16]
living_in / home_of / residence (residente_de) place_of (localizado_em)	[80]	[16, 46]
member / member-of (membro_de)	[80]	[27]
manufacturing / manufactured_ by (manufaturado_por) produces (produzir)	[18]	[16, 46]
organization-headquarters (organização-sede) headquarteredIn (sede_de)	[2, 18, 80]	–
employee-of (funcionário_de)	[78]	–
acquired (adquirido) /acquisition (aquisição) merge-acquisition (fusão-aquisição)	[3, 4, 18, 51]	–
wonAward / hasWonPrize (ganhou_prêmio)	[4, 64]	–
part-of / part / (parte-de) / part-whole (parte-todo) subsidiary (subsidiaria)	[51, 80, 83]	–
management (gerenciamento) / founder (fundador) affiliate-partner (filial-parceiro) /client (cliente)	[80]	–
social / parent / family_ relation (relação_familiar)	[29, 66, 80]	[16, 46, 89]
product_of (produto_de)	[78]	–
inclusion (inclusão)	–	[14, 16, 20]
quantified (quantificados) / results (resultados)	–	[41]
dead (morto) / wounded (ferido) / arrested (preso)	[100]	[100]
people_of (povo_de) / affiliation (vinculo_institucional) character_of (personagem_de) / name_of (nome_de) professional_relation (relação_profissional) ownership (proprietario_de / propriedade_de) date_of (data_de) / birth_date (data_nascimento) death_date (data_morte) / life_time (periodo_vida) representative_of (representante_de / representado_por) practised_in (praticado_em) / other_edition (outra_edição) participant_in (participante_em / ter_participação_de)	–	[16, 46]

Tabela 2.1 – Exemplos de relações propostas na literatura

Nesse contexto, um sistema de ER deve ser capaz de extrair, por exemplo, a relação "employment-organization" – proposta no programa ACE – entre as entidades "CEO" e "Microsoft" em "The CEO of Microsoft" [97]. Similarmente, temos para o Português a relação "papel-organização" entre as entidades nomeadas "diretor" e "Creative Commons" em "diretor da Creative Commons" [27]. Para um melhor entendimento, a Tabela 2.2 apresenta as entidades nomeadas e a extração das relações que ocorrem entre essas entidades contidas no trecho do texto descrito em (1), retirado da Coleção Dourada do Segundo HAREM.

> "No próximo sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate (...)."(1)

Entidades Nomeadas	**Relações**
No próximo sábado,	**Ronaldo Lemos** <*membro-de*> **Creative Commons**
Ronaldo Lemos <*Pessoa*>,	
diretor da **Creative**	**Ronaldo Lemos** <*desempenha*> **diretor**
Commons <*Organização*>,	
irá participar de um debate (...)	**Creative Commons** <*possui*> **diretor**

Tabela 2.2 – Exemplo de extração de relações de um trecho de texto em português.

Para extrair relações explícitas em textos, como a apresentada na Tabela 2.2, é necessária a análise de vários aspectos envolvendo a estrutura sintática e semântica da sentença. Alguns aspectos frequentemente analisados para ER são apresentados a seguir:

- A ocorrência de palavras que podem expressar uma relação particular em torno ou próximas às ENs. Por exemplo, "autor de" em "*Dan Brown* é o autor de *O Código Da Vinci*".

- Categorias lexicais providas pela anotação de *Part of Speech* (POS) podem auxiliar a identificar se uma palavra define uma relação ou não. Por exemplo, o verbo "fundar" em "*Microsoft* foi fundada por *Bill Gates*", que expressa uma relação de "fundador" entre as ENs "*Microsoft*" e "*Bill Gates*".

- Estruturas sintáticas da sentença podem expressar uma relação, tais como sintagmas preposicionais ou sintagmas verbais anotados por um parser. Por exemplo, a relação de "localização" em: "A opinião é do agrônomo Miguel Guerra, da *UFSC* de *Santa Catarina*" entre as ENs "*UFSC*" e "*Santa Catarina*".

Atualmente, a tarefa de ER é uma questão de pesquisa que envolve diferentes áreas, como PLN, Aprendizado de Máquina, Banco de Dados, Recuperação de Informação, entre outros. Para a tarefa de ER para EI, diferentes abordagens foram desenvolvidas: técnicas de aprendizado supervisionado utilizando corpus anotado; abordagens não supervionadas com base em padrões de extração genéricos; métodos semi-supervisionados, tais como *bootstrapping*, que necessita apenas de poucos exemplos anotados – e também a abordagem *Open Information Extraction* (Open IE) para extração das relações não definidas previamente.

Mais recentemente, a literatura tem apresentado o CRF como uma boa alternativa, em que ER é tratada como uma tarefa de etiquetagem de sequências de forma eficiente [29]. Neste trabalho, aplicamos o modelo CRF no processo de extração de descritores de relação em textos do Português, sendo que as relações classificadas pelo CRF não são conhecidas, somente as ENs são previamente reconhecidas (parâmetros da relação). Na próxima seção, iremos detalhar esse modelo e exemplificar o seu uso para a tarefa de ER.

2.3 *Conditional Random Fields*

Muitos trabalhos tratam tarefas de PLN como um problema de etiquetagem de sequências estruturadas e utilizam abordagens estatísticas de aprendizado de máquina como os modelos HMM [43], MEM [61], Modelos de Markov de Máxima Entropia (MEMMs) e CRFs [59].

Em especial, o modelo probabilístico CRF tem sido aplicado em uma variedade de tarefas de processamento de texto sequencial de forma eficiente [66], incluindo *part-of-speech* e *chunking* [99], segmentação de palavras [95], identificação de regras/papéis semânticos [88], REN [33] e, recentemente, ER [4, 29, 66].

CRFs são modelos baseados em grafos não direcionados utilizados para calcular a probabilidade condicional de valores atribuídos a nodos de saída a partir de determinados valores atribuídos a nodos de entrada [63]. Um modelo condicional especifica as probabilidades das sequências de etiquetas possíveis dada uma sequência de observação. Sendo que, a probabilidade condicional da sequência de etiquetas pode depender de *features* não independentes, arbitrárias da sequência de observação [70].

Inicialmente, os CRFs foram aplicados para resolver problemas de etiquetagem de textos [63], utilizando-se CRFs na forma de cadeias lineares. CRFs desse tipo ocorrem quando os nodos de saída do grafo são conectados por arestas que formam uma cadeia linear. CRFs de cadeia linear correspondem a máquinas de estado finito treinadas condicionalmente (*Finite State Machines* - FSM) [70].

A principal diferença entre as cadeias lineares HMMs, MEMMs e CRFs ocorre no diagrama que representa as dependências. A Figura 2.1 representa graficamente tais diferenças entre as cadeias lineares HMMs, MEMMs e CRFs, nesta ordem. O que mais faz diferir as cadeias lineares CRFs das MEMMs é que a etiqueta da observação corrente pode não depender apenas das etiquetas anteriores, mas também das etiquetas futuras [59]. Além disso, CRFs são modelos baseados em grafos não direcionados, enquanto HMMs e MEMMs são direcionados, conforme ilustrado na Figura 2.1.

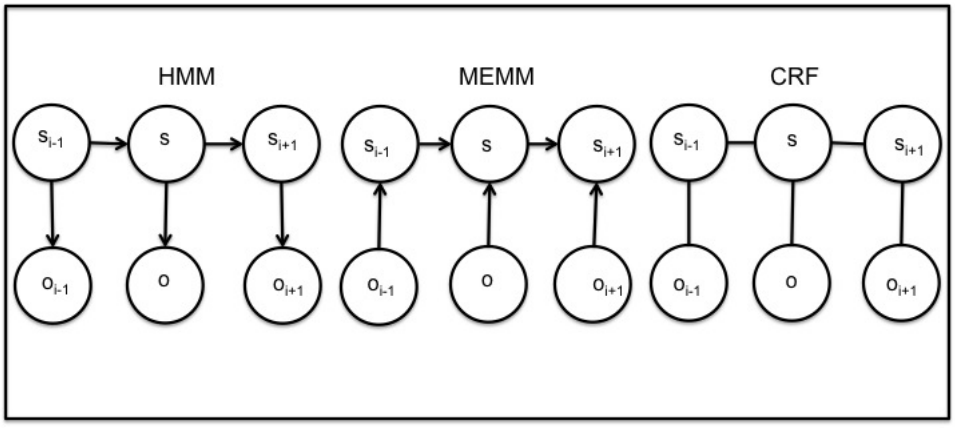

Figura 2.1 – Representação Gráfica das cadeias lineares HMMs, MEMMs e CRFs. (Fonte: [59])

CRF possui todas as vantagens do modelo MEMM, além de tratar o problema sob o viés do rótulo apresentado por esse modelo. Um motivador para a escolha entre as cadeias lineares CRFs e

MEMMs, é que MEMM utiliza modelos exponenciais por estado para as probabilidades condicionais dos próximos estados dado o estado atual, enquanto CRF tem um único modelo exponencial para a probabilidade conjunta de toda a sequência de etiquetas dada a sequência em observação. Assim, os pesos das diferentes *features* em diferentes estados podem ser trocados/negociados uns com os outros. CRFs são descritos em mais detalhes em [63, 70].

Para definirmos CRFs de cadeia linear, temos: $\mathbf{o} = (o_1, o_2, ..., o_T)$ como a lista de sequências de dados de entrada observados (valores dos T nodos de entrada), como, por exemplo, as sequências de palavras em um texto. S é um conjunto de estados FSM, em que para cada estado é associada uma etiqueta (etiqueta $L \in$ a um conjunto de etiquetas ζ), como, por exemplo, as categorias das entidades em um texto: Local, Pessoa etc. Por fim, $\mathbf{s} = (s_1, s_2, ..., s_T)$ é a lista de sequências de estados, que correspondem aos valores dos T nodos de saída.

CRFs de cadeia linear então definem a probabilidade condicional de uma sequência de estados, dada uma sequência de entrada, na forma de p ($\mathbf{s}|\mathbf{o}$):

$$p\ (\mathbf{s}|\mathbf{o}) = \frac{1}{Z_o} \exp\ (\ \sum_{t=1}\ \sum_{k=1}\ \lambda_k\ f_k(s_{t-1}, s_t, \mathbf{o}, t)),\ \text{onde:}$$

- Z_o é o fator de normalização sobre todos os estados;
- $f_k(s_{t-1}, s_t, \mathbf{o}, t)$ é uma função de *feature* arbitrária sobre esses argumentos,
- $\lambda_k \in (-\infty;\ +\infty)$ é o peso aprendido para cada função de *feature*.

Note que, o fator Z_o corresponde à soma dos *scores* de todas as possíveis sequências de estados, e que o número de sequências de estado é exponencial ao comprimento T da sequência de entrada:

$$Z_o = \sum \exp\ (\ \sum_{t=1}\ \sum_{k=1}\ \lambda_k\ f_k(s_{t-1}, s_t, \mathbf{o}, t)),\ \mathbf{s} \in S^T.$$

As funções de *features* f_k, em geral, podem realizar perguntas arbitrárias sobre a sequência de entrada, incluindo consultas sobre as palavras anteriores, as próximas palavras, assim como as suas combinações. Dessa forma, pode-se verificar as palavras numa janela de duas posições à direita e à esquerda da posição atual *t*. Entretanto, a função de *feature* não depende do valor de *t*, pois esse valor é utilizado somente como índice para as observações **o**. As funções de *features* podem ter valores de -∞ a ∞, entretanto, valores binários são mais tradicionais. Para exemplificar as funções de *feature*, duas definições são descritas a seguir:

- A observação **o** na posição *t* é a palavra "empresa":

  $f_k = \{$ *1*, se e somente se, $o_t =$ *"empresa"*;

  0, caso a condição não ocorra.$\}$

- A observação **o** na posição *t* é uma palavra contida em uma lista de países:

  $f_k = \{$ *1*, se e somente se, s_{t-1} é o estado #1 (por exemplo, a etiqueta OUTRA), e s_t é o estado #2 (por exemplo, a etiqueta LOCAL);

  0, para os demais casos.$\}$

As duas *features* apresentadas são *features* categoriais, uma vez que retornam um valor binário. No primeiro, é verificada a ocorrência da própria palavra em um determinado segmento, retornando *1* se isto ocorre, e *0* caso contrário. No segundo exemplo, a *feature* baseia-se num método simples como o uso de *gazetteer* como *feature*, ou seja, uma lista pré-definida de termos como nomes de países. Nesse caso, considerando que os pesos mais altos λ da função de *feature* tornam as suas correspondentes transições FSM mais prováveis, então o peso λ_k nesse exemplo poderá ser positivo, uma vez que as palavras que ocorrem numa lista de nomes de países tendem a ser consideradas como entidades da categoria Local.

Nesse contexto, para a aplicação do modelo probabilístico CRF para ER considera-se que cada palavra de uma sentença é uma observação $\boldsymbol{o}$, que recebe etiquetas L seguindo uma notação. Destaca-se a notação BIO [86], que pode ser utilizada para indicar as palavras que constituem uma relação [4, 66], podendo ser representada da seguinte forma: a etiqueta *B-R* indica o início da relação, a etiqueta *I-R* indica a continuidade da relação, e a etiqueta *O* indica que a palavra não faz parte da relação.

Para exemplificar, na Figura 2.2 é ilustrada a representação gráfica do CRF para ER de parte do texto descrito em (1), em que o destacado em negrito no texto corresponde a sequência de palavras que indicam a relação "vinculo_inst" (relação de vínculo institucional ou filiação) da trilha ReRelEM [46]. Essa sequência de palavras, apresentada em (1), representa as entradas observadas para o CRF. Em (2) são ilustradas as etiquetas de saída seguindo a anotação BIO e as entidades nomeadas das categorias Pessoa e Organização envolvidas nessa relação, as quais são representadas pelos argumentos *arg1* e *arg2*, respectivamente. Essas etiquetas BIO dadas para a sequência de palavras de entrada representam os nodos de saída do CRF. Como resultado, considerando a tripla (*argumento1*, *REL*, *argumento2*) temos (*Ronaldo Lemos*, *diretor de o*, *Creative Commons*).

"... Ronaldo Lemos, **diretor da** Creative Commons ..."(1)

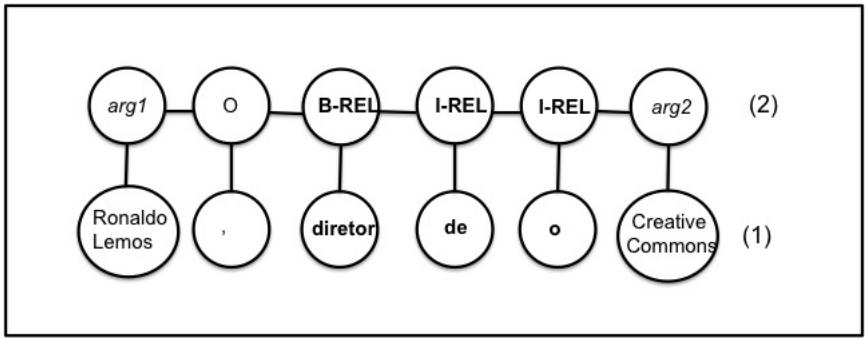

Figura 2.2 – Exemplo de Representação Gráfica do CRF para ER.

Conforme apresentado, a tarefa de ER pode ser tratada como um problema de etiquetagem de sequências, e o CRF, em especial, é um modelo probabilístico capaz de aprender a identificar as palavras que formam uma relação que ocorre entre ENs reconhecidas em textos. A seguir é definida a tarefa de extrair descritores de relação entre ENs do Português aplicando o modelo CRF no contexto desta tese.

2.4 Definição da Tarefa de Extração de Relações no Contexto desta Tese

Devido a variedade de abordagens para ER e os diferentes tipos de relações tratadas, conforme ilustrado na Seção 2.2, definimos nesta Seção o escopo específico da tarefa de ER deste trabalho.

Esta tese de doutorado baseou-se principalmente no trabalho de Li et al. [66] em que o modelo CRF é aplicado para a extração de descritores de relação, que expressam relações mais específicas contidas em textos da Língua Inglesa. Os autores estudam como extrair descritores de relação dados dois argumentos/parâmetros (entidades nomeadas das categorias Pessoa e Organização). Além deste trabalho, os trabalhos de [4, 29] mostraram-se relevantes para o contexto desta tese, os quais são apresentados em detalhes na Seção 3.3.

A partir do trabalho de Li et al., é proposto um processo para extração de descritores de relação entre ENs do domínio de Organizações em textos da Língua Portuguesa, aplicando o modelo CRF. Nesse contexto, definiu-se a tarefa de extração de descritores de relação da seguinte forma:

- Consideramos descritor de relação um segmento de texto de uma sentença, que descreve uma relação explícita entre duas ENs;
- O descritor de relação deve ocorrer no texto situado entre o par de ENs na sentença;
- As relações expressas entre o par de ENs não foram definidas previamente;
- As ENs envolvidas na relação, as quais representam os parâmetros da relação foram definidas previamente. Considerou-se as ENs das categorias Organização, Pessoa e Local, as quais mostraram-se relevantes para o domíno de Organizações [27].

É importante destacar que no trabalho de Li et al. são extraídos os descritores de relação entre as ENs considerando-se classes de relações pré-definidas (relações “employment" e “Personal/Social“), e nesta tese de doutorado são extraídos os descritores de relação que expressam qualquer tipo de relação entre as ENs.

Com base na definição da tarefa de ER apresentada, o processo proposto gera como etapa final um modelo baseado em CRF. Esse modelo é validado aplicando o método de validação cruzada sobre o corpus de referência da tese (ver Seção 5.5). Para isso, um conjunto de *features* foram definidas com base em trabalhos de ER para a Língua Inglesa [4, 22, 66, 67, 73], os quais são apresentados no Capítulo 3. Dentre eles, destaca-se o trabalho de Mintz et al. [73] que utiliza *features* baseadas em padrões lexicais, sintáticos e nas categorias das ENs. Destaca-se também o sistema O-CRF [4] que utiliza um compacto conjunto de padrões léxico-sintáticos para expressar relações não definidas previamente.

Este capítulo apresentou os fundamentos conceituais sobre as tarefas de REN e de ER; uma descrição sobre o modelo CRF. Por fim, a definição da tarefa de ER no escopo específico desta tese foi apresentada. Trabalhos relacionados à tarefa de ER são apresentados no Capítulo 3.

3. Trabalhos Relacionados

3.1 Sistemas/ Abordagens Computacionais para Extração de Relações

Atualmente, pesquisas em ER focam principalmente em aprendizado de padrões e na combinação de técnicas para ER entre pares de entidades a partir de textos livres, como uma coleção de artigos jornalísticos ou páginas Web, entre outros. Em abordagens de aprendizado de máquina, padrões para a identificação de relações não são escritos manualmente ("*handcrafted*"), mas são aprendidos a partir de exemplos etiquetados [75]. Diferentes abordagens de aprendizado de máquina têm sido utilizadas para ER, tais como HMM [43], CRF [63,66], MEM [61], SVM [30,50,101,109], K-Vizinhos mais próximos (*k-Nearest-Neighbors* - KNN) [109].

Em sistemas de ER, geralmente, a relação-alvo é dada como entrada juntamente com padrões de extração [2, 13]. Tais entradas são específicas para a relação em foco; consequentemente, o reconhecimento de uma nova relação requer que sejam desenvolvidos novos padrões de extração ou especificados novos exemplos de treinamento, ambos de forma manual. Dado esse cenário, a necessidade de intervenção manual cresce linearmente com o número de relações-alvo [4].

Para ER utilizando a abordagem supervisionada, os trabalhos existentes frequentemente utilizam a base de referência do ACE para avaliação [30, 50, 85, 107, 109]. Um conjunto de tipos de relações é definido e a tarefa de ER consiste em identificar pares de entidades que estão relacionados, e, posteriormente, classificar essas relações em tipos pré-definidos. Por exemplo, Zelenko et al. [107] introduziram os métodos de *kernel* para ER, em que o problema de ER é tratado como um problema de classificação de pares de entidades, utilizando para isso atributos de árvores de dependência superficial das sentenças (*shallow parse tree*). Cullota e Sorensen, em [30], estenderam a ideia da análise de árvores de dependência superficial, utilizando uma versão um pouco mais geral de *Kernel* de [107].

Recentemente, abordagens semi-supervisionadas e *bootstrapping* estão tendo uma atenção especial. Sistemas de ER baseados em *bootstrapping* são capazes de processar um grande corpus ou milhões de páginas retiradas da Web de forma eficiente, requerendo pouca intervenção humana [2, 13, 36, 37, 83, 97, 110].

Bootstrapping inicia com um pequeno número de exemplos, chamados sementes, e utiliza essas sementes para o treino do modelo inicial. Esse modelo é utilizado para treinar alguns dos dados não etiquetados. Então, o modelo é re-treinado utilizando as sementes originais e os exemplos auto-rotulados. Esse processo é repetido, e gradualmente é expandida a quantidade de dados etiquetados. Um exemplo disso é o sistema DIPRE (*Dual Iterative Pattern Relation Expansion*), que busca identificar padrões de citações de livros na Web [13] utilizando a técnica *bootstrapping*. As relações "autor-título" são extraídas de 24 milhões de páginas Web iniciando com um conjunto de 5 livros. O sistema SNOWBALL [2] extrai a relação "organization-headquarters" de páginas da Web e inclui o uso de REN.

A abordagem *bootstrapping* para tratar padrões genéricos também é proposta. Espresso é um sistema fracamente supervisionado, de propósito geral, que extrai relações semânticas de textos [83]. KnowItAll é um sistema não supervisionado, independente de domínio que extrai fatos da Web [36, 37], o qual possui a particularidade de utilizar uma nova forma de *bootstrapping* que não requer informação anotada manualmente para a etapa de treinamento.

Em geral, sistemas de ER baseados em *boostrapping* apresentam problemas semânticos (chamados de *semantic drift*) [31, 71] após muitas iterações. Em [18] é proposto que o problema possa ser tratado pelo algoritmo de aprendizado semi-supervisionado CBL (*Coupled Bootstrap Learner*), o qual possui como entrada uma ontologia inicial sobre o domínio de Empresas e Esportes, e um corpus composto por 200 páginas da Web. Motivados pelo fato de esses sistemas retornarem um número significativo de erros ou poucos exemplos de relações relevantes, um método baseado em grafos para ordenar as instâncias de relações retornadas pelo sistema de ER *bootstrapping* é proposto em [64]. Ordenar todos os pares de entidades extraídos pela relevância dada pelas sementes é útil para filtrar os exemplos irrelevantes.

Além do *boostrapping*, uma outra abordagem promissora fracamente supervisionada, também chamada de supervisão distante (do inglês "distant supervision"), é aplicada para ER. Segundo Mintz et al., em [73], se duas entidades participam de uma relação, qualquer sentença que contenha essas duas entidades também expressa essa relação particular. Mintz et al. utilizam *features* extraídas de diferentes sentenças contendo pares de entidades da base de conhecimento Freebase [12] para construir um vetor de *features*. Essas *features* baseiam-se em informações lexicais, sintáticas e de REN. Em [55] é apresentado o sistema MultiR para aprendizado multi-instâncias, o qual também utiliza a abordagem fracamente supervisionada a partir da base Freebase e utiliza as *features* propostas em [73].

Na literatura, encontramos sistemas que utilizam métodos que não necessitam de corpora anotados ou de um conjunto inicial de exemplos etiquetados, além de as relações não serem pré-definidas. Dentre eles, podemos citar o sistema FASTUS (*Finite State Automaton Text Understanding*), baseado em autômatos de estados finitos [54], e métodos totalmente não supervisionados, como o proposto por Hasegawa et al. em [51].

Uma abordagem para ER independente de relação foi proposta por Banko et al. em [3], denominando-se *Open Information Extraction* (Open IE). Essa abordagem é ideal para grandes corpora como a Web, que contêm um número expressivo de relações de interesse que não são previamente conhecidas e exploradas.

DIPRE, Snowball, Espresso e KnowItAll, apresentados anteriormente, são todos sistemas que tratam relações específicas. O primeiro sistema Open IE foi o sistema TextRunner [3, 106], o qual utiliza o classificador *Naive Bayes* juntamente com *features* baseadas em POS e *NP-chunker*.

Abordagens estatísticas de aprendizado de máquina são utilizadas em sistemas Open IE. O sistema SatSnowball [110], por exemplo, estende o sistema SNOWBALL com a adição do uso de método estatístico para ER entre entidades, especificamente *Markov Logic Network*. Em [4], Banko e Etzioni apresentam o sistema O-CRF baseado no modelo probabilístico CRF. Os autores mostram

que muitas relações podem ser categorizadas utilizando um compacto conjunto de padrões léxico-sintáticos. Em [102, 103] é proposta uma abordagem para Open IE que utiliza a Wikipedia como um recurso para o treinamento dos dados. Os autores apresentam o sistema WOE[1], o qual gera exemplos de treinamento a partir da informação dos *Infoboxes* da Wikipedia e do correspondente texto. A partir desses exemplos o sistema WOE pode aprender dois tipos de extratores: WOE$_{\text{PARSE}}$, um classificador de padrões que aprende a partir das *features* geradas por árvores de dependência sintáticas (*dependency-parse trees*); e WOE$_{\text{POS}}$, um extrator que é treinado aplicando o modelo CRF a partir de *features* baseadas na anotação de POS.

Um trabalho recente, descrito em [39] mostra que a saída de sistemas Open IE (como TextRunner e WOE) possui muitas extrações incoerentes. Para tratar esses problemas, os autores implementaram restrições sintáticas e lexicais no sistema ReVerb. Tais restrições servem para dois propósitos: (i) eliminam extrações incoerentes, e (ii) reduzem o número de extrações informativas por meio da identificação de relações nas sentenças que apresentam a combinação Verbo-Substantivo.

Um sistema Open IE multilíngue baseado em dependência (DepOE) foi proposto em [47], e utiliza o parser de dependência DepPattern[2]. DepOE foi utilizado para extrair triplas da Wikipedia em quatro línguas: Português, Inglês, Espanhol e Galego.

Em [105] foi desenvolvido um protótipo para ER utilizando padrões léxico-sintáticos de textos anotados com POS, denominado de LSOE (*Lexical-Syntactic patterns based Open Extractor*). A estratégia proposta baseia-se em dois tipos de padrões: (i) padrões genéricos para identificação de relações não específicas, e (2) regras baseadas na proposta de Cimiano e Wenderoth [26] para aprender estruturas Qualia. Os resultados da extração foram comparados aos dos sistemas ReVerb e DepOE.

Conforme apresentado, os trabalhos com Open IE, em geral, utilizam *features* sintáticas para a ER. Christensen et al. [24, 25] investigam o uso de *features* semânticas para a tarefa de Open IE, especificamente a aplicação de *Semantic Role Labeling* (SRL). SRL consiste em detectar argumentos semânticos associados a um verbo em uma sentença, e classificá-los como *agente*, *paciente*, entre outros. Em Mausam et al. [69] é apresentado o sistema OLLIE (*Open Language Learning for Information Extraction*) e um comparativo de seu desempenho com o trabalho de Christensen et al. [25], que utiliza SRL, e com os sistemas WOE$_{\text{PARSE}}$ e ReVerb, que representam o estado da arte em Open IE. Os autores tratam duas limitações dos sistemas Open IE: expandem o escopo sintático, identificando, além de relações verbais, também relações expressas por nomes e adjetivos, e incluem informações do contexto das sentenças na etapa de extração.

Diferentemente dos demais sistemas Open IE, uma abordagem baseada em um conjunto de orações (e seus tipos) que ocorrem nas sentenças é apresentada em [32]. Considera-se oração uma parte da sentença que expressa uma informação coerente, e é constituída de um sujeito, um verbo e, opcionalmente, de um objeto indireto, um objeto direto, um complemento, e um ou mais advérbios. O sistema ClausIE (*Clause-based Open Information Extraction*) baseia-se em um *parser*

[1] *Wikipedia-based Open Extractor*
[2] http://gramatica.usc.es/pln/tools/deppattern.html

de dependências e em um pequeno conjunto de léxicos independente de domínio.

Várias abordagens têm sido utilizadas para a ER para o Inglês, conforme descrito nas seções anteriores. Em contraste, existem poucas propostas para ER para o Português. No contexto deste trabalho, entre as abordagens apresentadas, destaca-se o uso do modelo probabilístico CRF, uma vez que na literatura encontramos diferentes aplicações deste modelo com sucesso, que incluem ER. Tomando como base os trabalhos relacionados de ER para o Inglês, adaptamos para o Português diferentes conjuntos de *features* para a aplicação do modelo CRF. Em especial as *features* descritas nos trabalhos de Mintz et al. [73], Banko e Etzioni [4], Chen et al. [22], Liang e Weld [67] e Li et al. [66]. Os trabalhos que aplicam CRF [4, 22, 66, 67] são apresentados em detalhe na Seção 3.3.

3.2 Sistemas de Extração de Relações para o Português

Conforme dito anteriormente, muitas abordagens têm sido propostas para ER para a Língua Inglesa, as quais foram apresentadas na seção anterior. Em comparação, existem poucas propostas de ER para o Português. Um dos principais obstáculos para o avanço das pesquisas é a falta de recursos disponíveis como dados anotados em Língua Portuguesa. Existe também uma demanda para o desenvolvimento de novas técnicas, ferramentas e recursos mais especializados como bases lexicais e ontologias de domínio. Um estudo sobre o estado da arte de ER incluindo sobre a Língua Portuguesa foi publicado em [1], no qual são abordados os avanços e as dificuldades da área.

Nesta seção são apresentadas abordagens utilizadas na tarefa de ER para o Português pelos sistemas que participaram da trilha ReRelEM [14, 16, 20], e também trabalhos que abordam essa tarefa disponíveis na literatura [6, 41, 48, 89, 93, 100, 104]. É importante salientar que, o conjunto de relações foi previamente definido na maioria dos sistemas que tratam o Português. O único sistema que aplica a abordagem Open IE para o Português é o sistema DepOE, apresentado anteriormente na Seção 3.1.

Geralmente, abordagens de ER utilizadas por sistemas para o Português são baseadas em regras [14, 16, 20]. De acordo com [92], muitas tarefas de extração podem ser executadas utilizando um conjunto de regras, que podem ser codificadas manualmente ("*hand-coded*") ou aprendidas por meio de exemplos. Estes sistemas aplicam heurísticas simples que exploram evidências de relações entre ENs em textos, abrangendo diferentes análises: análise lexical, sintática e semântica, tipo de ENs, informações de fontes externas. Como fonte externa é importante destacar a Wikipedia em Português que fornece um grande número de informações estruturadas, bem como ontologias que provêem nomes.

O sistema REMBRANDT (Reconhecimento de Entidades Nomeadas Baseado em Relações e ANálise Detalhada do Texto) [16] foi desenvolvido para reconhecer todo o tipo de ENs em textos em Português e tipos de relações entre elas. Esse sistema utiliza a Wikipedia em Português como um recurso externo, bem como regras gramaticais que descrevem evidências internas e externas sobre as ENs.

O trabalho de Cardoso [17] descreve que o REMBRANDT é agora uma ferramenta madura, e que

pode ser utilizado pela comunidade de PLN em várias tarefas de EI. O REMBRANDT é composto da ferramenta REMBRANDT NER (participante do Segundo HAREM); RE-NOIR, um módulo de consulta semântica para recuperação de documentos; SASKIA, uma base de conhecimento para todos os recursos de conhecimento e dados armazenados; um indexador que gera termos padrão e índices semânticos para todas as entidades extraídas, e um módulo de recuperação e de ordenação, chamado de LGTE[3] (*Lucene with GeoTemporal Extensions*).

O sistema SeRELeP (Sistema de Reconhecimento de RElações em textos de Língua Portuguesa) [14] objetiva reconhecer três relações da trilha ReRelEM: "identidade", "inclusão" e "localização".As etapas de identificação/classificação das ENs foram realizadas utilizando o parser PALAVRAS [8]. Para o reconhecimento das relações o SeRELeP utiliza regras heurísticas simples baseadas nas informações contidas no texto e informações adicionais fornecidas pelo parser PALAVRAS.

Diferentemente, o sistema SEI-Geo [20] foca no enriquecimento de ontologia [21]. SEI-Geo é um sistema de extração que trata REN, especificamente ENs da categoria Local e suas relações. O SEI-Geo utiliza Geo-ontologias, possibilitando a exploração de relações existentes entre locais reconhecidos em textos, a partir das relações contidas na ontologia.

Da mesma forma que o SEI-Geo, sistemas que focam na ER a partir de textos podem auxiliar na procura de instâncias para ontologias. Na literatura, encontramos poucos estudos que tratam a população de ontologias para o Português. Xavier e Lima [104] apresentam um método semi-automático para extrair e popular ontologias de domínio a partir da estrutura das categorias da Wikipedia em Português. Nesse trabalho, a extração/população é de uma ontologia de turismo contendo classes, relações e instâncias de Localização, especificamente as relações "localizado-em" e "é-um". A tarefa de instanciação é executada no mesmo estágio em que as relações entre os conceitos são extraídas por meio de heurísticas.

Em [41] um sistema para extração de informações a partir de relatórios médicos foi apresentado. Os autores relatam uma coleção dourada no escopo do projeto MedAlert, em que os documentos clínicos relativos aos episódios de hospitalização são anotados com suas múltiplas entidades e relações. As entidades de interesse do MedAlert foram definidas como um objeto real referido no texto, por exemplo, o medicamento mencionado, os exames realizados etc. As relações são as conexões entre essas entidades, como por exemplo, os resultados de um exame ou um medicamento indicado para determinada patologia. Para a extração automática de entidades e de relações, foi utilizado o sistema REMMA[4] (Reconhecimento de Entidades Mencionadas do MedAlert).

Em [89] é apresentado um sistema que identifica relações familiares em textos em Português. Documentos históricos e biográficos são exemplos de textos ricos neste tipo de relação. No HAREM, a categoria Família foi apresentada como uma subcategoria de Outra [46], não sendo tratados os tipos específicos de relação familiar como "pai", "mãe" etc. Nesse trabalho, os tipos de relações familiares foram tratados utilizando a abordagem baseada em regras. Um conjunto de *features* para a extração de tais relações em textos biográficos foi proposto, envolvendo informação de POS,

[3] http://lucene.apache.org/

[4] O sistema REMMA foi inicialmente desenvolvido para participar do Segundo HAREM.

estruturas sintáticas como construções de aposto, núcleo da sentença, entre outros.

Uma metodologia multilíngue para adaptar um sistema de extração automática de eventos para novas línguas, incluindo o Português é apresentada em [100]. Essa tarefa compreende a identificação das ENs e as relações entre elas. Os autores criaram o sistema NEXUS, que faz parte do *Europe Media Monitor Family of Applications*[5] (EMM). O NEXUS objetiva identificar eventos violentos, desastres naturais e crises humanitárias em textos de notícias. Atualmente, o NEXUS trata quatro línguas (Inglês, Francês, Italiano e Russo). Como resultado desse trabalho, o sistema foi adaptado para as línguas portuguesa e espanhola.

As enciclopédias como a Wikipédia cada vez têm sido mais utilizadas para extrair informações, como relações semânticas. Em [48] um sistema de extração de relações semânticas a partir de resumos da versão portuguesa da Wikipédia é apresentado. Esse sistema é centrado num conjunto de gramáticas semânticas, construídas a partir de padrões, e enquadra-se num projeto para construção automática de uma ontologia lexical para o Português.

Uma proposta de abordagem de supervisão distante para a classificação de relações entre duas ENs em artigos da Wikipédia é apresentada em [6]. Essa abordagem baseia-se na ideia de encontrar as relações mais semelhantes numa determinada base de dados de relações-exemplo previamente anotadas. O procedimento refere-se ao desenvolvimento e à aplicação de um classificador baseado na votação ponderada dos *K-vizinhos mais próximos*, em que cada instância de relação tem um peso correspondente à semelhança com a relação a ser classificada. Para isso, exemplos de treino foram extraídos da Wikipédia, correspondendo a frases que expressam relações entre pares de ENs da DB-Pédia. Um conjunto de 10 tipos de relações foram selecionadas, as quais foram extraídas e revisadas manualmente, como por exemplo, as relações "localizado-em", "influenciado-por", "sucessor-de", entre outros.

Existem trabalhos para o Português que investigam tarefas semelhantes, como a extração de citações, uma vez que combina diferentes opiniões em torno de novos tópicos, tais como ENs, ER e temas atuais. Extração de citações consiste da identificação da citação e de seu autor no texto [93]. Em [40] é apresentado o primeiro sistema de extração de citações (relação "autor-citação") que utiliza aprendizado de máquina para o Português.

De acordo com o apresentado nesta seção, a maioria dos sistemas de ER para o Português são baseados em heurísticas, utilizam poucos recursos externos como a Wikipedia ou ontologias de domínio (como, por exemplo, Geo-ontologias) e geralmente não usam técnicas de aprendizado de máquina, ao contrário do que ocorre em Inglês. Este trabalho apresenta um processo para a extração de descritores de relação entre ENs em textos do Português, aplicando o modelo probabilístico CRF. Na literatura, existem trabalhos que aplicam este modelo para REN em textos do Português, etapa necessária para ER. Podemos citar para o Português o trabalho de Batista et al. [5], que aplica o CRF para o reconhecimento de ENs geográficas, e, mais recentemente, o sitema NERP_CRF [33], que aplica o CRF no reconhecimento de ENs seguindo as categorias do HAREM. Já para a aplicação do CRF na ER para essa língua não encontramos nenhum trabalho na literatura.

[5] http://emm.newsbrief.eu/overview.html

3.3 Extração de Relações utilizando Conditional Random Fields

Existe um grande interesse em aplicar CRFs para uma variedade de domínios, envolvendo processamento de texto, bioinformática, visão computacional, entre outros. A primeira aplicação em larga escala do CRF para PLN foi realizada por Sha e Pereira em [95], a qual apresenta um analisador sintático ("*shallow parser*") para segmentação de sintagmas nominais em textos. Desde então, o modelo CRF tem sido aplicado em muitos problemas de PLN, destacando-se em diferentes aplicações para ER como a apresentada por Bellare e McCallum em [7] que extrai 12 relações bibliográficas aplicando um extrator CRF, o qual é treinado a partir de registros do BibTeX e pesquisas de citações em artigos. Em [22] é aplicado CRF para extrair relações entre elementos do conhecimento, envolvendo os tipos de relações de "pré-ordem", "ilustração" e "analogia". Algumas das *features* utilizadas por esses autores foram adaptadas para a aplicação do CRF no contexto desta tese de doutorado. Culotta et al. em [29] propõem a aplicação do modelo CRF para extrair relações familiares em textos biográficos (maiores detalhes estão na Seção 3.3.1). Li et al., em [66], também tratam relações familiares, uma vez que utilizam o modelo CRF para extrair relações específicas entre duas ENs baseando-se em relações mais gerais (ver Seção 3.3.3).

O CRF também é aplicado à tarefa de REN, a qual é um componente-chave para ER. Destaca-se o sistema FIGER de REN descrito por Liang e Weld em [67], o qual utilizou o CRF na etapa de segmentação. Os autores apresentam uma avaliação do desempenho do FIGER para a tarefa de ER, em que as categorias das entidades resultantes do FIGER foram utilizadas como *feature* para o sistema de ER denominado MultiR [55]. Destaca-se que algumas das *features* utilizadas pelo sistema FIGER foram adaptadas para a aplicação do CRF no contexto desta tese de doutorado.

Sistemas Open IE também utilizam o modelo CRF, de acordo com o apresentado na Seção 3.1. Podemos citar o sistema WOE$^{\text{POS}}$ [103], que utiliza recursos da Wikipedia e *features* baseadas na anotação de POS para o treinamento do modelo CRF. Destaca-se também o sistema O-CRF, que será detalhado na Seção 3.3.2.

Conforme apresentado, vários trabalhos utilizam CRF para ER para o Inglês. Para outras línguas como o Chinês encontramos alguns trabalhos que aplicam CRF para ER entre entidades [58, 108]. Entretanto, para o Português não temos conhecimento de trabalhos de ER com CRF. Destacaremos a seguir três trabalhos de ER com CRF para o Inglês que são relevantes para o contexto deste trabalho.

3.3.1 *Integrating Probabilistic Extraction Model and Data Mining to Discover Relations and Patterns in Text*

Cullota et al. em [29] propõem a integração de aprendizado de máquina supervisionado que aprende padrões contextual e relacional para extrair relações de textos biográficos. Para isso, um modelo de ER utilizando CRF é proposto, em que, para cada entidade encontrada num texto biográfico, pretende-se predizer que relação, caso houver, está ligada ao tópico da página a partir de um conjunto de relações previamente conhecidas.

É importante salientar que um texto biográfico trata principalmente de uma entidade, considerada entidade principal. Já as demais entidades referidas no texto são tratadas como entidades secundárias, cujo relacionamento se pretende identificar, caso houver, com a entidade principal.

Segundo os autores, esta formulação permite tratar a ER como uma tarefa de sequência de etiquetagem, como, por exemplo, a tarefa de REN. Entretanto, diferentemente da tarefa de REN, as entidades não são etiquetadas como Pessoa, Organização, entre outras categorias. Em vez disso, a etiqueta dada a uma entidade é a sua relação com a entidade principal. Para exemplificar, segue em (3) um trecho de um texto biográfico descrito em [29], no qual em negrito estão destacadas a entidade principal ("George W. Bush"), bem como as etiquetas das relações que ocorrem entre essa entidade e as entidades secundárias "George H. W. Bush"e "Barbara Bush" ("father" e "mother", respectivamente).

George W. Bush

"George is the son of George H. W. Bush (**father**) and Barbara Bush (**mother**)". (3)

Os experimentos com o modelo CRF para ER utilizou 1127 parágrafos retirados de 271 artigos da Wikipedia, nos quais foram anotadas 53 relações (mother, cousin, friend, education, boss, rival etc.), totalizando 4701 instâncias de relações. Para a avaliação dos resultados aplicando o modelo CRF, dividiu-se os dados em treino e teste (70-30, respectivamente), alcançando uma F-measure de 61.36%.

3.3.2 *The Tradeoffs Between Open and Traditional Relation Extraction*

Banko e Etzioni apresentam, em [4], o sistema Open IE denominado O-CRF baseado no modelo CRF. Os autores demonstram a capacidade de extrair uma variedade de relações semânticas entre entidades utilizando um compacto conjunto de padrões léxico-sintáticos. Por exemplo, a presença de um verbo no contexto de duas entidades pode ser um indicativo de uma relação entre elas (*Entidade 1 Verbo Entidade 2*).

Para o treinamento do sistema O-CRF, é aplicado um conjunto de heurísticas no Penn Treebank[6] [68], resultando num conjunto de exemplos etiquetados em forma de tuplas relacionais. Tais heurísticas são obtidas pela anotação da função sintática e semântica, como, por exemplo, a extração de sintagmas nominais participantes da relação *sujeito-verbo-objeto* apresentada em (4), em que as entidades envolvidas estão destacadas em negrito.

"**Einsten** received the **Nobel Prize** in 1921."(4)

O sistema O-CRF anota com a etiqueta ENT o par de entidades envolvidas na relação, e tal par serve para ancorar cada uma das extremidades da cadeia linear do CRF. Já as palavras que ocorrem no contexto da relação (entre as duas entidades em foco) são tratadas como pistas textuais que indicam a relação, e neste trabalho recebem a notação BIO. Para exemplificar, retomemos o

[6] http://www.cis.upenn.edu/ treebank/

exemplo anterior em (4), no qual a sequência de etiquetas dada pelo O-CRF é apresentada em (5). Notemos que as entidades recebem a etiqueta ENT, "received" recebe a etiqueta B-REL, indicando o início da relação, e na sequência "the" recebe a etiqueta I-REL por fazer parte da relação. Já as demais palavras que não fazem parte da relação explícita entre as entidades recebem a etiqueta O.

Einsten received the **Nobel Prize** in 1921

ENT B-R I-R ENT O O (5)

As *features* utilizadas pelo O-CRF são muito similares às utilizadas por sistemas de ER encontrados na literatura [3]. Dentre elas, destaca-se anotação de POS e NP-chunker: neste trabalho, utilizam-se os recursos do OpenNLP [81], palavras do contexto e expressões regulares para detectar pontuação, entre outras. Destaca-se que as *features* utilizadas pelo sistema O-CRF foram adaptadas para o Português e utilizadas na geração do modelo CRF no contexto desta tese de doutorado.

Para os experimentos um conjunto de 500 sentenças selecionadas randomicamente do corpus desenvolvido em [15] foi utilizado. Como resultado, O-CRF alcançou 88.3% de Precisão, 45.2% de Abrangência e 59.8% de F-measure utilizando os quatro padrões mais frequentes de relações observados entre duas entidades: *verbo*; *substantivo + preposição*; *verbo + preposição* e *infinitivo*.

Destaca-se que os autores comparam o resultado alcançado pelo sistema O-CRF com o TextRunner, sistema Open IE que utiliza o classificador Naive Bayes para predizer se as palavras que ocorrem entre duas entidades indicam uma relação ou não [3]. O-CRF alcançou o dobro do valor em Abrangência e um aumento na taxa de Precisão em relação ao TextRunner (Abrangência de 23.2% e Precisão de 86.6%).

3.3.3 *Extracting Relation Descriptors with Conditional Random Fields*

Em [66] é apresentado um estudo sobre o problema de ER em que tipos de relações são definidas em um nível geral. Entretanto, deseja-se extrair relações mais específicas contidas em textos em língua natural. Para exemplificar, tomemos a relação "Employment", uma das mais importantes relações da conferência ACE. Essa relação define o cargo/posição de uma EN do tipo Pessoa ocupado/exercido em uma EN do tipo Organização.

Segundo os autores, dependendo do objetivo da tarefa de ER em bases textuais, pode-se necessitar da informação exata do cargo/posição envolvendo tais ENs, caso essa informação seja mencionada explicitamente no referido texto. Um exemplo de Candidato a Instâncias da Relação, "Employment" entre as ENs Pessoa e Organização, é apresentado em (6), em que tais ENs são representadas por ARG-1 e ARG-2, respectivamente. Além disso, é apresentado o segmento do texto que descreve a relação específica entre as duas ENs relacionadas (entre ARG-1 e ARG-2), o qual os autores chamam de Descritor da Relação.

Candidato a Instância da Relação: "... said ARG-1 , a vice president at ARG-2 , which ...."(6)

Descritor da Relação: “a vice president"

Neste trabalho, os autores relatam o uso do modelo CRF para a extração de descritores de relações ocorridas nas relações Employment (Person, Organization) e Personal/Social (Person, Person). Entretanto, algumas alterações foram aplicadas no modelo CRF empregado, destacando-se a redução do espaço de possibilidades das sequências de etiquetas e a incorporação de *features* de longo alcance.

Para a avaliação do modelo CRF proposto, foi utilizado como *baseline* o modelo CRF padrão apresentado em [4]. Para isso, foram anotadas manualmente duas bases de dados: 150 artigos do New York Times, anotados com 536 instâncias da relação “Employment", e um conjunto de artigos da Wikipédia, utilizado em [29], anotado com 700 instâncias da relação “Personal/Social". Cabe salientar que as ENs Person e Organization foram identificadas com o uso do *Standard NER tagger* [42].

As *features* do *baseline* utilizam informações considerando etiquetas de uma janela de dois elementos (anteriores e subsequentes), dentre as quais destacam-se: a palavra, anotação de POS, estrutura da sentença com os valores das etiquetas seguindo a notação BIO (por exemplo, sintagma nominal - SN: B-SN, I-SN). Já o modelo CRF modificado proposto acrescentou as *features* de longo alcance, em que destaca-se o contexto do descritor da relação (representado como uma unidade única denominada de REL), tais como, a(s) palavra(s) que ocorre entre ARG-1 e REL (por exemplo: “ARG-1 is REL").

Como resultado é apresentado o comparativo entre o *baseline* e o CRF modificado considerando variações das *features* empregadas. Na avaliação do desempenho foram considerados dois critérios de extração correta: comparação exata do descritor da relação, e a comparação mais relaxada, que considera como correta a ocorrência de pelo menos uma palavra em comum com o descritor da relação, ambos considerando como referência a anotação manual. Destaca-se que, na avaliação de ambos os critérios, o CRF proposto alcançou melhores resultados do que o *baseline* (F-measure em torno de 80% para a relação “Employment" e entre 51-53% para a relação “Personal/Social").

Segundo os autores, devido à diversidade das relações extraídas e à independência de domínio, a extração Open IE pode não ser adequada para população de bases de dados ou bases de conhecimento. Já a extração de relações específicas entre duas ENs utilizando tipos de relações mais gerais e pré-definidas de um dado domínio é um caminho para assegurar que as instâncias extraídas possam ser utilizadas para popular bases de dados relacionais.

No contexto desta tese de doutorado, a definição da tarefa de extração de descritores de relação entre ENs do Português baseou-se no trabalho de Li et al., sendo que os parâmetros das relações foram previamente definidos (Organização, Pessoa e Local), mas as relações expressas entre essas ENs não são conhecidas.

Um outro tópico de ER que merece atenção é a forma de avaliação dos sistemas que tratam essa tarefa, bem como a dificuldade de comparação entre diferentes sistemas uma vez que aplicam distintas abordagens utilizando diferentes dados, línguas e formas de avaliação. A seguir será apresentada uma discussão sobre metodologias de avaliação usuais para a tarefa de ER.

3.4 Avaliação da Tarefa de Extração de Relações

A avaliação da tarefa de ER depende de corpora de referência ou bases de dados, os quais funcionam como termo de comparação para análise e avaliação de sistemas que lidam com essa tarefa.

Corpora de referência são necessários para fornecer uma norma com a qual se fará a comparação dos resultados do corpus de estudo. Tal corpus é denominado *Golden Standard*, e contém anotações – geralmente realizadas manualmente por mais de um especialista – seguindo *guidelines* que descrevem o esquema de anotação, bem como a forma de definir o consenso entre os anotadores para a tarefa específica. A base de dados de referência do MUC é um exemplo disso: é utilizada para avaliar as tarefas de REN e de ER dessa conferência. O sistema FASTUS, descrito na Seção 3.1, participou do MUC, e utilizou o corpus de referência dessa conferência para a avaliação.

Conferências de Avaliação Conjunta, como MUC, ACE, TAC, HAREM e *Evaluation Exercises on Semantic Evaluation* (SemEval), reúnem a participação de vários sistemas que são comparados ao executar uma mesma tarefa [90]. O objetivo de uma avaliação conjunta é melhorar o estado da arte da área. Na medida em que promove a pesquisa, produz como resultado metodologias de avaliação, recursos de avaliação reutilizáveis como bases de teste, entre outros. A realização de conferências dedicadas à avaliação de sistemas que envolvem as diferentes tarefas na compreensão da língua tem auxiliado no avanço da área de PLN. A seguir, uma breve descrição de conferências de avaliação conjunta que tratam REN e ER é apresentada.

A primeira importante conferência que definiu a tarefa de avaliação de REN foi a MUC. A sua primeira edição foi em 1987 com o objetivo de desenvolver uma avaliação conjunta na área de EI. Na sua sexta edição, ocorrida em 1995, foi iniciada a avaliação da tarefa de REN exclusivamente para a Língua Inglesa [77]. De uma forma geral, a tarefa de REN iniciada no MUC-6 consistiu em anotar as ENs em três categorias e tipos correspondentes: *Enamex* (tipos: *Person, Organization, Location*); *Timex* (tipos: *Date, Time*); *Numex* (tipos: *Money, Percent*). No MUC-7 foi acrescentada mais uma tarefa referente à identificação de relações entre as categorias (*Template Relation - TR*). Essa tarefa compreende a extração de fatos bem definidos em textos jornalísticos escritos em Inglês, nos quais as relações envolvem Organizações, que são ilustradas na Tabela 2.1 da Seção 2.2.

Outras iniciativas de avaliação que devem ser destacadas são o programa ACE, as sessões de avaliação da conferência TAC e SemEval. ACE teve sua primeira edição em 1999 com a realização de um estudo piloto para a Língua Inglesa. A partir de 2000-2001, o ACE expandiu a definição e escopo da tarefa de REN, envolvendo a identificação/classificação de entidades e expressões anafóricas para o Inglês e Chinês, denominada *Entity Detection and Tracking - EDT*. A definição das classes também foi diferente da proposta no MUC. A tarefa EDT considerou não somente as classes do MUC (*Person, Organization, Location*), mas também duas outras classes e correspondentes tipos[7]: *FAC - Facility* (tipos: *Airport, Building*) e *GPE - Geographical-Political Entity* (tipos: *Continent, District*).

O ACE em 2002-2003 incluiu a tarefa de reconhecimento das relações (Relation Detection and

[7] http://www.itl.nist.gov/iad/mig//tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf

Characterization - RDC) [34], a qual foi realizada até a edição de 2008 (ACE 2008) [80]. A tarefa RDC compreende a identificação/classificação de tipos de relações e correspondentes subtipos entre pares de entidades. A Tabela 2.1 da Seção 2.2 apresenta alguns tipos/subtipos das relações definidas no ACE.

Na sequência do MUC e do ACE, a conferência TAC iniciou em 2008. TAC é uma série de workshops de avaliação organizados para promover pesquisas em PLN e aplicações relacionadas, sendo que sua primeira edição (TAC 2008[8]) focou em três tarefas: (1) Trilha QA - sistemas que retornam respostas precisas de perguntas a partir de grandes coleções de documentos; (2) Trilha RTE - sistemas que reconhecem quando um trecho de texto implica outro; (3) Trilha Sumarização - sistemas que produzem sumários curtos e coerentes do texto.

A conferência TAC ocorre anualmente, e na TAC de 2009 a trilha sobre a população de bases de conhecimento foi adicionada (*Knowledge Base Population - KBP*[9]). Essa trilha promove a pesquisa em sistemas automatizados de detecção de informações sobre ENs (tais como Pessoas, Organizações e Locais) encontradas em grandes corpora, e acrescenta essa informação a uma base de conhecimento. Atualmente, a TAC 2012 foca na trilha KBP, que envolve três áreas (*Entity-Linking, Slot-Filling, Cold Start Knowledge Base Population*), todas com o objetivo de melhorar a capacidade de popular automaticamente bases de conhecimento a partir de textos, incluindo os idiomas Inglês, Chinês e Espanhol.

Recentemente, na avaliação SemEval-2010[10] [53], foi proposta dentre as suas tarefas uma derivada do reconhecimento de relações simples: a tarefa número 8 – *"Multi-Way Classification of Semantic Relations Between Pairs of Nominals"* – que compreende a classificação de relações semânticas entre pares de entidades previamente identificadas. O evento SemEval é uma série contínua de avaliação de sistemas de análise semântica.

Finalmente, para o Português apenas recentemente temos visto esforços similares de avaliação conjunta como o HAREM. O HAREM é uma conferência dedicada exclusivamente para a Língua Portuguesa, e que tem estudado expressões envolvendo nomes próprios. O primeiro evento de avaliação do HAREM iniciou em 2005 e seguiu os critérios de avaliação do MUC, mas com algumas modificações. O HAREM constituiu um marco para a avaliação conjunta para a Língua Portuguesa, uma vez que na literatura encontramos apenas os trabalhos de Palmer et al. [82] e de Bick [9] que são anteriores ao HAREM e tratam da avaliação de REN para o português.

A segunda edição do HAREM ocorreu em 2008 e os sistemas participantes puderam escolher as categorias, tipos/subtipos das ENs. Além disso, também foi adicionada a tarefa de detecção automática de relações semânticas entre ENs - ReRelEM [46]. As relações definidas na trilha ReRelEM são: Identidade (ENs com o mesmo referente, podendo ocorrer somente entre instâncias da mesma categoria), Inclusão (uma EN faz parte de uma outra En, sendo essas da mesma categoria); Localização (localização espacial de uma organização ou evento, ocorrendo entre as ENs das cate-

[8] http://www.nist.gov/tac/2008/index.html
[9] http://apl.jhu.edu/ paulmac/kbp.html
[10] http://semeval2.fbk.eu/

gorias Organização ou Acontecimento e a categoria Local), e Outra (relações que não ocorreram em nenhuma das descritas anteriormente).

Um importante pré-requisito para a avaliação apropriada de aplicações de PLN é conhecer amplamente o problema proposto, pois somente é possível o desenvolvimento de uma boa metodologia de avaliação se o problema analisado foi devidamente quantificado e se as vantagens da abordagem proposta forem identificadas. Em geral, as métricas de avaliação de desempenho utilizadas para avaliar REN e ER são as mesmas da área de Recuperação da Informação [35]. As medidas mais comumente utilizadas para tais avaliações são Precisão, Abrangência e F-measure, definidas da seguinte forma:

Precisão avalia o quanto o modelo acerta:

$$Precisão = \frac{número\ de\ itens\ corretamente\ classificados}{número\ total\ de\ itens\ classificados} \tag{3.1}$$

Abrangência avalia o quanto o modelo contabiliza:

$$Abrangência = \frac{número\ de\ itens\ corretamente\ classificados}{número\ de\ itens\ corretos\ da\ coleção} \tag{3.2}$$

F-measure combina as medidas de Precisão e Abrangência, obtendo um desempenho geral:

$$F - Measure = \frac{2 * Precisão * Abrangência}{Precisão + Abrangência} \tag{3.3}$$

Além das avaliações realizadas no âmbito dessas conferências, muitos trabalhos de pesquisa consideram diferentes bases de dados. Em sistemas supervisionados, a tarefa de ER é expressa como uma tarefa de classificação [75]. Portanto, as medidas (como Precisão, Abrangência e F-measure) podem ser utilizadas para avaliar esses sistemas, uma vez que sistemas supervisionados necessitam de dados de referência para o aprendizado, e esses dados podem ser utilizados para calcular tais medidas. A avaliação de sistemas que utilizam métodos não supervisionados também necessita de um corpus de referência com as informações de interesse anotadas para a sua validação ou da análise manual das relações extraídas automaticamente. Por exemplo, em [51] para a avaliação das relações detectadas automaticamente usando o método de *clustering*, os autores analisaram os dados manualmente.

De forma similar, na aplicação de métodos semi-supervisionados, dificilmente tem-se um conjunto de teste etiquetado para a validação do modelo aprendido. Além disso, métodos semi-supervisionados para ER são tipicamente aplicados para grandes quantidades de dados, tais como páginas da Web, e geralmente resultam em um grande número de novos padrões de relações (tais como Open IE). Portanto, a análise manual desses resultados seria uma tarefa muito custosa. O que se aplica usualmente é a análise manual de uma amostra dos dados. Esse subconjunto pode ser randomicamente

extraído ou baseado em um grupo específico de relações selecionadas de todo o conjunto. Por exemplo, o sistema DIRPE obteve como resultado uma lista com em torno de 15.000 livros, dos quais 20 foram selecionados randomicamente e analisados manualmente.

Na Tabela 3.1, é apresentada uma visão geral da avaliação e as bases de dados utilizadas por alguns trabalhos relacionados para o Inglês, descritos na Seção 3.1. Nota-se que ocorre uma variedade de tipo e tamanho de corpora/dados utilizados, bem como da forma de avaliação, assim os resultados apresentados não podem ser comparados.

Existem trabalhos de ER para o Português que também necessitam de uma avaliação manual das relações extraídas automaticamente, principalmente por não terem recursos disponíveis para o português, como um corpus de referência. Em [45] uma amostra aleatória das relações corretas extraídas automaticamente foi avaliada manualmente, seguindo uma pontuação para as relações (3: correta; 2: um pouco correta; 1: correta em termos gerais; 0: errada). Sistemas que participam de conferências de avaliação conjunta para o Português, como o HAREM, seguem as diretrizes da conferência. Por exemplo, os sistemas REMBRANDT, SEI-Geo e SeRELeP utilizaram a Coleção Dourada do ReRelEM durante a avaliação dessa trilha. Em geral, as relações anotadas por esses sistemas foram comparadas com as da Coleção Dourada, e cada tripla (EN Relação EN) foi avaliada como correta, em falta ou incorreta [44].

Uma outra dificuldade para avaliação dos trabalhos de ER do Português é a comparação dos resultados, pois a maioria dos trabalhos são para outras línguas e os poucos trabalhos para o Português utilizam diferentes recursos. Na Tabela 3.2 apresentamos os dados utilizados, o respectivo método de avaliação e os resultados alcançados pelos trabalhos para o Português, apresentados na Seção 3.2. Nota-se que alguns dos trabalhos ilustrados nessa tabela utilizaram a coleção dourada do HAREM, e assim podem ser comparados [14, 16, 20]. Entretanto, a maioria dos trabalhos utilizou diferentes dados de uma variedade de domínios, dificultando a sua comparação.

Conforme ilustrado nas tabelas, a maioria dos trabalhos que não possuía corpus de referência avaliou manualmente um subconjunto do corpus. Isso se deve à variedade de relações tratadas na literatura (ver Tabela 2.1), e pelo fato de a tarefa de anotação manual ser muito custosa e necessitar de mais de um especialista para o consenso da referida anotação.

Trabalhos	Dados/Corpora	Avaliação	Resultados, %
[54]	MUC-4, MUC-5, MUC-6.	Corpus de referência do MUC.	MUC-4 F= 47,7%; MUC-5 F= 42,67%; MUC-6 F= 51,12%.
[13]	24 milhões de páginas Web.	Avaliação manual de 20 livros selecionados de uma lista de 150,000.	19 livros corretos - 95%.
[2]	North American News.	Avaliação manual de um conjunto de 100 tuplas.	93 tuplas corretas - 93%.
[51]	Artigos do New York Times (NYT) de 1995.	Avaliação manual das relações.	Person-GPE F= 80%; Company-Company F= 75%.
[36, 37]	Páginas Web.	Avaliação automática utilizando bases externas: Tipster Gazetteer, Internet Movie Database.	City F= 85%; State F= 98%; Country F= 82%; Actor F= 90%; Film F= 65%.
[83]	Artigos do TREC-9 e CHEM.	Avaliação manual de 680 instâncias do corpus TREC e CHEM (2 especialistas).	TREC part-of P= 69,9%; sucession P= 49%. CHEM is-a P= 76%; reaction P= 91,4%; production P= 55,8%.
[18]	200 milhões de páginas Web.	Base Freebase como corpus de referência.	Média das categorias P= 83%; Média das relações P= 84%.
[64]	Wikipedia e projeto Yago.	5 tipos de relações extraída pelo projeto YAGO como corpus de referência.	Média das relações = 39%.
[3, 106]	Penn Treebank, 9 milhões de páginas Web.	Avaliação manual de 400 tuplas (3 especialistas).	80,4% tuplas corretas.
[4]	500 sentenças do corpus de EI [15].	Subconjunto do corpus anotado com 4 relações.	Open IE F= 59,8%; relações pré-específicas F= 29,5%.
[110]	Sent500 [15] e Web1M.	Avaliação manual das tuplas extraídas do Sent500.	F= 76,4%.
[102, 103]	WSJ do Penn Treebank, Wikipedia e páginas Web.	Avaliação manual de 300 sentenças de cada corpus (2 especialistas).	WSJ F= 64,7%; Wikipedia F= 57,2%; Web F= 65%.
[29]	1127 parágrafos de 271 artigos da Wikipedia.	Anotação manual de 53 relações familiares.	F = 61,36%
[66]	150 artigos do NYT, artigos da Wikipedia [29].	Anotação manual das relações.	NYT Employment F=80%. Wiki Personal/Social F=51%.
[39]	500 sentenças de páginas Web.	Avaliação manual das relações (2 especialistas).	F= 69,8%.
[47]	Wikipedia em Inglês, Espanhol, Galego e Português.	Avaliação manual de 200 sentenças da Wikipedia em Inglês (2 especialistas).	P= 68%.

Tabela 3.1 – Dados e métodos de avaliação para o Inglês.

Trabalhos	**Dados/Corpora**	**Avaliação**	**Resultados, %**
[14]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Identidade F= 68%, Inclusão F= 45%, Localização F= 31%.
[16]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Identidade F= 73%, Inclusão F= 33%, Localização F= 20%.
[20]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Inclusão F= 45%.
[104]	Textos da Categoria Turismo da Wikipedia.	Corpus de Referência do domínio de Turismo.	F= 85%.
[89]	Textos biográficos da Wikipedia, 110 sentenças do corpus CETEMPúblico.	Avaliação manual das relações familiares.	Wikipedia F= 29%. CETEMPúblico F= 36%.
[41]	corpus MedAlert	Corpus de Referência composto por 20 textos anotados manualmente	Inclusão F = 89%
[100]	artigos de notícias sobre eventos relacionados a desastres	Avaliação comparativa entre o *baseline* Português e os resultados	Feridos F = 51% Sequestrados F= 67% Mortos F = 69% Presos F = 47%
[6]	97.988 frases retiradas da DBPédia da Wikipédia em Português	Subconjunto de teste formado por 625 frases (89.601 relações anotadas manualmente entre entidades)	local-de-enterro F= 67% pessoa-chave-em F= 11% localizado-em F= 92% origem-de F = 81% antepassado-de F = 62% parte-de F = 62% sucessor-de F = 24% parceiro F = 28% outros F = 63%
[40]	GLOBOQUOTES retirados do Globo.com.	Sistema *Baseline* manualmente construído.	Citação-Autor F= 79,02%.

Tabela 3.2 – Dados e métodos de avaliação para o Português.

Neste capítulo foi apresentada uma extensa revisão da literatura sobre a tarefa de ER, envolvendo os trabalhos do Inglês e do Português. Trabalhos que aplicam o CRF na extração de diferentes tipos de relações foram descritos em detalhe. Uma discussão sobre formas de avaliação da tarefa de ER também foi apresentada. Destaca-se o uso de corpus de referência para a avaliação da tarefa de ER. No Capítulo 4 é apresentado o corpus de referência para ENs (HAREM) e a construção do corpus de referência para ER no contexto deste trabalho.

4. Corpus da Pesquisa

4.1 Corpus do HAREM

No âmbito das duas edições da conferência HAREM, destacam-se as coleções douradas de cada edição, ou seja, conjuntos de textos em que as ENs foram manualmente anotadas. A coleção dourada do Primeiro HAREM[1] é constituída de 129 textos pertencentes às variantes lusa e brasileira, envolvendo diferentes gêneros de texto, como por exemplo, jornalísticos, literários, políticos, entre outros.

A coleção dourada do Segundo HAREM[2] é formada por 129 textos escritos em Português do Brasil e Europeu. A essa coleção foram adicionados novos gêneros de texto, tais como blogs, wikis, texto de enciclopédias (Wikipédia), e perguntas utilizadas na avaliação de sistemas de perguntas e respostas, além dos gêneros de texto considerados no Primeiro HAREM.

Em ambas as coleções, as ENs contidas nos textos e seus correspondentes tipos/categorias foram anotados manualmente, seguindo as diretivas da conferência HAREM [91]. A primeira edição do HAREM considerou 10 categorias para as ENs: Abstração, Acontecimento, Coisa, Local, Obra, Pessoa, Organização, Tempo, Valor e Variado, totalizando 5.132 ENs anotadas na respectiva coleção dourada. As categorias utilizadas no Segundo HAREM são similares às do Primeiro HAREM. O número de categorias é o mesmo nas duas avaliações: 10 categorias, as quais não sofreram alterações em relação a sua designação, exceto na categoria Variado, que foi substituída por Outro [19]. Um total de 7.255 ENs foram anotadas na coleção dourada do Segundo HAREM. Destaca-se nessa coleção que, além das ENs também foi anotado manualmente um conjunto de relações expressas entre tais ENs nos textos, as quais foram definidas na trilha ReRelEM [46]. Mais detalhes da conferência HAREM são apresentados na Seção 3.4 .

Para exemplificar a anotação das ENs das coleções douradas do HAREM, retomemos a sentença descrita em (1) retirada de um texto do Segundo HAREM e a correspondente anotação em formato XML ilustrada em (7). Podemos identificar na anotação do trecho do texto, duas ENs destacadas em itálico: (*tag EM*) "*Ronaldo Lemos*" e "*Creative Commons*", a classificação nas categorias Pessoa e Organização, respectivamente (*tag CATEG)*.

> "No próximo sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate (...)."(1)

> <EM ID="ric-42664-163" CATEG="PESSOA" TIPO="INDIVIDUAL"> *Ronaldo Lemos*</EM>, diretor do <EM ID="ric-42664-170" CATEG="ORGANIZACAO" TIPO="INSTITUICAO"> *Creative Commons*</EM> (7)

[1] http://www.linguateca.pt/harem/
[2] http://www.linguateca.pt/HAREM/coleccoes/CDSegundo HAREMReRelEM.xml

Um subconjunto das coleções douradas do HAREM compõe o corpus de referência dessa tese. A esse subcorpus adicionamos outras anotações, tal como descrito na Seção 4.2.

4.2 Anotação dos Dados

A anotação dos dados foi realizada em duas etapas: seleção dos textos das coleções douradas do Primeiro e do Segundo HAREM, e posterior adição da anotação das relações expressas entre determinadas ENs contidas nos textos selecionados nas respectivas coleções.

No contexto desta tese de doutorado, analisou-se somente os textos das duas edições do HAREM, descritas na Seção 4.1, considerando os seguintes gêneros: notícia, opinião, jornalístico, político, expositivo, Web e blogue jornalístico. Esses gêneros foram escolhidos devido ao fato de o foco deste trabalho ser o domínio de Organizações; logo, os textos devem tratar assuntos relacionados a negócios, incorporações, entre outros.

Como resultado desta etapa, foram selecionados 82 textos para constituir o corpus de referência desta tese, os quais são utilizados na etapa de pré-processamento descrita a seguir. A Tabela 4.1 ilustra o número de textos selecionados de cada coleção.

Coleção Dourada	**Número de Textos**
Primeiro HAREM	51
Segundo HAREM	31
Total	**82**

Tabela 4.1 – Corpus de referência.

De posse dos textos que compõem o corpus de referência, o próximo passo é identificar as relações que ocorrem entre pares de ENs em cada sentença desses textos, informação necessária para a aplicação do processo proposto. Como a tarefa de extração de descritores de relação é recente, corpora com instâncias de relações anotadas não estão disponíveis, fazendo-se necessária a anotação dos dados [66].

A etapa de anotação manual das relações foi realizada por duas linguistas e seguiu as seguintes diretrizes: dadas duas ENs que ocorrem em uma mesma sentença, foi identificado o segmento de texto (descritor) que melhor descrevesse uma relação entre essas duas ENs. Cabe lembrar que as ENs já haviam sido identificadas e que as relações anotadas não haviam sido definidas previamente, conforme definido na Seção 2.4 .

A Figura 4.1 apresenta em detalhe os passos da anotação manual, os quais foram aplicados para cada sentença dos 82 textos do corpus de referência desta tese de doutorado. Cabe salientar que nem todas as sentenças dos textos foram consideradas, apenas as que possuíam o par de ENs de interesse. Após a anotação da relação entre os pares de ENs por cada anotador, ocorreu uma discussão das anotações para verificar o consenso entre eles. Os casos em que não ocorreu consenso foram discutidos e anotados novamente, conforme indicado na figura como “resolver diferenças". Por fim, após o consenso da anotação, as instâncias de relações foram consideradas positivas, caso

tenha sido anotado um descritor de relação, ou negativas na ausência deste. Um relato sobre a anotação manual dos dados é apresentado no Apêndice C.

Como resultado desta etapa de anotação de relações, quatro conjuntos de dados foram construídos de acordo com as categorias das ENs envolvidas na anotação da relação da seguinte forma:

- ORG-ORG: anotação das instâncias de relações que ocorrem entre duas ENs de Organização;
- ORG-PES: anotação das instâncias de relações que ocorrem entre o par de ENs Organização e Pessoa;
- ORG-LOCAL: anotação das instâncias de relações que ocorrem entre o par de ENs Organização e Local;
- ORG-PES-LOCAL: união da anotação das instâncias de relações dos três conjuntos: ORG-ORG, ORG-PES e ORG-LOCAL.

Na Tabela 4.2 é apresentado o total de instâncias de relações, o número de instâncias positivas, o número de instâncias negativas para cada conjunto de dados.

Dados	**Total**	**Positivos**	**Negativos**
ORG-ORG	175	90	85
ORG-PES	171	105	66
ORG-LOCAL	170	109	61
ORG-PES-LOCAL	516	304	212

Tabela 4.2 – Número de instâncias de relações dos conjuntos de dados.

Conforme dito anteriormente, neste trabalho as relações não foram previamente definidas, entretanto as relações anotadas nesta etapa foram classificadas como verbais (relações nas quais o principal elemento descritor é um verbo) e não verbais (relações nas quais o principal elemento descritor não é um verbo), as quais foram contabilizadas na Tabela 4.3.

Dados	**Total**	**Relações Verbais**	**Relações Não-Verbais**
ORG-ORG	90	66	24
ORG-PES	105	45	60
ORG-LOCAL	109	37	72
ORG-PES-LOCAL	304	148	156

Tabela 4.3 – Classificação das relações dos conjuntos de dados.

Destaca-se que o conjunto ORG-ORG possui o maior número de relações verbais, isto é, a maioria das relações entre pares de ENs da categoria Organização foram mediadas por um verbo. Os demais conjuntos apresentam o maior número de relações não verbais devido às características das relações identificadas entre os pares de ENs. No conjunto ORG-PES das 60 relações não-verbais, 25 expressam a relação de "vínculo-institucional" entre as ENs das categorias Pessoa e Organização. Já no conjunto ORG-LOCAL das 72 relações não-verbais, 26 indicam a relação de "localização" e 5 expressam a relação de "pertence-a" entre as ENs das categorias Organização e Local.

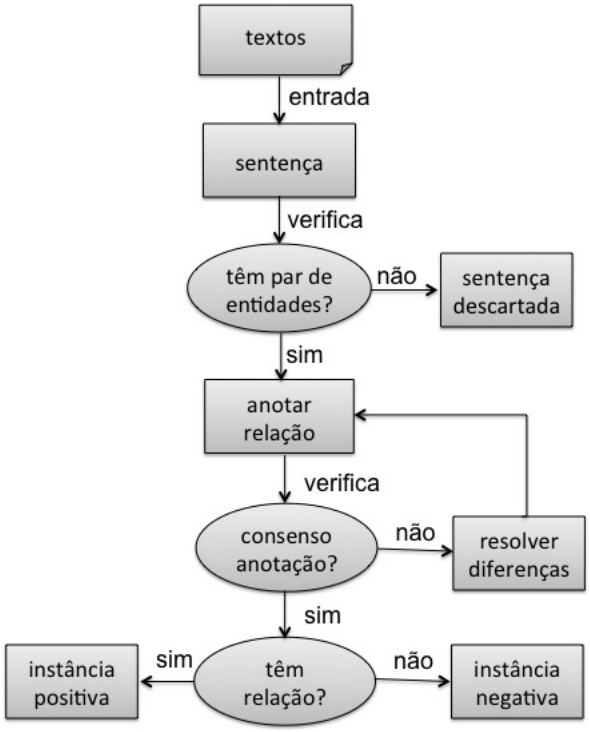

Figura 4.1 – Processo de anotação manual das relações.

Relações	**Dados**	**Porções de Texto de Instâncias de Relação Positivas**	**Descritores de Relação**
Não-verbais	ORG-ORG	... *Comissão de Controle e Gestão Fiscal*, **do** *Ministério da Fazenda*	**do**
	ORG-PES	*Mario Lúcio Vaz*, **diretor da** *Central Globo de Controle de Qualidade* ...	**diretor da**
	ORG-LOCAL	... *Departamento Municipal de Limpeza Urbana* **de** *Porto Alegre* ...	**de**
Verbais	ORG-ORG	... *Força Internacional de Assistência e Segurança* **constitui a** *Força de Reacção Rápida do Comandante* ...	**constitui a**
	ORG-PES	... *Amílcar Cabral* **criou o** *Partido Africano para a Independência da Guiné e Cabo Verde* ...	**criou o**
	ORG-LOCAL	... *Goa Tourism Development Corporation Office* **organiza excursões a** *Goa* ...	**organiza excursões a**

Tabela 4.4 – Exemplos de porções de texto de instâncias de relação positivas.

A Tabela 4.4 e a Tabela 4.5 apresentam exemplos de porções do texto que representam instâncias de relação positivas e instâncias de relações negativas, respectivamente, os quais são utilizados no processo de extração de descritores de relação. No Apêndice A, mais exemplos de instâncias de relações positivas das bases de dados são apresentados.

Dados	Porções de Texto de Instâncias de Relação Negativas
ORG-ORG	*... mudança fulcral em que os contestatários se apoiam para acusar aTransgás proteger a Sonae.*
	... em consequência da reestruturação orgânica operada na Marinha, passou a integrar o Arquivo Central da Marinha ...
ORG-PES	*Nada nos move contra a Transgás, emendou o dirigente agrícola Orlando Gonçalves.*
	Os censores de Zé Gregori elaboram um ofício advertindo a Rede Record ...
ORG-LOCAL	*Uma nova reunião no Monte Sobral dá origem ao Movimento das Forças Armadas.*
	O handebol chegou ao nosso país em meados da década de cinqüenta através dos funcionários da Volkswagen que vindos da Alemanha ...

Tabela 4.5 – Exemplos de porções de texto de instâncias de relação negativas.

Neste capítulo foi apresentado as coleções douradas do HAREM, as quais foram utilizadas como corpus de referência para as ENs. A partir dessas coleções douradas foi selecionado um subconjunto de textos que compõem o corpus de referência dessa tese. As etapas de anotação das relações entre ENs contidas nesse subcorpus foram descritas, resultando em um corpus de referência para extração de descritores de relação do Português. No Capítulo 5, um processo para ER entre ENs do Português que utiliza esse corpus de referência é descrito em detalhe.

5. Processo Proposto

Neste capítulo é descrito o processo de extração de descritores de relação entre ENs da Língua Portuguesa para o domínio de Organizações proposto no âmbito desta tese de doutorado. Para isso, considera-se como descritor de relação o segmento do texto situado entre o par de ENs na sentença e que descreve uma relação explícita entre essas ENs (Pessoa, Organização e Local).

Uma visão geral do processo proposto é ilustrada na Figura 5.1. As etapas de pré-processamento, as *features* utilizadas e a geração do modelo probabilístico CRF são apresentadas em detalhe a seguir.

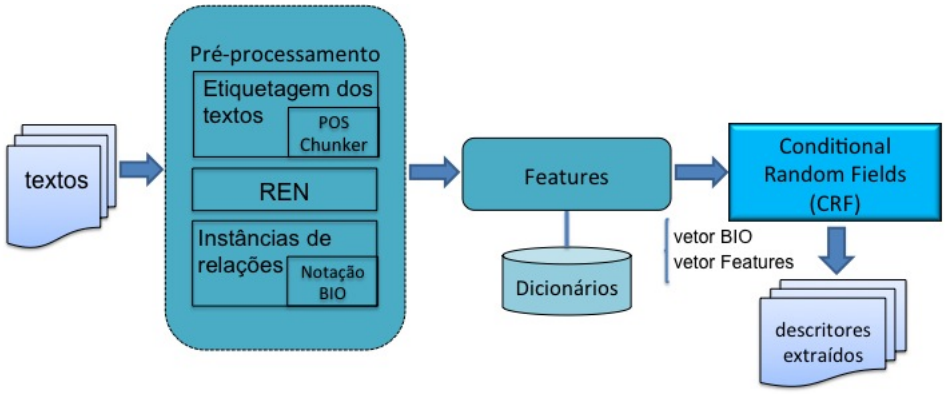

Figura 5.1 – Visão geral do processo proposto.

5.1 Etiquetagem dos Textos

Em geral, no processo de ER a partir de textos se faz necessária uma etapa de pré-processamento, na qual são aplicadas ferramentas de PLN para etiquetagem dos textos. Tais ferramentas provêm informações de partes do discurso (*Part-Of-Speech* - POS), sintáticas (*chunker*) e semânticas dos textos.

Para o Português, existem ferramentas de PLN para etiquetagem dos textos, como, por exemplo, o parser PALAVRAS [8], o etiquetador de POS disponível na biblioteca OpenNLP[1], a anotação de POS baseada na biblioteca FreeLing[2], entre outros.

No contexto desta tese de doutorado, além da anotação de referência provida pelos textos selecionados da coleção dourada das duas edições do HAREM (ver Seção 4.1), é necessária a anotação da informação de POS, sintática e semântica desses textos. Neste trabalho, os textos foram anotados com o parser PALAVRAS [8], o qual provê tais informações. Destaca-se que o parser PALAVRAS fornece etiquetas semânticas para substantivos, nomes próprios, verbos e alguns adjetivos. Podemos citar a etiqueta <Hprof> que indica uma profissão/cargo, que foi utilizada nesta tese de doutorado.

[1]Disponível em: http://opennlp.apache.org/
[2]Disponível em: http://nlp.lsi.upc.edu/freeling/

Em (8) temos um exemplo de anotação do parser PALAVRAS em formato *Constraint Grammar* (CG) aplicado ao trecho do texto ilustrado em (1). Destaca-se que para cada palavra temos as informações dadas pelo parser PALAVRAS indicadas nesta ordem: a própria palavra; a forma canônica da palavra (refere-se ao infinitivo, para os verbos, e ao singular masculino, para as outras palavras variáveis, como por exemplo, artigos e substantivos, entre outros); a etiqueta semântica da palavra; a anotação de POS, seguida da anotação sintática da palavra.

> "No próximo sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate (...). " (1)
>
> (...)
> Ronaldo=Lemos [Ronaldo=Lemos] <hum> PROP @SUBJ>
> ,
> diretor [diretor] <Hprof> N @N<PRED
> de [de] PRP @N<
> o [o] DET @>N
> Creative=Commons [Creative=Commons] <org> PROP @P<
> (...) (8)

Retomando o exemplo ilustrado em (8), temos para a EN "Ronaldo Lemos" as seguintes informações: a própria palavra "Ronaldo=Lemos"; após, a sua forma canônica entre colchetes "[Ronaldo=Lemos]"; a etiqueta semântica <hum>, indicando que a palavra representa o nome de uma Pessoa; a anotação de POS "PROP" indica que se trata de um nome próprio; e, por fim, a anotação sintática "@SUBJ>", indicando que tem função de sujeito da sentença. Cabe salientar que o parser PALAVRAS anota os nomes próprios compostos como uma única EN, como ocorreu com as ENs *"Ronaldo=Lemos"* e *"Creative=Commons"*, procedimento utilizado em outros trabalhos que tratam da ER entre ENs [66]. Já a palavra "diretor" possui a anotação de POS "N" que indica um substantivo e a etiqueta semântica <Hprof> que indica uma profissão/cargo. Essa anotação das ENs será utilizada na próxima etapa do processo de ER proposto.

5.2 Reconhecimento das Entidades Nomeadas

Diversos sistemas que tratam de extração de relações semânticas iniciam o processo com a aplicação de REN para identificar as ENs e os possíveis argumentos de relações contidas em uma sentença, e posteriormente extrair ou classificar o tipo de relação, conforme apresentado no Capítulo 3.

No processo proposto neste trabalho, REN é uma das etapas de pré-processamento e objetiva a identificação das ENs relevantes do domínio de Organizações. Para a definição de que tipos/categorias de ENs devem ser considerados, um estudo sobre o domínio de Organizações foi realizado, no qual foi representado o conhecimento deste domínio de interesse por meio de uma

ontologia [27]. A partir desse estudo, identificou-se os tipos de ENs relevantes para a extração das relações tratadas nesse trabalho, destacando-se, além da categoria Organização, as ENs das categorias Pessoa e Local. Podemos citar a relação de "localização", que ocorre entre as ENs das categorias Organização e Local, a qual é considerada em vários trabalhos, conforme apresentado na Tabela 2.1.

Na literatura existem sistemas de REN que tratam o Português, dentre os quais podemos citar o PALAVRAS-NER [9, 10], REMBRANDT [16], NERP-CRF [33], Freeling[3], Language Tasks[4], que poderiam ser utilizados nesta etapa do processo. Entretanto, nesta tese foi utilizado o corpus de referência de ENs do HAREM, conforme descrito no Capítulo 4, que já possui as ENs anotadas, não necessitando da aplicação automática de um sistema de REN.

Nesse contexto, a identificação dos pares de ENs de interesse foi realizada da seguinte forma: identificou-se em cada sentença dos textos o par de ENs em foco (Organização e Pessoa / Organização e Local / Organização e Organização, não necessariamente nessa ordem) que ocorre mais próximo na mesma sentença do texto, com base na anotação das ENs do corpus de referência da tese. Contudo, somente é considerada uma ocorrência de cada par de ENs na mesma sentença. Assim, quando ocorreram mais de um par de ENs, essas sentenças foram duplicadas considerando os diferentes pares de ENs (Organização e Pessoa / Organização e Local / Organização e Organização).

Para um melhor entendimento, retomemos o trecho do texto (1) anotado pelo parser PALAVRAS em (8). Foi adicionada a essa anotação uma coluna com a categoria da EN (PES, ORG, LOCAL), conforme ilustrado em (9). Destaca-se em negrito no exemplo a EN *"Ronaldo=Lemos"* que recebeu a anotação da categoria PES e a EN *"Creative=Commons"* que recebeu a anotação da categoria ORG.

> "No próximo sábado, Ronaldo Lemos, diretor da Creative Commons, irá participar de um debate (...)."(1)
>
> (...)
> **Ronaldo=Lemos [Ronaldo=Lemos]** <hum> PROP @SUBJ> **PES**
> ,
> diretor [diretor] <Hprof> N @N<PRED
> de [de] PRP @N<
> o [o] DET @>N
> **Creative=Commons [Creative=Commons]** <org> PROP @P< **ORG**
> (...) (9)

Como resultado dessa etapa, os pares de ENs identificados nas sentenças do texto são considerados candidatos a argumentos das relações, e serão analisados na próxima etapa.

[3]Disponível em: http://nlp.lsi.upc.edu/freeling/
[4]Disponível em: http://ltasks.com/

5.3 Representação das Instâncias de Relações

Conforme apresentado no Capítulo 2, a tarefa de extrair descritores de relação entre ENs pode ser tratada como um problema de etiquetagem de sequências. Sendo assim, as palavras que descrevem uma relação (descritores) devem ser etiquetadas utilizando uma notação que as represente.

No contexto desta tese, optou-se por utilizar a notação BIO [86] geralmente utilizada na etiquetagem de sequências, definindo-se então um conjunto de etiquetas para o descritor de relação da seguinte forma:

- **B-REL**: indica o início do descritor da relação;
- **I-REL**: indica que faz parte do descritor da relação;
- **O**: indica que não faz parte do descritor da relação.

De posse da anotação das instâncias de relações positivas e negativas de cada um dos conjuntos de dados (ver Seção 4.2), e da definição da notação BIO, temos que aplicar para cada instância de relação a sequência de etiquetas BIO conforme as seguintes diretrizes:

- Para instâncias positivas, deve-se etiquetar as palavras que constituem o descritor de relação seguindo a notação BIO (B-REL, I-REL, O);
- Para instâncias negativas, ou seja, o segmento de palavras entre o par de ENs não descreve uma relação, deve-se etiquetar cada palavra desse segmento com a etiqueta O (a palavra não faz parte da relação);
- Por fim, todas as demais palavras, pontuações, etc. inclusive as ENs que são os argumentos da relação, devem ser etiquetadas com a etiqueta O, uma vez que não fazem parte do descritor.

Adicionou-se a anotação do descritor da relação ao exemplo ilustrado em (9). Na última posição de cada linha foi adicionada a etiqueta BIO, conforme apresentado no exemplo (10). As linhas que formam o descritor de relação estão destacadas em negrito. Podemos notar que, mesmo as palavras que ocorrem entre o par de ENs, mas não fazem parte do descritor, recebem a etiqueta O, como no exemplo (10): a vírgula não faz parte do descritor e recebeu a etiqueta O. Além disso, as ENs *"Ronaldo=Lemos"* e *"Creative=Commons"* também receberam a etiqueta O.

(...)
Ronaldo=Lemos [Ronaldo=Lemos] <hum> PROP @SUBJ> PES O
, O
diretor [diretor] <Hprof> N @N<PRED B-REL
de [de] PRP @N< I-REL
o [o] DET @>N I-REL
Creative=Commons [Creative=Commons] <org> PROP @P< ORG O
(...) (10)

A quantidade de etiquetas da anotação BIO aplicada às palavras das instâncias de relações dos conjuntos de dados é apresentada na Tabela 5.1. Podemos notar que, em todas as bases ocorreram mais casos da etiqueta O em relação às demais, devido ao fato de que, para os exemplos negativos, deve-se atribuir essa etiqueta para cada palavra que ocorre entre o par de ENs correspondente.

Dados	**B-REL**	**I-REL**	**O**
ORG-ORG	90	281	1702
ORG-PES	105	312	1548
ORG-LOCAL	109	292	1765
ORG-PES-LOCAL	304	885	5012

Tabela 5.1 – Número de etiquetas da anotação BIO no conjunto de dados.

Nesse contexto, temos que, para cada instância de relação anotada (positiva ou negativa), é gerado um vetor com as etiquetas BIO de cada palavra (chamado de "vetor BIO"), com base nos arquivos no formato ilustrado em (10). Para um melhor entendimento do vetor BIO, é apresentado em (11) o vetor correspondente à sequência das etiquetas BIO do trecho de texto descrito em (1):

Vetor BIO:

... Ronaldo=Lemos , **diretor de o** Creative=Commons ... (1)

```
[... O        O  B-REL  I-REL  I-REL     O ... ] (11)
```

Como neste trabalho é utilizado o aprendizado supervisionado no processo de ER proposto, a sequência de etiquetas BIO dos descritores de relação representa as etiquetas/anotação de referência, ou seja, as etiquetas de saída que devem ser preditas na etapa de aprendizado utilizando o CRF. Sendo assim, os vetores BIO são uma das entradas para a etapa de geração e validação do modelo probabilístico CRF, descrita na Seção 5.5.

Nessa etapa de pré-processamento, também gerou-se um vetor com algumas informações das instâncias de relação necessárias para a etapa de geração das *features*, descrita na Seção 5.4. Sendo assim, para cada instância de relação anotada (positiva ou negativa) é gerado um vetor contendo as seguintes informações de cada palavra, nesta ordem:

- **'sintatica':** informação sintática da palavra. Caso não possua essa informação, retorna 'nulo';
- **'POS':** informação de POS da palavra. Caso seja uma pontuação (como vírgula, dois pontos, etc.), retorna essa pontuação;
- **'semantica':** informação semântica da palavra. Caso não possua essa informação, retorna 'nulo';
- **'dicionario':** informação que indica se a palavra está contida no dicionário externo; retorna 'sim/não';

- **'semanticaProf':** informação que indica se a etiqueta semântica é de profissão/cargo; retorna a própria etiqueta (<Hprof> ou <Htit>) ou, caso não possua essa informação, retorna 'nulo';

- **'gerarFeatures':** informação que indica se vão ser geradas as *features* para a palavra em foco; retorna um valor booleano.

- **'categoria':** informação referente à categoria da EN (PES ou ORG ou LOCAL). Caso a palavra não seja uma EN, retorna 'nulo';

- **'palavra':** informação referente à forma canônica da palavra (reduz os verbos à forma infinitiva e as outras palavras variáveis ao singular masculino).

Para a geração desse vetor (chamado de "vetor pre-processo"), extraiu-se tais informações das instâncias de relações dos arquivos no formato ilustrado em (10). Retomando o exemplo, em (12) é apresentado o vetor pre-processo correspondente ao trecho de texto descrito em (1) :

Vetor pre-processo:

... Ronaldo=Lemos , **diretor de o** Creative=Commons ... (1)

... ['sintatica': '@SUBJ>', 'POS': 'PROP', 'semantica': 'hum', 'dicionario': 'nao', 'semanticaProf': 'nulo', 'gerarFeatures': True, 'categoria': 'PES', 'palavra': 'Ronaldo=Lemos'],

['sintatica': 'nulo', 'POS': ',', 'semantica': 'nulo', 'dicionario': 'nao', 'semanticaProf': 'nulo', 'gerarFeatures': True, 'categoria': 'nulo', 'palavra': ','],

['sintatica': '@N<PRED', 'POS': 'N', 'semantica': 'Hprof', 'dicionario': 'sim', 'semanticaProf': 'hprof', 'gerarFeatures': True, 'categoria': 'nulo', 'palavra': **'diretor'**],

[sintatica': '@N<', 'POS': 'PRP', 'semantica': 'sam-', 'dicionario': 'nao', 'semanticaProf': 'nulo', 'gerarFeatures': True, 'categoria': 'nulo', 'palavra': **'de'**],

['sintatica': '@>N', 'POS': 'DET', 'semantica': '-sam', 'dicionario': 'nao', 'semanticaProf': 'nulo', 'gerarFeatures': True, 'categoria': 'nulo', 'palavra': **'o'**],

['sintatica': '@P<', 'POS': 'PROP', 'semantica': 'org', 'dicionario': 'nao', 'semanticaProf': 'nulo', 'gerarFeatures': True, 'categoria': 'ORG', 'palavra': 'Creative=Commons'], ... (12)

No contexto deste trabalho, os vetores pre-processo são importantes para a geração das *features* que descrevem as instâncias de relação, etapa fundamental para o processo de ER proposto.

5.4 Definição das *Features*

O CRF, por ser um método de aprendizado supervisionado, necessita de vetores de atributos que descrevam as características/aspectos dos dados de entrada [74].

No caso do CRF para ER, gera-se os vetores de *features* para as ENs que são os parâmetros da relação e para as palavras que ocorrem entre esse par de ENs na sentença do texto. Retomando o exemplo do trecho do texto descrito em (1), os vetores de *features* são gerados para as ENs: *"Ronaldo=Lemos"* e *"Creative=Commons"* (parâmetros da relação) e para cada uma das palavras que compõem a sequência que ocorre entre tais ENs: *", diretor de o"*.

No contexto desta tese de doutorado, diferentes conjuntos de *features* foram definidos com base na literatura [4, 22, 66, 67, 73], conforme apresentado nas tabelas: Tabela 5.2; Tabela 5.3; Tabela 5.4; Tabela 5.5; Tabela 5.6; Tabela 5.7 e Tabela 5.8.

Para um melhor entendimento da aplicação das *features*, na sua descrição, consideramos que a posição *i* refere-se à posição atual da sequência.

***Features* baseadas em POS**	**Descrição**
Baseada na anotação de POS	a anotação de POS da palavra, na posição *i*
	a anotação de POS da palavra, na posição *i-1*
	a anotação de POS da palavra, na posição *i+1*
	a anotação de POS da palavra, na posição *i-2*
	a anotação de POS da palavra, na posição *i+2*
Baseada em duas consecutivas anotações de POS	a anotação de POS da palavra e da próxima palavra, posições *i* e *i+1*
	a anotação de POS da palavra e da palavra anterior, posições *i* e *i-1*
	a anotação de POS das duas palavras anteriores, posições *i-1* e *i-2*
	a anotação de POS das duas palavras posteriores, posições *i+1* e *i+2*

Tabela 5.2 – Conjunto de *Features* baseadas em POS, adaptado de [66, 73].

***Features* Baseadas no Item Lexical**	**Descrição**
Baseada no Item Lexical	a forma canônica da palavra, na posição *i*
	a forma canônica da palavra, na posição *i-1*
	a forma canônica da palavra, na posição *i+1*
	a forma canônica da palavra, na posição *i-2*
	a forma canônica da palavra, na posição *i+2*
Baseada nos dois Itens Lexicais	a forma canônica da palavra e da próxima palavra, posições *i* e *i+1*
	a forma canônica da palavra e da palavra anterior, posições *i* e *i-1*
	a forma canônica das duas palavras anteriores, posições *i-1* e *i-2*
	a forma canônica das duas palavras posteriores, posições *i+1* e *i+2*
Tamanho do Segmento	número de palavras que compõem o segmento. O segmento contém o par de ENs e a sequência de palavras que ocorrem entre essas ENs.

Tabela 5.3 – Conjunto de *Features* Baseadas no Item Lexical, adaptado de [22, 66, 67, 73].

***Features* Sintáticas**	**Descrição**
Baseada na anotação sintática	a anotação sintática da palavra, na posição *i* a anotação sintática da palavra, na posição *i-1* a anotação sintática da palavra, na posição *i+1* a anotação sintática da palavra, na posição *i-2* a anotação sintática da palavra, na posição *i+2*
Baseada em duas anotações sintáticas consecutivas	a anotação sintática da palavra e da próxima palavra, posições *i* e *i+1* a anotação de sintática da palavra e da palavra anterior, posições *i* e *i-1* a anotação sintática das duas palavras anteriores, posições *i-1* e *i-2* a anotação sintática das duas palavras posteriores, posições *i+1* e *i+2*
Baseada no Núcleo	se a palavra é o núcleo do segmento, na posição *i*. O núcleo é o termo da oração ao qual o predicado designa as propriedades ou relações.
Baseada no Núcleo do Aposto	se a palavra é o núcleo de um aposto, na posição *i*. Aposto é um ou mais termos que se referem a um substantivo ou pronome explicando-o.
Baseada no Aposto	se a palavra faz parte de um aposto, na posição *i* se a palavra faz parte de um aposto, na posição *i-1* se a palavra faz parte de um aposto, na posição *i+1* se a palavra faz parte de um aposto, na posição *i-2* se a palavra faz parte de um aposto, na posição *i+2*
Baseada no Objeto Direto	se a palavra têm função de Objeto Direto, na posição *i* . Objeto direto é o complemento direto de um verbo transitivo.

Tabela 5.4 – Conjunto de *Features* Sintáticas, adaptado de [66, 67].

***Features* Baseadas em Padrões**	**Descrição**
Baseada no Verbo	se a palavra é um Verbo, na posição *i* se a palavra é um Verbo, na posição *i-1* se a palavra é um Verbo, na posição *i+1* se a palavra é um Verbo, na posição *i-2* se a palavra é um Verbo, na posição *i+2*
Baseada no Verbo + Preposição	se a palavra é um Verbo, na posição *i* e a próxima palavra é uma Preposição, na posição *i+1*
Baseada no Verbo + Artigo	se a palavra é um Verbo, na posição *i* e a próxima palavra é um Artigo, na posição *i+1*
Baseada no Verbo + Preposição + Artigo	se a palavra é um Verbo, na posição *i* e a próxima palavra é um Preposição, na posição *i+1*, seguida de um Artigo, na posição *i+2*
Baseada no Substantivo + Preposição	se a palavra é um Substantivo, na posição *i* e a próxima palavra é uma Preposição, na posição *i+1*
Baseada no Advérbio	se a palavra é um Advérbio, na posição *i*
Baseada no Advérbio + Preposição	se a palavra é um Advérbio, na posição *i* e a próxima palavra é uma Preposição, na posição *i+1*
Baseada no Advérbio + Preposição + Artigo	se a palavra é um Advérbio, na posição *i* e a próxima palavra é uma Preposição, na posição *i+1*, seguida de um Artigo, na posição *i+2*

Tabela 5.5 – Conjunto de *Features* Baseadas em Padrões, adaptado de [4].

***Features* Baseadas na Sequência Frasal**	**Descrição**
Baseada na Sequência de POS	a anotação de POS da sequência de palavras entre o par de ENs envolvidos em uma relação (parâmetros)
Baseada na Sequência de POS + POS das ENs	a anotação de POS da sequência de palavras entre o par de ENs envolvidos em uma relação (parâmetros), incluindo a anotação de POS dessas ENs

Tabela 5.6 – Conjunto de *Features* Baseadas na Sequência Frasal, adaptado de [66, 73].

***Features* Semânticas**	**Descrição**
Baseado na anotação semântica provida pelo parser PALAVAS [10]	se a palavra possui a etiqueta semântica de profissão, na posição *i* se a palavra possui a etiqueta semântica de profissão, na posição *i-1* se a palavra possui a etiqueta semântica de profissão, na posição *i+1* se a palavra possui a etiqueta semântica de profissão, na posição *i-2* se a palavra possui a etiqueta semântica de profissão, na posição *i+2*
Baseado na anotação da categoria da EN	se a palavra é uma EN, na posição *i* , retornar a categoria da EN

Tabela 5.7 – Conjunto de *Features* Semânticas, adaptado de [73].

***Features* de Dicionários**	**Descrição**
Lista de cargos	lista de cargos/profissão tipicamente utilizadas, e de títulos de pessoas.
Lista de pistas de Localização	lista de palavras tipicamente usadas para indicar uma localização e nomes de locais.

Tabela 5.8 – Conjunto de *Features* baseadas em Dicionário.

Para exemplificar os vetores de *features*, na Tabela 5.9 são apresentadas as nove *features* baseadas em POS, que compõem os vetores correspondentes à EN *"Ronaldo Lemos"* e à palavra *"diretor"*, em que a primeira é parâmetro da relação e a segunda faz parte do descritor da relação *"diretor de o"*, ilustrado anteriormente no trecho do texto em (1). Os respectivos vetores de *features* são ilustrados em (13) e em (14), em que é apresentada a identificação de cada *feature* aplicada seguida do seu valor.

Vetor de *Features*:

... Ronaldo=Lemos , **diretor de o** Creative=Commons ... (1)

[1: 'PROP', 2: ',', 3: ',', 4: 'N', 5: 'N', 6: 'PROP , ', 7: 'PROP ,', 8: 'N ,', 9: ', N'] (13)

[1: 'N', 2: ',', 3: 'PRP', 4:'PROP', 5: 'DET', 6: 'N PRP', 7: 'N ,', 8: 'PROP ,', 9: 'PRP DET'] (14)

Features	***'Ronaldo Lemos'***	***'diretor'***
1: informação de POS da palavra na posição *i*	'PROP'	'N'
2: informação de POS da palavra na posição *i-1*	','	','
3: informação de POS da palavra na posição *i+1*	','	'PRP'
4: informação de POS da palavra na posição *i-2*	'N'	'PROP'
5: informação de POS da palavra na posição *i+2*	'N'	'DET'
6: anotação de POS da palavra e da próxima palavra, posições *i* e *i+1*	'PROP , '	'N PRP'
7: a anotação de POS da palavra e da palavra anterior, posições *i* e *i-1*	'PROP ,'	'N , '
8: a anotação de POS das duas palavras anteriores, posições *i-1* e *i-2*	'N ,'	'PROP ,'
9: a anotação de POS das duas palavras posteriores, posições *i+1* e *i+2*	', N'	'PRP DET'

Tabela 5.9 – Exemplos de vetor de *features*.

Nesta etapa, utilizam-se as informações contidas nos vetores pre-processo (ver Seção 5.3), em especial o parâmetro "gerarFeatures", que indica se as *features* devem ser geradas para determinada palavra. Cabe salientar que os vetores de *features* gerados são utilizados como entrada para a próxima etapa de geração e validação do modelo probabilístico CRF.

5.5 Geração e Validação do Modelo Probabilístico CRF

A etapa de geração e validação do modelo probabilístico CRF utiliza como entrada o vetor BIO, juntamente com os vetores de *features* que descrevem cada instância de relação dos dados de entrada, resultantes das etapas anteriores. O modelo sequencial CRF é gerado a partir dos vetores de *features*, em que para cada *feature* é atribuído um peso, resultando numa matriz de pesos.

O CRF, a partir dessa matriz de pesos gerada, é capaz de classificar/etiquetar corretamente as palavras que indicam uma menção explícita de uma relação em novos textos, ainda não etiquetados. Sendo assim, o modelo probabilístico CRF gerado no final do processo de extração de descritores de relação proposto pode ser testado/validado ao ser aplicado em novos textos, e utiliza o mesmo vetor de *features* dessa etapa de geração do modelo.

Nesse contexto, para a validação do modelo probabilístico CRF, é utilizado o método de validação cruzada (*cross validation*) descrito da Seção 6.2. Cabe lembrar que foi utilizado o corpus de referência descrito no Capítulo 4, necessário para a geração desse modelo e para sua validação por meio da validação cruzada.

5.6 Descritores de Relação Extraídos

Neste capítulo foi descrito um processo de extração de descritores de relação entre ENs do Português utilizando o modelo CRF. Conforme dito anteriormente, a etapa de aprendizado desse processo tem como entrada o vetor BIO e o vetor de *features* que descreve cada instância de relação.

Na aplicação de validação cruzada, as instâncias de relação de treinamento utilizam esses dois vetores para gerar o modelo CRF, já as instâncias de relação de teste possuem como entrada somente

o vetor de *features*. As etiquetas BIO dos exemplos de teste são preditas pelo modelo CRF gerado no treinamento, correspondendo ao conjunto de etiquetas BIO com maior probabilidade de ocorrerem para cada descritor. Para exemplificar os descritores de relação extraídos, retomemos um trecho do exemplo (1), em que a saída correspondente é apresentada em (15):

... Ronaldo=Lemos , **diretor de o** Creative=Commons ... (1)

Descritor de Relação Extraído:

Ronaldo=Lemos<O>, **diretor**<B-REL>, **de**<I-REL>, **o**<I-REL>, *Creative=Commons*<O> (15)

Neste capítulo foi apresentado um processo para extração de descritores de relação em textos da Língua Portuguesa, os quais descrevem relações explícitas entre ENs do domínio de Organizações (Pessoa, Organização e Local) utilizando o modelo probabilístico CRF. Uma avaliação experimental foi realizada para avaliar os descritores de relação extraídos, considerando a anotação de referência descrita na Seção 4.2. A avaliação experimental é descrita no Capítulo 6.

6. Avaliação Experimental

Neste capítulo é apresentada a avaliação experimental do processo proposto da seguinte forma. Na Seção 6.1 é definida a configuração da avaliação experimental. A avaliação e a discussão dos resultados são apresentadas, respectivamente, na Seção 6.2 e na Seção 6.3. A análise de erros é descrita na Seção 6.4. Por fim, uma comparativo dos resultados alcançados é apresentado na Seção 6.5.

O objetivo da avaliação experimental é aplicar o modelo probabilístico CRF na etapa de aprendizado com base nas *features* definidas. Para isso, as bibliotecas NLTK[1] e Mallet[2] foram utilizadas para implementar o algoritmo CRF.

A avaliação experimental foi realizada considerando que os descritores de relação extraídos foram avaliados com base na anotação de referência descrita na Seção 4.2, seguindo duas diretrizes de avaliação [66]:

- **Descritores Corretos:** o descritor de relação extraído deve ser igual ao descritor de relação positivo anotado manualmente, ou seja, devem ter a mesma sequência de etiquetas BIO;

- **Descritores Parcialmente Corretos:** o descritor de relação extraído deve ter, no mínimo, a mesma palavra etiquetada como B-REL do descritor de relação positivo anotado manualmente.

Cabe salientar que, os descritores parcialmente corretos também foram considerados na avaliação experimental com o objetivo de permitir a avaliação mais inclusiva dos casos em que o CRF consegue identificar que existe uma relação entre os pares de ENs, apesar de não identificar todos os elementos que formam o descritor de relação positivo conforme a referência manual.

Para um melhor entendimento, na Tabela 6.1 é apresentado um exemplo de instância de relação da base ORG-ORG (par de ENs destacado em itálico e descritor da relação destacado em negrito), ilustrando como seria a sua avaliação como descritor correto e como descritor parcialmente correto.

Instância de Relação	**Descritor Correto**	**Descritor Parcialmente Correto**
... o *PSD* passa entre as sombras, ou ficando pura e simplesmente silencioso, ou murmurando umas críticas de circunstância que ninguém ouve, ou, em muitos casos, **concordando com o** *Governo* ...	concordar <B-REL> com<I-REL> o<I-REL>	concordar <B-REL> com<O> o<O>

Tabela 6.1 – Exemplo de Critério de Avaliação dos descritores de relação.

[1] Disponível em: http://nltk.org/
[2] Disponível em: http://mallet.cs.umass.edu/

Conforme o exemplo, para o descritor “concordar com o" ser considerado correto, a sequência de palavras que o formam deve ser anotada com uma etiqueta B-REL, seguida de etiquetas I-REL, nessa ordem. Em contrapartida, para o descritor “concordar com o" ser considerado parcialmente correto, pelo menos o verbo “concordar" deve receber a etiqueta B-REL, seguida ou não de etiquetas I-REL. Podemos notar que, em ambos os critérios o descritor em foco expressa uma relação verbal entre as ENs “PSD" e “Governo", uma vez que, o verbo “concordar" pode vir acompanhado ou não de uma preposição.

6.1 Configuração da Avaliação Experimental

Na avaliação Experimental proposta, diferentes configurações de *features* de entrada para o CRF foram avaliadas. Tais configurações envolveram os diferentes conjuntos de *features* descritos na Seção 5.4, e que foram utilizados na etapa de geração e validação do modelo CRF da seguinte forma:

- *F1=POS:* utilizou-se somente o conjunto de *features* baseadas em POS, descrito na Tabela 5.2;
- *F2=POS+LEX:* adicionou-se o conjunto de *features* lexicais, descrito na Tabela 5.3;
- *F3=POS+LEX+SINT:* adicionou-se o conjunto de *features* sintáticas, descrito na Tabela 5.4;
- *F4=POS+LEX+SINT+PAD:* adicionou-se o conjunto de *features* baseadas em padrões, descrito na Tabela 5.5;
- *F5=POS+LEX+SINT+PAD+FR:* adicionou-se o conjunto de *features* baseadas na sequência frasal, descrito na Tabela 5.6;
- *F6=POS+LEX+SINT+PAD+FR+SEM:* adicionou-se o conjunto de *features* semânticas, descrito na Tabela 5.7.

Cabe salientar que, no processo proposto, o uso do conjunto de *features* baseadas em dicionários (ver Tabela 5.8) foi utilizado em todas as configurações acima, porém em apenas algumas das bases. A base ORG-ORG não utilizou essas *features* baseadas em dicionários. Para a base ORG-PES, foi utilizada uma lista de cargos. Para a base ORG-LOCAL foi utilizada uma lista de pistas de localização. Por fim, a base ORG-PES-LOCAL utilizou ambas as listas disponíveis.

6.2 Avaliação dos Resultados

Nesta seção é apresentada a avaliação dos resultados da extração de descritores de relação entre ENs do Português, considerando as configurações de *features* descritas anteriormente. Para um melhor entendimento, os resultados do processo proposto são apresentados da seguinte forma:

Bases:

- ORG-ORG;
- ORG-PES;
- ORG-LOCAL;
- ORG-PES-LOCAL.

Para cada base:

- Medidas de desempenho: número de corretos (#C), Abrangência (A), Precisão (P) e F-measure (F).
- Taxa/nível de significância do valor alcançado para cada configuração de *features* em relação à configuração anterior por meio do teste de hipótese *T-test* [56] (os valores que apresentaram melhoria significativa foram indicados com * nas tabelas).
- Método de avaliação: validação cruzada em *r-folds* (número de partições) em que os exemplos são aleatoriamente divididos em *r folds* mutuamente exclusivas. Os exemplos nos *(r-1) folds* são utilizados para treinamento e a hipótese induzida é testada no *fold* remanescente. Este processo é repetido *r* vezes, cada vez considerando um *fold* diferente para teste. No final das iterações do processo de validação cruzada, consideramos neste trabalho a soma dos resultados dos *r folds* de teste. Nesse caso, a soma dos resultados das 5 bases de teste, uma vez que aplicamos validação cruzada com *5-folds*, devido ao tamanho reduzido das bases, ampliando assim a proporção de casos de teste. Entretanto, a configuração mais tradicional de *10-folds* também foi considerada e está disponível no Apêndice B.
- Matriz de confusão BIO: a matriz de confusão é apresentada considerando a soma dos resultados de cada *fold* de teste (validação cruzada em *r-folds*), ou seja, a matriz final é a soma das matrizes individuais. Portanto, cada matriz de confusão inclui todos os exemplos da base de teste representando assim cada um dos classificadores.

ORG-ORG:

Na classificação BIO da base ORG-ORG, ilustrada na Tabela 6.2, podemos notar que a configuração *F4* teve um impacto positivo no aprendizado, uma vez alcançou o melhor valor de Abrangência (45%).

Em relação às diferentes configurações aplicadas à base ORG-ORG (ver Tabela 6.3), podemos notar que a configuração *F5* apresentou ganhos significativos em Precisão em relação à configuração anterior (taxa de significância de 95%), alcançando taxas de Precisão de 53% e de 80% considerando descritores corretos e descritores parcialmente corretos, respectivamente.

ORG-ORG (*5-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	36	3	51	0.40	0.65	0.49
	I-REL	0	90	191	0.32	0.53	0.40
	O	19	74	1609	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	28	4	58	0.31	0.65	0.42
	I-REL	0	86	195	0.30	0.45	0.36
	O	15	100	1587	0.93	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	36	3	51	0.40	0.69	0.50
	I-REL	0	86	195	0.30	0.45	0.36
	O	16	100	1586	0.93	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	**41**	2	47	**0.45**	0.69	0.55
	I-REL	0	87	194	0.30	0.46	0.37
	O	18	97	1587	0.93	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	36	0	54	0.40	0.80	0.53
	I-REL	0	71	210	0.25	0.58	0.35
	O	9	50	1643	0.96	0.86	0.91
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	38	0	52	0.42	0.79	0.55
	I-REL	0	74	207	0.26	0.58	0.36
	O	10	53	1639	0.96	0.86	0.91

Tabela 6.2 – Classificação BIO de ORG-ORG por conjunto de *features*.

Por fim, a configuração *F6* apresentou os melhores resultados para descritores corretos em relação às demais configurações. Já para os descritores parcialmente corretos, a configuração *F6* manteve a taxa de Precisão de 80%, e apresentou um aumento da F-measure em relação à configuração anterior (de 53% para 55%).

ORG-ORG	Descritores Corretos				Descritores Parcialmente Corretos			
(5-folds)	#C	A	P	F	#C	A	P	F
F1=POS	23	0.25	0.41	0.31	36	0.40	0.65	0.49
F2=POS+LEX	19	0.21	0.44	0.28	28	0.31	0.65	0.42
F3=POS+LEX+SINT	24	0.26	0.46	0.33	36	0.40	0.69	0.50
F4=POS+LEX+SINT+PAD	24	0.26	0.40	0.32	**41**	**0.45**	0.69	0.54
F5=POS+LEX+SINT+PAD+FR	24	0.26	0.53*	0.35	36	0.40	**0.80***	0.53
F6=POS+LEX+SINT+PAD+FR+SEM	**27**	**0.30**	**0.56**	**0.39**	38	0.42	**0.79**	**0.55**

Tabela 6.3 – Resultados de ORG-ORG por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

ORG-PES:

Na Tabela 6.4, notam-se boas taxas de Precisão para todas as etiquetas da classificação BIO para a base ORG-PES, refletindo a baixa taxa de falsos-positivos. Destacam-se os 56 casos etiquetados com B-REL aplicando-se a configuração *F6* e, consequentemente, uma alta taxa de Precisão (80%).

ORG-PES (*5-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	51	6	48	0.48	0.63	0.55
	I-REL	4	127	181	0.40	0.54	0.46
	O	25	99	1425	0.91	0.86	0.88
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	44	8	53	0.41	0.69	0.52
	I-REL	3	135	174	0.43	0.53	0.47
	O	16	110	1423	0.91	0.86	0.88
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	56	5	44	0.53	0.77	0.63
	I-REL	1	147	164	0.47	0.56	0.51
	O	15	110	1424	0.91	0.87	0.89
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	56	5	44	0.53	0.76	0.62
	I-REL	1	150	161	0.48	0.56	0.51
	O	16	112	1421	0.91	0.87	0.89
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	55	3	47	0.52	0.78	0.62
	I-REL	2	120	190	0.38	0.62	0.47
	O	13	70	1466	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	**56**	2	47	0.53	**0.80**	0.64
	I-REL	2	127	183	0.40	0.63	0.49
	O	12	70	1467	0.94	0.86	0.90

Tabela 6.4 – Classificação BIO de ORG-PES por conjunto de *features*.

ORG-PES	Descritores Corretos				Descritores Parcialmente Corretos			
(5-folds)	#C	A	P	F	#C	A	P	F
F1=POS	33	0.31	0.41	0.35	51	0.48	0.63	0.55
F2=POS+LEX	37	0.35	0.58*	0.44	44	0.41	0.69	0.52
F3=POS+LEX+SINT	47	0.44	0.65	0.53	**56**	**0.53***	0.77	**0.63**
F4=POS+LEX+SINT+PAD	45	0.42	0.61	0.50	**56**	**0.53**	0.76	0.62
F5=POS+LEX+SINT+PAD+FR	45	0.42	0.64	0.51	55	0.52	0.78	0.62
F6=POS+LEX+SINT+PAD+FR+SEM	**50**	**0.47**	**0.71**	**0.57**	**56**	**0.53**	**0.80**	**0.63**

Tabela 6.5 – Resultados de ORG-PES por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

Nos resultados alcançados para a base ORG-PES por conjunto de *features*, tivemos um ganho significativo em Precisão na configuração *F2* para descritores corretos em relação à configuração anterior (taxa de significância de 99%), conforme ilustrado na Tabela 6.5. A configuração *F3* apresentou ganhos em Abrangência para descritores parciamente corretos em relação à configuração anterior (taxa de significância de 95%), e a melhor taxa de F-measure (63%). Destaca-se a configuração *F6* por alcançar as melhores taxas de desempenho.

ORG-LOCAL:

ORG-LOCAL (*5-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	40	1	68	0.36	0.66	0.47
	I-REL	0	67	225	0.22	0.47	0.30
	O	20	74	1671	0.94	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	**47**	1	61	**0.43**	0.73	0.54
	I-REL	1	59	232	0.20	0.41	0.27
	O	16	81	1668	0.94	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	46	4	59	0.42	0.71	0.53
	I-REL	2	73	217	0.25	0.41	0.31
	O	16	97	1652	0.93	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	44	3	62	0.40	0.68	0.50
	I-REL	2	72	218	0.24	0.40	0.30
	O	18	102	1645	0.93	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	43	2	64	0.39	0.72	0.51
	I-REL	1	63	228	0.21	0.53	0.30
	O	15	53	1697	0.96	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	45	3	61	0.41	0.80	0.54
	I-REL	2	73	217	0.25	0.55	0.34
	O	9	55	1701	0.96	0.85	0.90

Tabela 6.6 – Classificação BIO de ORG-LOCAL por conjunto de *features*.

ORG-LOCAL	Descritores Corretos				Descritores Parcialmente Corretos			
(5-folds)	#C	A	P	F	#C	A	P	F
F1=POS	30	0.27	0.50	0.35	40	0.36	0.66	0.47
F2=POS+LEX	41	0.37	0.64*	0.47*	**47**	**0.43**	0.73	**0.54**
F3=POS+LEX+SINT	39	0.35	0.60	0.45	46	0.42	0.71	0.53
F4=POS+LEX+SINT+PAD	37	0.33	0.57	0.42	44	0.40	0.68	0.50
F5=POS+LEX+SINT+PAD+FR	39	0.35	0.66	0.46	43	0.39	0.72	0.51
F6=POS+LEX+SINT+PAD+FR+SEM	**43**	**0.39**	**0.76***	**0.52**	45	0.41	**0.80**	**0.54**

Tabela 6.7 – Resultados de ORG-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

Na classificação BIO resultante para a base ORG-LOCAL, destaca-se a configuração *F2*, a qual classificou 47 exemplos com a etiqueta B-REL, alcançando a melhor taxa de abrangência (43%), conforme ilustrado na Tabela 6.6. Os resultados da base ORG-LOCAL (Tabela 6.7) apresentaram ganhos significativos para os descritores corretos: a configuração *F2*, comparada à configuração anterior alcançou ganhos em Precisão e em F-measure (grau de significância de 95% e 97.5%, respectivamente). Destacou-se também a configuração *F6* com ganho em Precisão comparado à configuração anterior (grau de significância de 95%). Essa configuração de *feature* apresentou as melhores taxas de Precisão e de F-measure para os descritores corretos e parcialmente corretos.

ORG-PES-LOCAL:

ORG-PES-LOCAL (*5-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	120	11	173	0.39	0.64	0.48
	I-REL	3	213	669	0.24	0.47	0.31
	O	63	229	4720	0.94	0.84	0.89
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	129	11	164	0.42	0.73	0.53
	I-REL	0	307	578	0.34	0.56	0.42
	O	46	225	4741	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	**132**	9	163	**0.43**	0.71	0.53
	I-REL	1	347	537	0.39	0.56	0.46
	O	52	262	4698	0.93	0.87	0.90
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	**132**	8	164	**0.43**	0.72	0.54
	I-REL	1	337	547	0.38	0.55	0.45
	O	48	259	4705	0.93	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	117	4	183	0.38	0.72	0.50
	I-REL	0	265	620	0.29	0.64	0.40
	O	45	144	4823	0.96	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	125	5	174	0.41	0.75	0.53
	I-REL	1	271	613	0.30	0.65	0.41
	O	39	140	4833	0.96	0.85	0.90

Tabela 6.8 – Classificação BIO de ORG-PES-LOCAL por conjunto de *features*.

ORG-PES-LOCAL	Descritores Corretos				Descritores Parcialmente Corretos			
(5-folds)	#C	A	P	F	#C	A	P	F
F1=POS	71	0.23	0.38	0.28	120	0.39	0.64	0.48
F2=POS+LEX	101	0.33*	0.57*	0.42*	129	0.41	0.69	0.52
F3=POS+LEX+SINT	105	**0.34**	0.56	0.42	**132**	**0.43**	0.71	0.53
F4=POS+LEX+SINT+PAD	104	**0.34**	0.57	0.42	**132**	**0.43**	0.72	**0.54**
F5=POS+LEX+SINT+PAD+FR	101	0.33	0.62	0.43	117	0.38	0.72	0.49
F6=POS+LEX+SINT+PAD+FR+SEM	**106**	**0.34**	**0.64**	**0.45**	125	0.41	**0.75**	0.53

Tabela 6.9 – Resultados de ORG-PES-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

A base ORG-PES-LOCAL, diferentemente das demais bases, alcançou um melhor desempenho com a validação cruzada de *10-folds* em comparação à aplicação de *5-folds*. Esse comportamento era esperado, já que essa base é constituída pela união dos exemplos das três bases; logo, a validação cruzada com *10-folds* é mais apropriada. Por esse motivo, além dos resultados da base ORG-PES-LOCAL com validação cruzada de *5-folds*, apresentamos também os resultados de *10-folds*.

A classificação BIO referente à base ORG-PES-LOCAL, tanto com validação cruzada de *5-folds* como com *10-folds*, apresentou o maior número de casos etiquetados com B-REL com as configurações *F3* e *F4*, conforme ilustrado na Tabela 6.8 e Tabela 6.10, respectivamente.

Consequentemente, as configurações *F3* e *F4* alcançaram as melhores taxas de abrangência, 43% e 46%, com a aplicação da validação cruzada com *5-folds* e com *10-folds*, respectivamente.

Os resultados da base ORG-PES-LOCAL com validação cruzada de *5-folds* são apresentados na Tabela 6.9. Ganhos com a configuração *F2* para os descritores corretos foram alcançados para todas as taxas em relação à configuração anterior, com grau de significância de 99.5%.

Em geral, para os descritores corretos, a configuração *F6* apresentou as melhores taxas de desempenho. Do total de 106 casos de descritores corretos, 20 exemplos correspondem à base ORG-ORG, 40 exemplos são da base ORG-PES, e 46 exemplos são da ORG-LOCAL. Para os descritores parcialmente corretos, a melhor taxa de F-measure (54%) ocorreu com a configuração *F4*, a taxa de 75% de Precisão foi alcançada pela configuração *F6*.

ORG-PES-LOCAL (*10-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	121	9	174	0.39	0.63	0.48
	I-REL	3	203	679	0.22	0.49	0.31
	O	200	68	4744	0.94	0.84	0.89
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	133	7	164	0.43	0.74	0.55
	I-REL	2	318	565	0.35	0.56	0.44
	O	44	235	4733	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	**140**	5	159	**0.46**	0.71	0.55
	I-REL	1	334	550	0.37	0.56	0.45
	O	56	250	4706	0.93	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	**141**	6	157	**0.46**	0.71	0.56
	I-REL	1	342	542	0.38	0.56	0.45
	O	56	256	4700	0.93	0.87	0.90
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	122	6	176	0.40	0.72	0.51
	I-REL	2	276	607	0.31	0.63	0.41
	O	45	154	4813	0.96	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	133	5	166	0.43	0.74	0.55
	I-REL	3	287	595	0.32	0.64	0.43
	O	43	152	4817	0.96	0.86	0.90

Tabela 6.10 – Classificação BIO de ORG-PES-LOCAL por conjunto de *features*.

Os resultados da base ORG-PES-LOCAL com validação cruzada de *10-folds*, de uma maneira geral, alcançaram taxas de desempenho mais altas do que as apresentadas para essa base considerando *5-folds*. Podemos notar, na Tabela 6.11, ganhos com a configuração *F2*: para os descritores corretos temos ganhos nas taxas de Abrangência, Precisão e F-measure em comparação à configuração anterior (grau de significância de 99.5%), e para descritores parcialmente corretos temos a melhor taxa de Precisão (74%) com grau de significância de 99.5% em relação à configuração anterior.

ORG-PES-LOCAL	Descritores Corretos				Descritores Parcialmente Corretos			
(10-folds)	#C	A	P	F	#C	A	P	F
F1=POS	78	0.25	0.40	0.31	121	0.39	0.63	0.48
F2=POS+LEX	107	0.35*	0.59*	0.44*	133	0.43	**0.74***	0.54
F3=POS+LEX+SINT	108	0.35	0.54	0.43	140	**0.46**	0.71	0.55
F4=POS+LEX+SINT+PAD	108	0.35	0.54	0.43	**141**	**0.46**	0.71	**0.56**
F5=POS+LEX+SINT+PAD+FR	103	0.33	0.60	0.43	122	0.40	0.72	0.51
F6=POS+LEX+SINT+PAD+FR+SEM	**113**	**0.37**	**0.63**	**0.46**	133	0.44	**0.74**	0.55

Tabela 6.11 – Resultados de ORG-PES-LOCAL por conjunto de *Features*. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

Nesta base, a configuração *F6* para os descritores corretos manteve-se com as melhores taxas de desempenho, e com valores superiores de Abrangência e F-measure aos alcançados com validação de *5-folds*. Destaca-se que os 113 exemplos identificados como descritores corretos estão distribuídos da seguinte forma nas bases: 24 deles são da base ORG-ORG, 43 exemplos são da base ORG-PES, e 46 exemplos da base ORG-LOCAL foram extraídos corretamente. Para os descritores parcialmente corretos, manteve-se uma melhor taxa de F-measure (56%) com a configuração *F4*.

6.3 Discussão dos Resultados

Nesta seção uma discussão sobre os resultados do processo proposto de extração de descrição de relação entre ENs é apresentada. Destaca-se dos resultados apresentados na Seção anterior, o número de descritores de relação classificados corretamente para cada base, considerando a sua classificação como verbais e não verbais (ver Seção 4.2).

O gráfico da Figura 6.1 apresenta, para cada base, o número de descritores de relação verbais e não verbais extraídos corretamente com a melhor configuração de *features* (configuração *F6*), considerando os dois critérios de avaliação (descritores corretos e descritores parcialmente corretos), e o número total correspondente de descritores de referência.

Podemos notar que, para a base ORG-ORG (validação cruzada de *5-folds*) foram etiquetados um maior número de descritores de relação parcialmente corretos em comparação aos corretos (16 e 26, respectivamente, de um total de 66 exemplos), principalmente para as relações verbais. Isso se deve ao fato de os descritores de relação desta base, na sua maioria, serem formados por várias palavras, assim dificultando a etiquetagem de todos os elementos que formam tais descritores. Em comparação às demais bases, os descritores da base ORG-ORG são bem mais extensos, ou seja, formados por um número maior de palavras.

A base ORG-PES (validação cruzada de *5-folds*) foi a que apresentou o maior número de descritores de relação etiquetados corretamente. Das 60 relações não-verbais de referência, temos 40 exemplos etiquetados corretamente e 42 exemplos identificados parcialmente corretos. Das 45 relações verbais de referência, temos 10 exemplos corretos e 12 exemplos parcialmente corretos.

Cabe enfatizar que para a base ORG-LOCAL, temos um maior número de relações de referência não-verbais do que verbais (72 e 37, respectivamente) devido às características das relações contidas nesta base, como as relações de “localização" e “pertence-a".

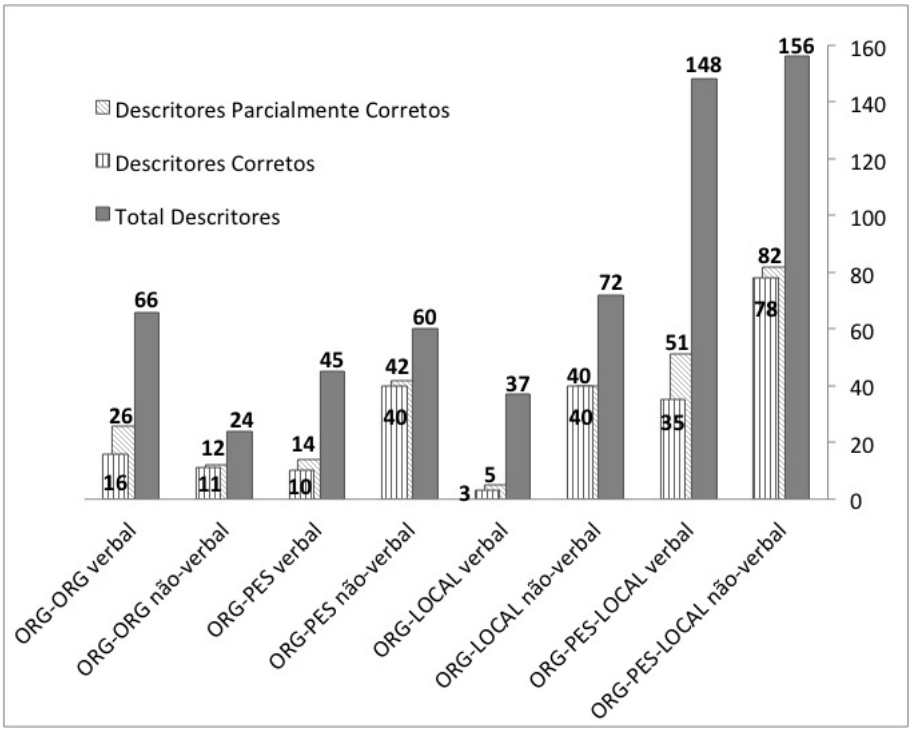

Figura 6.1 – Melhores resultados de descritores de relação.

Tais aspectos foram observados no gráfico 6.1, no qual, para a base ORG-LOCAL (validação cruzada de *5-folds*), foram etiquetados apenas 3 e 5 casos de descritores de relação verbal corretos e parcialmente corretos, respectivamente. Já para os descritores de relação não verbais tivemos 40 casos corretos de um total de 72 exemplos.

Por fim, a base ORG-PES-LOCAL apresentou o melhor número de exemplos corretos com a aplicação da validação cruzada de *10-folds* (ver Figura 6.1). Para descritores de relação verbais, de 148 exemplos de referência, alcançou-se um total de 35 exemplos corretos (distribuídos nas bases: ORG-ORG: 14 exemplos; ORG-PES: 11 exemplos; e ORG-LOCAL: 10 exemplos) e 51 exemplos parcialmente corretos (distribuídos nas bases: ORG-ORG: 19 exemplos; ORG-PES: 18 exemplos; e ORG-LOCAL: 14 exemplos). Para descritores de relação não verbais, dos 156 exemplos de referência classificou-se 82 exemplos como parcialmente corretos distribuídos nas bases: 11 casos da ORG-ORG; 34 casos da ORG-PES; e 37 casos da ORG-LOCAL.

Na avaliação dos resultados dos diferentes modelos CRFs gerados com base nas configurações de *features* propostas, temos para a base ORG-ORG a configuração *F4* em destaque, a qual alcançou 54% de F-measure, conforme gráfico da Figura 6.2. Isso ocorre em razão de a maioria dos descritores de relação dessa base serem extensos e constituídos por verbos. Assim, as *features* baseadas em padrões auxiliam na identificação das palavras que formam esses descritores. A configuração *F6* se manteve apresentando as melhores taxas de F-measure para descritores corretos e parcialmente corretos (39% e 55%, respectivamente). Isso se deve ao fato de a *feature* semântica baseada na categoria da EN prover uma informação valiosa sobre o tipo de descritor que se deseja identificar.

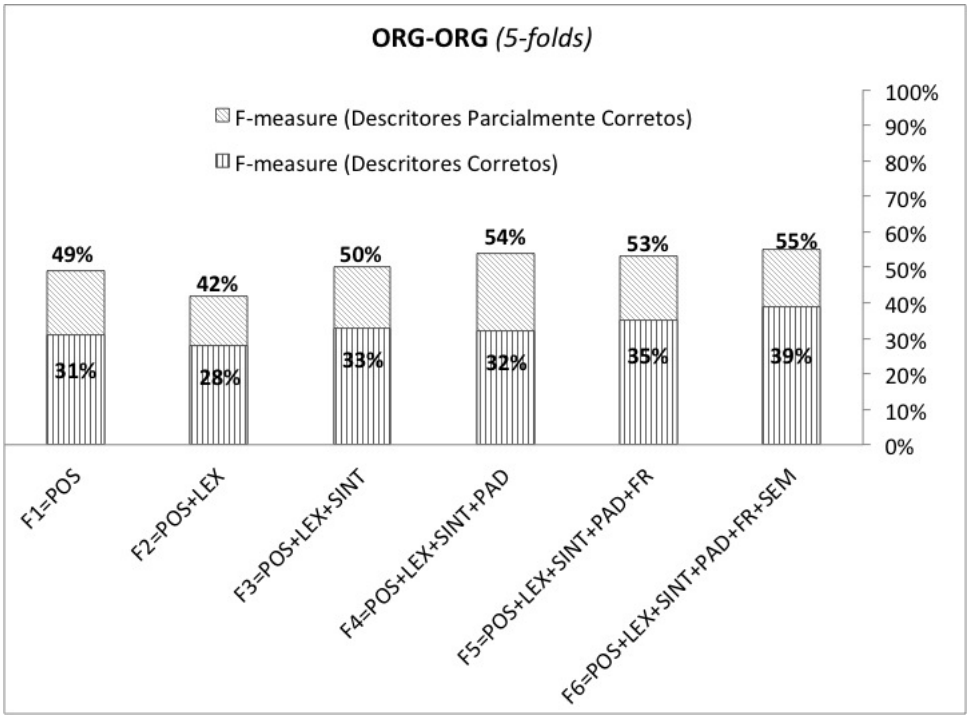

Figura 6.2 – Comparativo de F-measure da base ORG-ORG entre as diferentes configurações de *Features* com validação cruzada de *5-folds*.

No gráfico da Figura 6.3 podemos notar que a base ORG-PES apresentou boas taxas de F-measure, destacando-se a taxa de F-measure resultante da aplicação da configuração *F3*. Essa configuração apresentou a mesma taxa da configuração *F6* para descritores parcialmente corretos. Isso ocorre porque as *features* sintáticas auxiliam na identificação dos descritores de relação em foco, em especial as *features* baseadas no aposto, no objeto direto e no núcleo, as quais ocorrem em vários exemplos de descritores de relação da base ORG-PES. A configuração *F6* apresentou também a melhor taxa de F-measure para descritores corretos, uma vez que a *feature* baseada na anotação semântica de cargo/profissão auxilia na identificação das relações de "vínculo institucional".

A base ORG-LOCAL apresentou a melhor taxa de F-measure (54%) considerando a configuração *F2*, mesmo valor alcançado pela configuração *F6*, de acordo com o gráfico da Figura 6.4. A partir da análise desta base, verificou-se que a adição das *features* baseadas nos itens lexicais já auxiliam na identificação dos descritores de relação, uma vez que tais descritores, na sua maioria, descrevem relações não verbais de "localização" e "pertence-a" expressas geralmente por uma preposição.

Assim, uma simples *feature* que expresse esse padrão já traz ganhos na etiquetagem dos exemplos. Destaca-se também a configuração *F6* (ver Figura 6.4), a qual alcançou uma boa taxa de F-measure para descritores de relação corretos (52%).

Conforme apresentado na Seção 6.2, a base ORG-PES-LOCAL apresentou melhores taxas de F-measure com a validação cruzada de *10-folds*, diferentemente das demais bases. Isso se deve ao fato dessa base ser constituída por mais exemplos para treino-teste, já que é formada pela união das outras 3 bases, e assim adequando-se mais para a técnica de validação cruzada com mais *folds*.

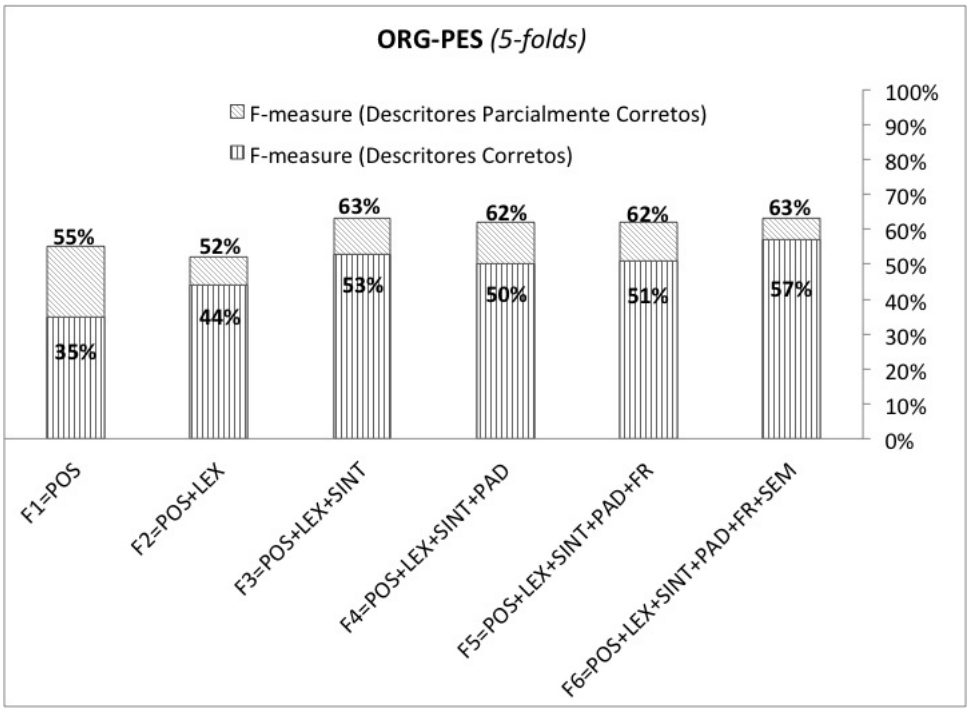

Figura 6.3 – Comparativo de F-measure da base ORG-PES entre as diferentes configurações de *Features* com validação cruzada de *5-folds*.

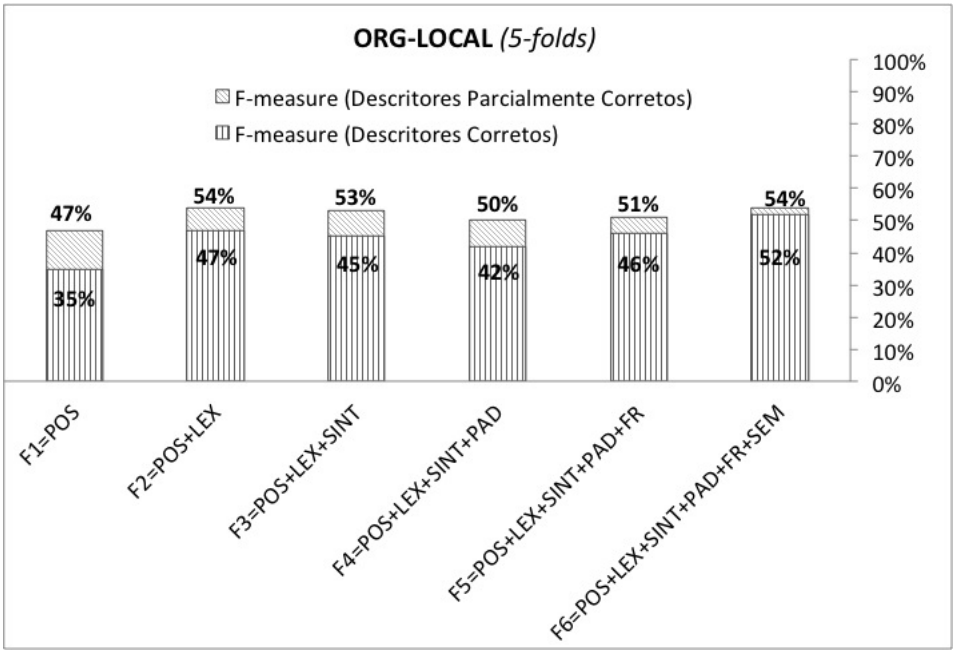

Figura 6.4 – Comparativo de F-measure da base ORG-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *5-folds*.

Dentre as configurações de *features* para a base ORG-PES-LOCAL, destaca-se a configuração *F4*, com F-measure de 54% com validação cruzada de *5-folds*, e uma taxa mais alta (56%), com *10-folds* (Figura 6.5 e Figura 6.6, respectivamente). Já para os descritores corretos, a configuração *F6* alcançou a melhor F-measure (45% com *5-folds* - Figura 6.5, 46% com *10-folds* - Figura 6.6).

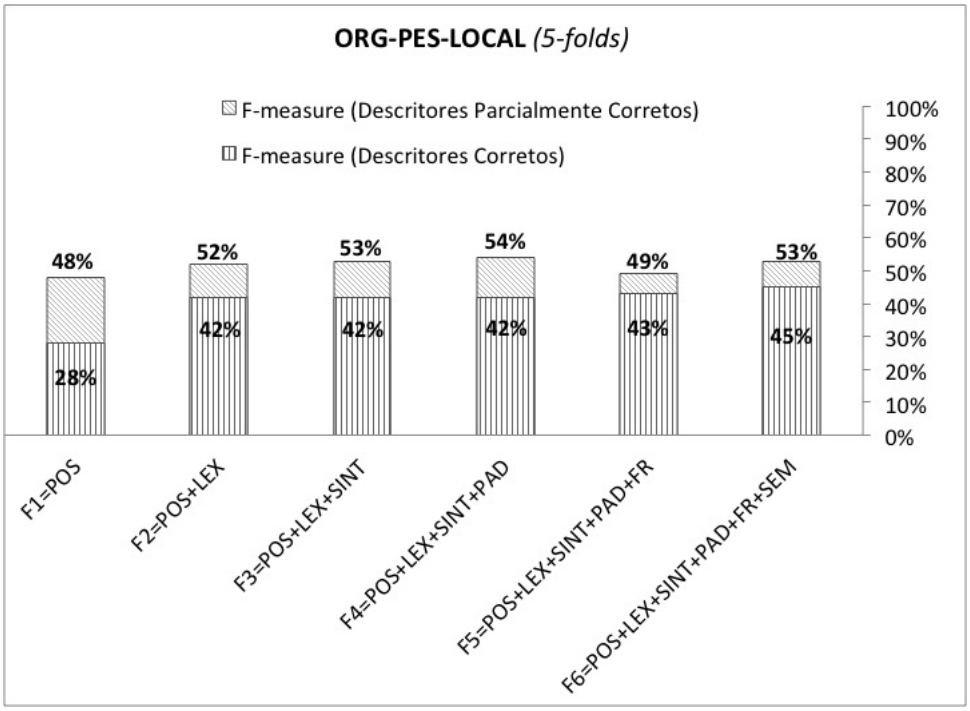

Figura 6.5 – Comparativo de F-measure da base ORG-PES-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *5-folds*.

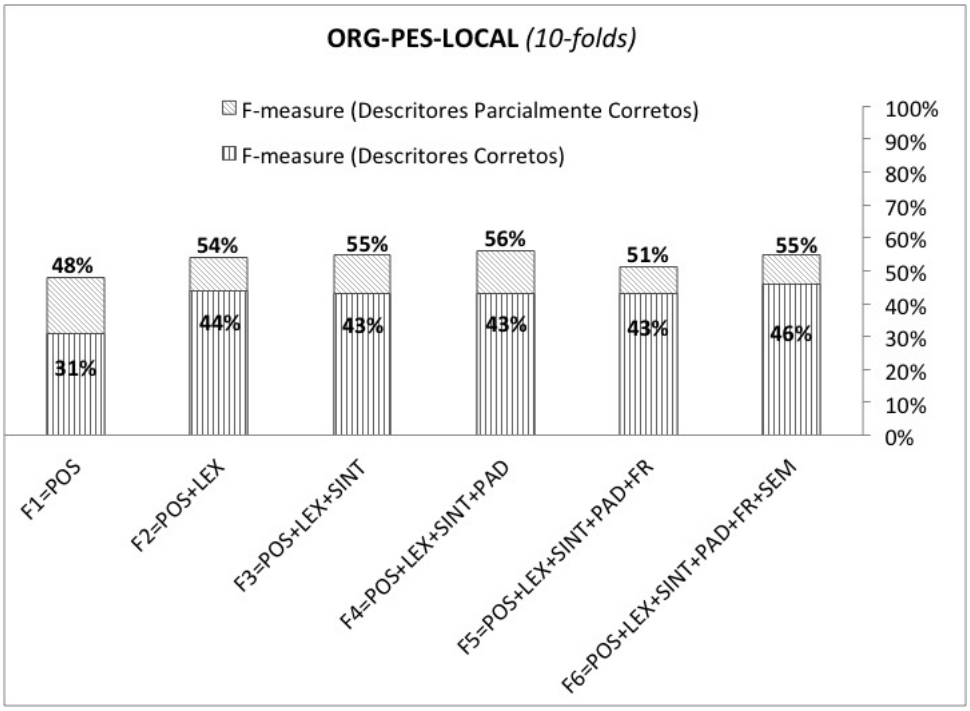

Figura 6.6 – Comparativo de F-measure da base ORG-PES-LOCAL entre as diferentes configurações de *Features* com validação cruzada de *10-folds*.

6.4 Análise de Erros

A partir da avaliação dos resultados apresentados na Seção 6.2, alguns erros na classificação BIO dos descritores de relação foram identificados e são discutidos a seguir:

Exemplos Falso-Positivos:

Na avaliação dos resultados, percebeu-se a identificação de alguns exemplos falso-positivos, ou seja, descritores de relação formados por palavras que não expressam uma relação entre as ENs. Assim, tais palavras não deveriam ter sido etiquetadas/classificadas com a etiqueta B-REL, seguida ou não por etiquetas I-REL, e sim com a etiqueta O. Cabe salientar que, em geral, as bases de dados apresentaram poucos casos de falso-positivos, e isso ocorre em razão de o modelo probabilístico CRF utilizado no processo proposto ser muito preciso na etiquetagem dos descritores de relação.

Um dos principais erros na classificação de exemplos como falso-positivos foi a identificação de descritores de relação constituídos por verbos, e que não expressam uma relação explícita entre pares de ENs. Dentre as bases, a ORG-ORG foi a base em que ocorreram mais casos desse tipo de erro.

A Tabela 6.12 ilustra exemplos de casos falso-positivos da base ORG-ORG, em que as ENs estão destacadas em itálico, seguida da coluna que indica as etiquetas de saída do processo proposto, e da coluna que indica as etiquetas de referência. Nos dois exemplos apresentados nessa tabela, as palavras que ocorrem entre as ENs deveriam ter recebido a etiqueta O por não expressarem uma relação explícita no texto. No primeiro exemplo, interpretou-se a EN "Associação Industrial da Região de Viseu" como referente do verbo "ser", quando, na verdade, o seu referente é a EN anterior *"Almeida Henriques"* ("Almeida Henriques, presidente da Associação Industrial da Região de Viseu (AIRV), é o novo rosto do Conselho Empresarial do Centro."). O mesmo ocorre no segundo exemplo, em que o verbo "passou" é interpretado de forma incorreta como referente da EN "Marinha", seguido do verbo "integrar". Nesse caso o referente é "BCM", que está no enunciado anterior ("BCM possui actualmente cerca de 80.000 volumes correspondendo a perto de 45.000 títulos. A partir de 1994, em consequência da reestruturação orgânica operada na *Marinha*, passou a integrar o *Arquivo Central da Marinha*").

Instâncias de Relação Negativas	Etiquetas BIO de Saída	Etiquetas BIO de Referência
Almeida Henriques, presidente da *Associação Industrial da Região de Viseu* (AIRV), é o novo rosto do *Conselho Empresarial do Centro*.	ser<B-REL>	ser<O>
A partir de 1994, em consequência da reestruturação orgânica operada na *Marinha*, passou a integrar o *Arquivo Central da Marinha*	integrar<B-REL> o<I-REL>	integrar<O> o<O>

Tabela 6.12 – Exemplos classificados como falso-positivos na base ORG-ORG.

A base ORG-PES apresentou descritores de relação não verbais classificados como falso-positivos. Isso ocorre porque há um maior número de relações não verbais nessa base. Alguns desses casos classificados como falso-positivos são exemplos formados por nomes de cargos/profissões. Conforme apresentado na Seção 4.2, temos na base ORG-PES 25 exemplos de descritores de relação não verbais compostos por nomes de cargos/profissões, os quais representam a relação de "vínculo institucional" entre Pessoa e Organização. Além disso, temos a *feature* de *Dicionário* (lista de cargos/profissões) e a *feature* semântica (anotação semântica de cargo/profissão), as quais adicionam mais relevância às informações de profissão como candidatas a constituintes de um descritor de relação.

Instâncias de Relação Negativas	**Etiquetas BIO de Saída**	**Etiquetas BIO de Referência**
O escritor *Clive Cussler*, autor das aventuras de Dirk Pitt, assinou um contrato com a *Simon & Schuste* ...	autor<B-REL>	autor<O>
... *Saraiva Dias*, vereador substituto do presidente da autarquia, referiu . ao *PÚBLICO* ...	vereador<B-REL>	vereador<O>

Tabela 6.13 – Exemplos classificados como falso-positivos na base ORG-PES.

Para exemplificar tais casos, na Tabela 6.13 são apresentados dois exemplos classificados como descritores falso-positivos na base ORG-PES. Em ambos os casos, a informação referente à profissão ("autor" e "vereador", respectivamente) ocorreu em construções de aposto, sendo interpretadas como informações mais relevantes, e que expressam uma relação entre os respectivos pares de ENs. De fato, estabeleceu-se uma relação entre a EN ("*Clive Cussler*") e sua profissão ("autor"), já que, normalmente, na base ORG-PES a relação descrita é não-verbal. Contudo, o correto, neste caso, é uma relação idiossincrática para o sistema: no primeiro caso, o verbo "assinar" estabeleceria a relação entre as ENs: *"Clive Cussler"* e *"Simon & Schuste"* ; no segundo caso, o verbo "referir" deveria relacionar *"Saraiva Dias"* e *"PÚBLICO"*.

Na base ORG-LOCAL, como temos vários casos de descritores de relação positivos que expressam a relação de "localização" por meio de uma preposição (relações não-verbais), ocorreu a classificação de descritores falso-positivos formados por preposição. Na Tabela 6.14, são ilustrados alguns desses casos de falso-positivos.

Instâncias de Relação Negativas	**Etiquetas BIO de Saída**	**Etiquetas BIO de Referência**
... embaixador de *Portugal* em *Espanha* ...	em<B-REL>	em<O>
... *Inglaterra* de que nenhum governo imposto por Napoleão em *Portugal* ...	em<B-REL>	em<O>

Tabela 6.14 – Exemplos classificados como falso-positivos na base ORG-LOCAL.

A base ORG-PES-LOCAL, em geral, apresentou uma distribuição uniforme de casos falso-positivos entre cada par de ENs considerado (ORG-ORG, ORG-PES e ORG-LOCAL).

A Tabela 6.15 ilustra exemplos de descritores falso-positivos dessa base. No primeiro exemplo, temos um descritor de relação falso-positivo envolvendo as ENs das categorias Local e Organização, em que as palavras que o formam deveriam ter recebido a etiqueta O. Isso se deve ao fato de a preposição "por" ter função de adjunto adverbial de local, assim sendo relacionada à EN *"Setúbal"*. O segundo exemplo ilustra um descritor de relação classificado como falso-negativo, o qual ocorreu entre duas ENs da categoria Organização. Nesse caso, o verbo "anunciar" foi interpretado de forma errada como referente da EN "Iraque".

Instâncias de Relação Negativas	**Etiquetas BIO de Saída**	**Etiquetas BIO de Referência**
... logotipo do porto de *Setúbal*, por decisão do *Conselho de Administração* ...	por<B-REL> decisão<I-REL> de <I-REL> o<I-REL>	por<O> decisão<O> de<O> o<O>
... disparos da artilharia anti-aérea do *Iraque*, anunciou o *Pentágono* ...	anunciar<B-REL> o<I-REL>	anunciar<O> o <O>

Tabela 6.15 – Exemplos classificados como falso-positivos na base ORG-PES-LOCAL.

Exemplos Falso-Negativos:

Na avaliação dos resultados, realizou-se uma análise dos exemplos falso-negativos, ou seja, descritores de relação que expressam uma relação entre as ENs e que não foram etiquetados nas bases de dados (as palavras que formam tais descritores receberam a etiqueta O).

Para a base ORG-ORG a maioria dos exemplos falso-negativos foi de descritores de relações verbais, haja vista que tais descritores são extensos, o que dificulta a sua etiquetagem. Para exemplificar, a Tabela 6.16 apresenta um descritor de relação verbal classificado como falso-negativo, o qual é composto por uma oração relativa[3]. Já o segundo exemplo ilustra um caso não verbal etiquetado de forma incorreta. Nesse caso, a palavra "armada" foi anotada pelo *parser* PALAVRAS como verbo, em vez de adjetivo, e consequentemente impactou na etiquetagem dos elementos que formam o descritor não verbal: "ala armada do" (a palavra "ala" recebeu a etiqueta O, seguida de etiquetas I-REL).

A base ORG-PES apresentou mais exemplos falso-negativos verbais do que não verbais. Destaca-se que, dos 25 descritores de relação não verbais que expressam a relação de "vínculo-institucional", foram etiquetados corretamente 17 desses casos (apenas 8 falso-negativos dessa relação não verbal).

Na Tabela 6.17 ilustramos no primeiro exemplo um caso da relação de "vínculo-institucional", classificado como falso negativo: "líder do".

[3]As orações relativas são iniciadas por um pronome relativo que concorda em gênero/número com o seu antecedente, e exerce função de modificador desse antecedente.

Instâncias de Relação Positivas	Etiquetas BIO de Saída	Etiquetas BIO de Referência
... da *Biblioteca Houghton*, **que guarda as obras raras de** *Harvard* ...	que<O> guarda<O> as<O> obras<O> raras<O> de<O>	que<B-REL> guarda<I-REL> as<I-REL> obras<I-REL> raras<I-REL> de<I-REL>
A *Resistência Islâmica*, **ala armada do** *Hizbollah* ...	ala<O> armar<I-REL> de<I-REL> o<I-REL> Hizbollah<I-REL>	ala<B-REL> armar<I-REL> de<I-REL> o<I-REL> Hizbollah<I-REL>

Tabela 6.16 – Exemplos classificados como falso-negativos na base ORG-ORG.

Nesse caso, o descritor não foi etiquetado corretamente: a profissão “líder" recebeu a etiqueta O, em vez de B-REL. Isso é consequência da construção de aposto “(PSE,DE)", que ocorre entre a EN “Martin Schulz" e o descritor de relação em foco, não ter sido corretamente anotada pelo parser PALAVRAS. Na sequência, exemplificamos na Tabela 6.17 um caso de falso-negativo verbal, em que o descritor: “dirigido pelo" não foi identificado; logo, os elementos que o formam receberam a etiqueta O. Isso se deve à anotação do parser PALAVRAS, que anotou o verbo “dirigir" como adjetivo. Nota-se que a forma do particípio desse verbo (“dirigido") é a mesma que a do adjetivo, o que pode ter causado o erro.

Instâncias de Relação Positivas	Etiquetas BIO de Saída	Etiquetas BIO de Referência
... *Martin SCHULZ* (PSE, DE), **líder do** *PSE* ...	líder<O> de<I-REL> o<I-REL>	líder<B-REL> de<I-REL> o<I-REL>
... *Escola de Pilotagem* ajudou a formar dezenas de pilotos particulares e profissionais, sendo actualmente **dirigida pelo** *Cmdt. João Filhó*	dirigido<O> por<O> o<O>	dirigido<B-REL> por<I-REL> o<I-REL>

Tabela 6.17 – Exemplos classificados como falso-negativos na base ORG-PES.

Na Tabela 6.18, temos, primeiramente, um exemplo falso-negativo de descritor de relação verbal da base ORG-LOCAL. Nesse exemplo, ocorreram elementos interpostos entre a EN “Legião da Boa Vontade" e o descritor da relação (“foi fundada no"), o que dificultou a sua identificação. O segundo caso falso negativo é de um descritor de relação não verbal, o qual expressa uma relação de “destino" entre o par de ENs. Entretanto, ocorreu um erro na anotação provida pelo parser PALAVRAS, em que “Setembro" foi anotado como uma localidade geográfica (etiqueta <top>). Consequentemente, temos dois adjuntos adverbiais de local, dificultando assim a anotação da preposição “em" com a etiqueta B-REL.

Instâncias de Relação Positivas	Etiquetas BIO de Saída	Etiquetas BIO de Referência
A *Legião da Boa Vontade*, instituição educacional, cultural, beneficente e filantrópica, **foi fundada no** *Brasil* ...	ser<O> fundar<O> em<O> o<O>	ser<B-REL> fundar<I-REL> em<I-REL> o<I-REL>
O *Conselho Europeu Metodista* reuniu no passado mês de Setembro **em** *Stuttgart* ...	em<O>	em<B-REL>

Tabela 6.18 – Exemplos classificados como falso-negativos na base ORG-LOCAL.

Instâncias de Relação Positivas	Etiquetas BIO de Saída	Etiquetas BIO de Referência
... o *PSD* **aproximou-se politicamente do** *PS* ...	aproximar<O> se<O> político<O> de<O> o<O>	aproximar<B-REL> se<I-REL> político<I-REL> de<I-REL> o<I-REL>
Dr. José Getúlio Lima (advogado e professor, pai do Dr. José Carlos de Barros Lima), que atuou muito na área educacional da região tendo sido **o fundador**, do ginásio *Conselheiro Lafayette*	o<O> fundador<O>	o<B-REL> fundador<I-REL>
A administração, vendas e a fábrica da *MARFINITE* **ficam em** *Itaquaquecetuba* ...	ficar<O> em<B-REL>	ficar <B-REL> em<I-REL>

Tabela 6.19 – Exemplos classificados como falso-negativos na base ORG-PES-LOCAL.

Para a base ORG-PES-LOCAL, por ser a união das demais bases, a distribuição dos exemplos falso-negativos se manteve similar aos apresentados anteriormente para cada base. A Tabela 6.19 apresenta três desses casos da base ORG-PES-LOCAL. No primeiro, o descritor de relação é composto pelo pronome reflexivo "se" que acompanha o verbo, seguido de um advérbio e de uma preposição. Esse padrão de descritor ocorreu entre duas ENs da categoria Organização, e apenas essa ocorrência foi encontrada na base, o que dificulta a sua identificação. No segundo, ocorreram elementos interpostos entre a EN "Dr. José Getúlio Lima" e o descritor ("o fundador"), o que dificultou a sua identificação. Além disso, a profissão "fundador" exerce a função de complemento do sujeito, nesse caso, a EN "Dr. José Getúlio Lima"; porém, essa EN não foi anotada como sujeito da oração pelo parser PALAVRAS, impactando também na identificação.

No terceiro exemplo da Tabela 6.19, ocorreu um erro na etiquetagem BIO, em que o verbo "ficar" não recebeu a etiqueta B-REL, mas somente a preposição "em", a qual deveria ter sido etiquetada com I-REL. Isso ocorre em razão de o verbo "ficar" referir-se aos núcleos "administração", "vendas" e fábrica", e não à EN "MARFINITE".

6.5 Comparação dos Resultados do Processo Proposto com Sistemas de ER do Português

Nesta seção é apresentada uma comparação dos resultados do processo proposto com outros sistemas de ER do Português, considerando uma relação pré-definida, especificamente a relação de Localização ("ocorre_em" / "sede_de") definida na trilha ReREIEM do Segundo HAREM.

Conforme descrito nas seções anteriores, é difícil realizar uma avaliação entre sistemas de ER para o Português, principalmente pelo fato de os poucos trabalhos que tratam tal tarefa para essa língua utilizarem diferentes recursos, como corpus e parser. Entretanto, a relação de Localização foi tratada por dois sistemas participantes da trilha ReREIEM (SeRELeP e REMBRANDT) utilizando a coleção dourada do Segundo HAREM.

Para a realização do comparativo, selecionamos da mesma coleção dourada (Segundo HAREM) um subconjunto constituído de 40 descritores positivos que descrevem a relação de Localização e de 40 descritores negativos, considerando a anotação de referência dessa coleção.

Na Tabela 6.20 é apresentado o comparativo da extração da relação de Localização da trilha ReREIEM entre os seguintes resultados: processo proposto; sistema SeRELeP e sistema REMBRANDT.

	Abrangência	**Precisão**	**F-Measure**	**Máx. Sistema**	**Sistema**
Processo proposto	**0.40**	**0.76**	**0.52**	40	21
SeRELeP [14]	0.27	0.36	0.31	384	140
REMBRANDT [16]	0.13	0.40	0.20	42	17

Tabela 6.20 – Comparativo dos resultados da relação de Localização.

Cabe salientar que nos resultados consideradou-se a diferença do número de instâncias da relação de Localização utilizadas (Máx. Sistema: número total de instâncias, Sistema: número de instâncias encontradas por cada sistema).

Em comparação aos sistemas SeRELeP e REMBRANDT, a taxa de Precisão alcançada é bem superior em relação a esses sistemas, bem como a taxa de F-measure (52%) alcançada pelo processo proposto. Diferentemente do processo proposto, os sistemas SeRELeP e REMBRANDT não utilizam aprendizado de máquina.

Neste capítulo foi apresentada a avaliação experimental do processo para extração de descritores de relação entre ENs do Português, as diferentes configurações para o modelo CRF foram apresentadas, seguida da avaliação e da discussão dos resultados. Uma análise de erros das extrações dos descritores de relação foi descrita, e por fim, uma comparação entre o processo proposto e sistemas de ER do Português considerando a uma relação pré-definida (relação de Localização) foi apresentado. No Capítulo 7 são apresentadas as considerações finais deste trabalho.

7. Considerações Finais

Esta tese de doutorado teve como objetivo tratar da tarefa de ER a partir de textos da Língua Portuguesa, um desafio para a área de EI. É importante destacar que o português é uma língua carente de recursos, o que torna a tarefa ainda mais desafiadora.

Conforme apresentado ao longo deste trabalho, diferentes abordagens computacionais têm sido estudadas e aplicadas, e recursos linguísticos têm sido considerados para a solução desse problema. Além disso, há uma variedade considerável de formas e métodos de avaliação de sistemas de ER.

Uma análise detalhada do estado da arte sobre a tarefa de extração automática de relações semânticas foi apresentada no Capítulo 3. Esse estudo resultou na proposta de um processo para extração de descritores de relação entre ENs do domínio de Organizações (Organização, Pessoa e Local) em textos da Língua Portuguesa, apresentado no Capítulo 5.

Um dos objetivos de aplicar aprendizado supervisionado para ER é estudar as diferentes *features* associadas aos exemplos positivos e negativos sobre uma coleção de documentos anotados. Além disso, pretendia-se definir *features* capazes de identificar as instâncias de relação. Na Seção 5.4, diferentes conjuntos de *features* foram definidos baseados em informações de POS; no item lexical; sintáticas; semânticas, e na categoria da EN. Esses foram adaptados para o Português [4, 22, 66, 67, 73]. O corpus de referência utilizado é composto por um subconjunto das coleções douradas do HAREM, a qual foram adicionadas outras anotações (Seção 4.2).

Na avaliação experimental, diferentes configurações de *features* de entrada foram avaliadas, gerando diferentes modelos CRFs (Capítulo 7). Dentre os resultados alcançados de *F-measure*, destaca-se a configuração F6 com os melhores resultados na classificação de descritores corretos para todas as bases de dados (ORG-ORG, ORG-PES, ORG-LOCAL e ORG-PES-LOCAL). Isso ocorre em razão da *feature* semântica baseada na categoria da EN prover uma informação importante sobre o tipo de descritor de relação que se deseja identificar.

Na avaliação dos descritores parcialmente corretos para as bases ORG-ORG, ORG-PES e ORG-LOCAL mantiveram-se os melhores resultados de F-measure com a configuração F6. Para a base ORG-PES destacou-se também a configuração F3, porque *features* sintáticas baseadas em aposto, objeto direto e núcleo mostraram-se relevantes para a identificação dos descritores de relação dessa base. Já para a base ORG-LOCAL, destacou-se também a configuração F2, uma vez que, as *features* baseadas em itens lexicais já conseguem auxiliar na identificação de descritores que expressam a relação de "localização" que ocorre nessa base. Por fim, a base ORG-PES-LOCAL apresentou o melhor resultado de F-measure para descritores parcialmente corretos com a configuração F4. O bom desempenho ocorreu porque essa *feature* conseguir caracterizar os elementos que formam os descritores dessa base.

Nota-se que, em geral, o maior número de descritores etiquetados corretamente em todas as bases foram os que expressam relações não-verbais (ver gráfico da Figura 6.1). A base ORG-PES foi a que alcançou o maior número de descritores etiquetados corretamente, além das melhores taxas

de desempenho em relação às demais bases, considerando os descritores corretos e os descritores parcialmente corretos. Cabe salientar que, devido as características da base ORG-PES, a maioria dos descritores de relação é não-verbais, como a relação de "vínculo-institucional" contida nessa base.

Já a base ORG-ORG apresentou taxas de desempenho mais baixas se comparada às demais bases. Além disso, diferentemente do que ocorreu com as demais bases, a base ORG-ORG apresentou mais exemplos parcialmente corretos. Isso é consequência do tipo/aspectos dos descritores de relação dessa base, os quais, na sua maioria, são verbais e formados por vários elementos.

A partir da análise dos resultados, conclui-se que a tarefa de extração de descritores de relação entre ENs do Português pode ser tratada como um problema de etiquetagem de dados, uma vez que, o modelo CRF alcançou taxas de desempenho competitivas para essa tarefa. Os trabalhos relacionados de ER do Português reportam resultados com base em outros corpora, o que torna a comparação entre dos resultados difícil. No entanto, apresentamos na Tabela 7.1 os sistemas que utilizam a coleção do HAREM/ReRelEM [6, 14, 16, 20] e o trabalho de Batista et al. [6], os quais tratam ER entre ENs. Assim, podemos ter uma ideia dos níveis alcançados em outros estudos. Diferentemente desses trabalhos, as relações classificadas pelo CRF não são conhecidas, somente as categorias das ENs são definidas previamente (parâmetros da relação).

Trabalhos	**Dados/Corpora**	**Avaliação**	**Resultados, %**
SeRELeP [14]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Identidade F= 68%, Inclusão F= 45%, Localização F= 31%. Todas as relações F= 36%.
REMBRANDT [16]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Identidade F= 73%, Inclusão F= 33%, Localização F= 20%. Todas as relações F= 45%.
SEI-Geo [20]	Coleção Dourada do HAREM/ReRelEM.	Coleção Dourada anotada manualmente.	Inclusão F= 45%. Todas as relações F= 27% .
Batista et al. [6]	97.988 frases retiradas da DBPédia da Wikipédia em Português	subconjunto de teste formado por 625 frases (89.601 relações anotadas manualmente entre ENs)	local-de-enterro F= 67% pessoa-chave-em F= 11% localizado-em F= 92% origem-de F = 81% antepassado-de F = 62% parte-de F = 62% sucessor-de F = 24% parceiro F = 28% outros F = 63% Média das relações F= 55.6%.
Processo Proposto	82 textos das Coleções Douradas do HAREM	304 relações anotadas manualmente entre ENs	ORG-ORG corretas F= 39% ORG-ORG parcial F = 55% ORG-PES corretas F = 57% ORG-PES parcial F = 63% ORG-LOCAL corretas F = 52% ORG-LOCAL parcial F = 54% ORG-PES-LOCAL corretas F = 46% ORG-PES-LOCAL parcial F = 55%

Tabela 7.1 – Resultados de trabalhos de ER do Português.

Com base na revisão da literatura, destacaram-se trabalhos recentes para o Inglês que utilizam o modelo probabilístico CRF em diversas aplicações de PLN, em especial para a tarefa de ER são alcançados resultados promissores [4,29,59]. O modelo CRF é aplicado eficientemente em problemas de etiquetagem de sequências estruturadas, como o texto em linguagem natural. Esse modelo pode obter uma etiquetagem ideal globalmente, e utilizar características simples para expressar complexas estruturas de contexto. Para nosso conhecimento, este trabalho é o primeiro que utiliza CRFs de cadeias lineares para ER do Português.

Seguindo estas observações, apresentamos um processo para extração de descritores de relação entre ENs em textos da Língua Portuguesa que compreende as etapas de pré-processamento dos textos (anotação de POS, Chunker, Semântica, REN, BIO), geração das *features* (POS, sintáticas, semânticas), e por fim a aplicação do modelo CRF para a extração destes descritores. Definimos um conjunto de *features* capaz de classificar os descritores de relação (configuração *F6*), destacando-se a *feature* semântica baseada na categoria da EN por caracterizar melhor o tipo de relação que se deseja identificar entre o par de ENs.

7.1 Contribuições

A proposta de um processo para extração de descritores de relação, em particular entre ENs em textos do Português é uma das principais contribuições deste trabalho. Conforme apresentado na Seção 3.2, em geral, os sistemas de ER para o Português são baseados em heurísticas. Na literatura, existem poucos trabalhos em ER para essa língua que usam técnicas de aprendizado de máquina [6, 40], ao contrário do que ocorre para o Inglês. Segundo [59], soluções que representam o estado da arte de REN e ER utilizam métodos de aprendizado de máquina.

Dentre as contribuições, destacam-se também a preparação e a disponibilização de um subcorpus para a tarefa de ER para o Português. A um subconjunto das coleções douradas do HAREM foi adicionada a anotação dos descritores de relação envolvendo as ENs das categorias Pessoa, Organização e Local. Esse é um importante recurso para a Língua Portuguesa, uma vez que, a tarefa de extração de descritores de relação é recente e corpora com instâncias de relações anotadas não estão disponíveis [66].

A partir de um estudo do estado da arte sobre a tarefa de extração automática de relações semânticas, definiu-se diferentes conjuntos de *features* para o aprendizado. Portanto, outra contribuição para este trabalho é a definição das *features*, posto que uma das dificuldades para a aplicação de aprendizado supervisionado é a escolha de *features* que melhor representem os exemplos.

Por fim, a avaliação experimental contribuiu para a avaliação do processo proposto, bem como para a análise dos conjuntos de *features*. Diferentes configurações de *features* de entrada para o CRF foram avaliadas utilizando as métricas de desempenho usuais, e o teste de hipótese *T-test* foi aplicado para verificar os ganhos significativos entre cada *configuração* testada. A partir da análise de cada configuração de *feature*, foi possível verificar quais delas são realmente relevantes para cada base de dados.

7.2 Trabalhos Futuros

Como trabalhos futuros, pretende-se aplicar outros métodos de aprendizado de máquina para ER, os quais foram utilizados em trabalhos como HMM [43], MEM [61], SVM [30], KNN [6].

Pretende-se também expandir as bases de dados, adicionando mais textos anotados. Outro objetivo é aplicar processo aprendido num novo corpus constituído de textos de outros domínios. Além disso, pretende-se analisar, de maneira individual, o potencial das *features*, e ampliar o escopo do texto para observação das *features*.

Acrescenta-se que se pretende realizar uma extensão do processo proposto de ER também para outras línguas.

Bibliografia

[1] Sandra Collovini Abreu, Tiago Luis Bonamigo, and Renata Vieira. A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, pages 1–19, 2013.

[2] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *5th ACM International Conference on Digital Libraries*, pages 85–94, 2000.

[3] Michele Banko, Michael J Cafarella, Stephen Soderl, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *In IJCAI*, pages 2670–2676, 2007.

[4] Michele Banko and Oren Etzioni. The tradeoffs between open and traditional relation extraction. In Kathleen McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *ACL*, pages 28–36. The Association for Computer Linguistics, 2008.

[5] David Batista, Mário J. Silva, Francisco Couto, and Bibek Behera. Geographic signatures for semantic retrieval. In *6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, 2010.

[6] David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário Silva. Extracção de relações semânticas de textos em português explorando a DBpédia e a Wikipédia. *linguamatica*, 5(1):41–57, 2013.

[7] Kedar Bellare and Andrew Mccallum. Learning Extractors from Unlabeled Text using Relevant Databases. In *Sixth International Workshop on Information Integration on the Web (IIWeb)*, 2007.

[8] Eckhard Bick. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus, 2000.

[9] Eckhard Bick. Multi-level ner for portuguese in a cg framework. In *PROPOR 2003*, volume 2721 of *Lecture Notes in Computer Science*, pages 118–125, Faro, Portugal, 2003. Springer.

[10] Eckhard Bick. Functional aspects in portuguese ner. In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira, and Maria Carmelita Dias, editors, *PROPOR*, volume 3960 of *Lecture Notes in Computer Science*, pages 80–89. Springer, 2006.

[11] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: High-performance learning name-finder. In *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, 1997.

[12] Kurt D. Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*, pages 1247–1250, 2008.

[13] Sergey Brin. Extracting patterns and relations from the world wide web. In Paolo Atzeni, Alberto O. Mendelzon, and Giansalvatore Mecca, editors, *WebDB*, volume 1590 of *Lecture Notes in Computer Science*, pages 172–183. Springer, 1998.

[14] Mírian Brucksen, José Guilherme Camargo Souza, Renata Vieira, and Sandro Rigo. Sistema serelep para o reconhecimento de relações entre entidades mencionadas. In C. Mota and D. Santos, editors, *Segundo HAREM*, chapter 14, pages 247–260. Linguateca, 2008.

[15] Razvan C. Bunescu. Learning to extract relations from the web using minimal supervision. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics – ACL 2007*, 2007.

[16] Nuno Cardoso. Rembrandt – reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto. In C. Mota and D. Santos, editors, *Segundo HAREM*, chapter 11, pages 195–211. Linguateca, 2008.

[17] Nuno Cardoso. Rembrandt - a named-entity recognition framework. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[18] Andrew Carlson, Justin Betteridge, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupling semi-supervised learning of categories and relations. In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 1–9, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[19] Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, Diana Santos, and Cláudia Freitas. Segundo harem: Modelo geral, novidades e avaliação. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. 2008.

[20] Marcirio S. Chaves. Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem. In C. Mota and D. Santos, editors, *Segundo HAREM*, chapter 13, pages 231–245. Linguateca, 2008.

[21] Marcirio S. Chaves, Mário J. Silva, and Bruno Martins. A geographic knowledge base for semantic web applications. In C. A. Heuser, editor, *20th Brazilian Symposium on Databases*, pages 40–54, 2005.

[22] Yingying Chen, Qinghua Zheng, Wei Wang, and Yan Chen. Knowledge element relation extraction using conditional random fields. In *CSCWD*, pages 245–250, 2010.

[23] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition with a maximum entropy approach. In *In Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pages 160–163, 2003.

[24] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM–LbR '10, pages 52–60, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[25] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. An analysis of open information extraction based on semantic role labeling. In *K–CAP*, pages 113–120, 2011.

[26] Philipp Cimiano and Johanna Wenderoth. Automatically learning qualia structures from the web. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 28–37, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[27] Sandra Collovini, Fernando Grando, Marlo Souza, Larissa Freitas, and Renata Vieira. Semantic relations extraction in the organization domain. In *Proceedings of IADIS International Conference on Applied Computing 2011*, pages 99–106, Rio de Janeiro, RJ, 2011.

[28] Sandra Collovini and Renata Vieira. Análise de expressões referenciais em corpus anotado da língua portuguesa. In *V Best MSc dissertation/PhD thesis contest (CTDIA'2006)*, Ribeirão Preto, SP, 2006. Proceedings of the SBIA-IBERAMIA.

[29] Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 296–303, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[30] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 423–429, Barcelona, Spain, 2004.

[31] J. R. Curran, T. Murphy, and B. Scholz. Minimising semantic drift with mutual exclusion bootstrapping. *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pages 172–180, 2007.

[32] Luciano Del Corro and Rainer Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 355–366, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[33] Daniela do Amaral and Renata Vieira. O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, Fortaleza, Brazil, 2013. SBC.

[34] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ace) program: Tasks, data, and evaluation. In Maria Tereza Lino, Maria Francisca Xavier, Fatima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation – LREC 2004*, pages 837–840, Lisboa, 2004.

[35] Nelson Francisco Favilla Ebecken, Maria Celia Santos Lopes, and Myrian Christina de Aragão Costa. Mineração de textos. In Solange Oliveira Rezende, editor, *Sistemas Inteligentes: fundamentos e aplicações*, chapter 13, pages 337–370. Manole, Barueri, SP, 2005.

[36] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. *Artiicial Intelligence*, 165(1):91–134, 2005.

[37] Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web–scale information extraction in knowitall: (preliminary results). In *WWW*, pages 100–110, 2004.

[38] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. Open information extraction: The second generation. In *Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI*, pages 3–10, 2011.

[39] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545, 2011.

[40] William Paulo Ducca Fernandes, Eduardo Motta, and Luiz Milidiú. Quotation extraction for portuguese. In *8th Brazilian Symposium in Information and Human Language Technology - STIL'2011*, pages 204–208, Cuiabá, Brasil, 2011.

[41] Liliana Ferreira, César Oliveira, Antônio Teixeira, and João Cunha. Extração de informação de relatórios médicos. *linguamatica*, 1(1):89–101, 2009.

[42] Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370, 2005.

[43] Dayne Freitag and Andrew Mccallum. Information extraction with hmm structures learned by stochastic optimization. In *In Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 584–589. AAAI Press, 2000.

[44] Cláudia Freitas, Cristina Mota, Diana Santos, Hugo Gonçalo Oliveira, and Paula Carvalho. Second harem: Advancing the state of the art of named entity recognition in portuguese. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[45] Cláudia Freitas and Violeta Quental. Subsídios para a elaboração automática de taxonomias. In *V Workshop em Tecnologia da Informação e da Linguagem Humana - TIL 2007*, Rio de Janeiro, Brasil, 2007.

[46] Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, and Cristina Mota. *Relações semânticas do ReRelEM: além das entidades no Segundo HAREM*, chapter 4, pages 75–94. Linguateca, 2008.

[47] Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, Avignon, France, 2012. Association for Computational Linguistics.

[48] Hugo Gonçalo Oliveira, Hernani Costa, and Paulo Gomes. Extracção de conhecimento léxico-semântico a partir de resumos da Wikipédia. In *INForum 2010 - II Simpósio de Informática, Track on Gestão e Tratamento de Informação*, INForum'10, pages 537–548, Braga, Portugal, September 2010.

[49] Ralph Grishman. The nyu system for muc-6 or where's the syntax? In *MUC*, pages 167–175, 1995.

[50] Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[51] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 415, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[52] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[53] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 33–38, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[54] Jerry R Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel, and Mabry Tyson. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. *English*, pages 1–25, 1997.

[55] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, pages 541–550, 2011.

[56] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics*. Macmillan, New York, USA, 1978.

[57] Hideki Isozaki and Hideto Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[58] J. F. Jiang and S. X. Wang. A bootstrapping method for acquisition of bi-relations and bi-relations patterns. *Journal of Chinese Information Processing*, 19(2):71–77, 2005.

[59] Jing Jiang. Information extraction from text. In *Mining Text Data*, pages 11–41. 2012.

[60] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Education Ltd., London, 2 edition, 2009.

[61] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, ACL demo '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[62] Ralf Krestel, Sabine Bergler, and René Witte. Minding the source: Automatic tagging of reported speech in newspaper articles. In *LREC*, 2008.

[63] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[64] Haibo Li, Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Using graph based method to improve bootstrapping relation extraction. In *CICLing (2)*, pages 127–138, 2011.

[65] Xin Li and Dan Roth. Learning question classifiers. In Tsuei-Er Chen and Yi-Fen Liu, editors, *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 556–562. Morgan-Kaufman Publishers, San Francisco, CA, USA, 2002.

[66] Yaliang Li, Jing Jiang, Hai Leong Chieu, and Kian Ming A. Chai. Extracting relation descriptors with conditional random fields. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 392–400, Chiang Mai, Thailand, 2011. Asian Federation of Natural Language Processing.

[67] Xiao Ling and Daniel S. Weld. Fine-grained entity recognition. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI*, Toronto, Ontario, Canada, 2012. AAAI Press.

[68] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[69] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In *EMNLP-CoNLL*, pages 523–534. ACL, 2012.

[70] Andrew Mccallum. Efficiently inducing features of conditional random fields. In *Uncertainty in Artificial Intelligence*, pages 403–410, San Francisco, CA, 2003. Morgan Kaufmann.

[71] Tara McIntosh and James R. Curran. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 396–404, Suntec, Singapore, 2009. Association for Computational Linguistics.

[72] Andrei Mikheev, Moens Moens, and Claire Grover. Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, Norway, 1999.

[73] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[74] Maria Carolina Monard and José Augusto Baranauskas. Conceitos sobre aprendizado de máquina. In Solange Oliveira Rezende, editor, *Sistemas Inteligentes: fundamentos e aplicações*, chapter 4, pages 90–114. Manole, Barueri, SP, 2005.

[75] Guillermo Moncecchi, Jean-Luc Minel, and Dina Wonsever. A survey of kernel methods for relation extraction. In *Workshop on Natural Language Processing and Web-based technologies in conjunction with IBERAMIA 2010*, 2010.

[76] Cristina Mota, Diana Santos, and Elisabete Ranchhod. Avaliação de reconhecimento de entidades mencionadas: Princípio de harem. In Diana Santos, editor, *Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa*, chapter 14, pages 161–176. IST Press, 2007.

[77] MUC-6. Coreference task definition. In *Proceedings of the Sixth Message Understanding Conferece - MUC-6*, 1995.

[78] MUC-7. Coreference task definition. In *Proceedings of the Seventh Message Understanding Conferece - MUC-7*, 1997.

[79] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.

[80] NIST and ACE. Automatic content extraction 2008 evaluation plan (ace08). Technical report, NIST, 2008.

[81] OpenNLP. open-source framework to develop natural language applications, 2010.

[82] David D. Palmer and David S. Day. A statistical profile of the named entity task. In *ANLP*, pages 190–193, 1997.

[83] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *International Conference on Computational Linguistics/Association*, pages 113–120, Sydney, Australia, 2006. ACL Press.

[84] Bruno Pouliquen, Ralf Steinberger, and Clive Best. Automatic detection of quotations in multilingual news. In *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, pages 487–492, Borovets, Bulgaria, 2007.

[85] Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *COLING*, pages 697–704, 2008.

[86] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora*, pages 82–94. Cambridge MA, USA, 1995.

[87] Yael Ravin and Nina Wacholder. Extracting names from natural-language text. Technical report, IBM, 1996.

[88] Dan Roth and Wen tau Yih. Integer linear programming inference for conditional random fields. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 736–743, 2005.

[89] Daniel Santos, Nuno Mamede, and Jorge Baptista. Extraction of family relations between entities. In Miguel P. Correia Luís S. Barbosa, editor, *Proceedings of the INForum 2010 - II Simpósio de Informática*, pages 549–560, Braga, Portugal, 2010.

[90] Diana Santos. Avaliação conjunta. In Diana Santos, editor, *Avaliação Conjunta: Um novo paradigma no processamento computacional da língua portuguesa*, chapter 1, pages 1-12. IST Press, 2007.

[91] Diana Santos and Nuno Cardoso. *Breve introdução ao HAREM*, chapter 1, pages 1–16. Linguateca, 2007.

[92] Sunita Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377, 2008.

[93] Luís Sarmento and Sérgio Nunes. Automatic extraction of quotes and topics from news feeds. In *14th Doctal Symposium on Informatics Engineering*, 2009.

[94] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*, pages 104–107, 2004.

[95] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 134–141, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[96] Lucia Specia and Enrico Motta. A hybrid approach for extracting semantic relations from texts. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 57–64, Sydney, Australia, 2006. Association for Computational Linguistics.

[97] Ang Sun. A two-stage bootstrapping algorithm for relation extraction. In *Proceedings of RANLP 2009 - Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2009.

[98] Ang Sun, Ralph Grishman, and Satoshi Sekine. Semi-supervised relation extraction with large-scale word clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 521–529, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[99] Charles Sutton. Conditional probabilistic context-free grammars. Master's thesis, University of Massachusetts, 2004.

[100] Hristo Tanev, Vanni Zavarella, Jens Linge, Mijail Kabadjov, Jakub Piskorski, Martin Atkinson, and Ralf Steinberger. Exploiting machine learning techniques to build an event extraction system for portuguese and spanish. *linguamatica*, 1(2):55–66, 2009.

[101] Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, and Ji Wang. Automatic extraction of hierarchical relations from text. In *Proceedings of the 3rd European conference on The Semantic Web: research and applications*, ESWC'06, pages 215–229, Berlin, Heidelberg, 2006. Springer-Verlag.

[102] Fei Wu. Machine reading: form wikipedia to the web. Master's thesis, University of Washington, Seattle, WA, USA, 2010.

[103] Fei Wu and Daniel S. Weld. Open information extraction using wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[104] Clarissa Castellã Xavier and Vera Lúcia Strube de Lima. A semi-automatic method for domain ontology extraction from portuguese language wikipedia's categories. In *The Brazilian Symposium on Artificial Intelligence (SBIA)*, volume 1, pages 11–20, São Bernardo do Campo, 2010. Advances in Artificial Intelligence - SBIA 2010.

[105] Clarissa Castellã Xavier, Vera Lúcia Strube de Lima, and Marlo Souza. Open information extraction based on lexical-syntactic patterns. In *Brazilian Conference on Intelligent Systems (BRACIS 2013)*, Fortaleza, Brazil, 2013. SBC.

[106] Alexander Yates, Michele Banko, Matthew Broadhead, Michael J. Cafarella, Oren Etzioni, and Stephen Soderland. Textrunner: Open information extraction on the web. In *HLT-NAACL (Demonstrations)*, pages 25–26, 2007.

[107] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106, 2003.

[108] Suxiang Zhang, Suxian Zhang, and Guoyang Gao. Automatic entity relation extraction based on conditional random fields. In *FSKD (2)*, pages 286–290, 2008.

[109] Shubin Zhao and Ralph Grishman. Extracting relations with integrated information using kernel methods. In *ACL*. The Association for Computer Linguistics, 2005.

[110] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: a statistical approach to extracting entity relationships. In *18th International Conference on World wide web*, pages 101–110, New York, NY, USA, 2009. ACM.

8. Apêndice A

Exemplos de descritores de relação positivos das bases são ilustrados nas Tabelas: 8.1, 8.2, 8.3.

Relações	Instâncias de Relação	Descritores de Relação
Não-verbais	... *Núcleo Interactivo de Astronomia* (NUCLIO) **no** *Centro de Interpretação Ambiental da Ponta do Sal* ...	**no**
	... *Câmara Municipal* **de** *Cascais* ...	**de**
	... *Macquarie Futures* **dos** *EUA* ...	**dos**
	... *Hospital de São João*, **no** *Porto* ...	**no**
	... *Creative Commons* **no** *Brasil* ...	**no**
	... *Comissão de Controle e Gestão Fiscal*, do Ministério da Fazenda, publicou, **no** *Diário Oficial* ...	**no**
	... *Porto Fino*, **na rua** *Padrão* ...	**na rua**
	... *McDonald's* **na** *Avenida Jorge Amado* ...	**na**
	... *Brooks* **da** *Força Aérea* ...	**da**
Verbais	... *Serrambi Viagens e Turismo* **promove eventos de** grande importância para o turismo de *Pernambuco* ...	**promove eventos de**
	... *Legião da Boa Vontade*, instituição educacional, cultural, beneficente e filantrópica, **foi fundada no** *Brasil* ...	**foi fundada no**
	... *Goa Tourism Development Corporation Office* **organiza excursões a** *Goa* ...	**organiza excursões a**
	... *França* **renunciasse a quaisquer reclamações sobre a foz do** *Amazonas* ...	**renunciasse a quaisquer reclamações sobre a foz do**
	... *CICA 1* liderada pelo Tenente-Coronel Carlos Azeredo **toma o** *Quartel General da Região Militar do Porto* ...	**toma o**
	... *BC9* de Viana do Castelo **tomam o** *Aeroporto de Pedras Rubras* ...	**tomam o**
	... *Waterford*, **visite a** fábrica de cristais *Waterford* ...	**visite a**
	... *World Alliance of Reformed Churches* - **condena a guerra no** *Iraque* ...	**condena a guerra no**
	... *CCB* **apresentou ao público de** *Lisboa* ...	**apresentou ao público de**

Tabela 8.1 – Exemplos positivos da base ORG-LOCAL.

Relações	Instâncias de Relação	Descritores de Relação
Não-verbais	... *Observatório Astronómico* **da** *Universidade de Coimbra* ...	**da**
	... *Centro de Astronomia* **da** *Universidade do Porto* ...	**da**
	... *Conselho* **da** *UE* ...	**da**
	... *Santa Isabel*(que acolhe meninas desprotegidas), **da** *Madre Teresa de Calcutá* ...	**da**
	... *CBKC* (Confederação Brasileira de Cinofilia), **órgão filiado ao** *FCI* ...	**órgão filiado ao**
	... *Ministério da Indústria* **do** *Governo* ...	**do**
	... *Comissão de Controle e Gestão Fiscal*, **do** *Ministério da Fazenda* ...	**do**
	... *Falintil* - Forças de Defesa de Timor-Leste, **da** *Polícia Nacional de Timor-Leste* ...	**da**
	... *Europa* **para a** *África* ...	**para a**
	... *Joint Non-Lethal Weapons Program*(JNLWP), **do** *Pentágono* ...	**do**
	... *Secretaria de Desenvolvimento Sustentável* **do** *Governo* ...	**do**
Verbais	... *OLP*, **dominada pela** *Fatah* ...	**dominada pela**
	... *Hamas* a **mandar em** *Gaza* ...	**mandar em**
	... a *Biblioteca Central da Marinha* (BCM) **é sucessora da** antiga *Biblioteca da Real Academia dos Guardas-Marinhas* ...	**é sucessora da**
	... *PSDB* **Vinha negociando um acordo com a** *Igreja Universal* ...	**vinha negociando um acordo com a**
	... O governo da *Rússia* **vai apoiar uma ação da** *Otan*	**vai apoiar uma ação da**
	... *México* **prende assessor de deputado do** *PRI*	**prende assessor de deputado do**
	... *Corpo Nacional de Escutas* (CNE) **atribuiu ontem à** *Câmara de Santo Tirso* ...	**atribuiu ontem à**
	... *Portugal* **conseguiu da** *Espanha* ...	**conseguiu da**
	... *CIOE* **tomam a** *RTP* ...	**tomam a**

Tabela 8.2 – Exemplos positivos da base ORG-ORG.

Relações	Instâncias de Relação	Descritores de Relação
Não-verbais	... *Francis WURTZ* , **da** *CEUE/EVN* ...	**da**
	... *Brian CROWLEY* , **da** *UEN* ...	**da**
	... *Nauman Barakat*, **vice-presidente da** *Macquarie Futures* ...	**vice-presidente da**
	... *Moncef Kaabi* , **da** *Natixis* ...	**da**
	... *Hugo Doménech*, **professor da** *Universidade Jaume de Castellón* ...	**professor da**
	... *Mario Lúcio Vaz*, **diretor da** *Central Globo de Controle de Qualidade* ...	**diretor da**
	... *Ministro dos Negócios Estrangeiros*, **embaixador de** *Portugal* ...	**embaixador de**
	... *António Ribeiro*, **em declarações ao** *PÚBLICO* ...	**em declarações ao**
	... *Fernando Gomes*, **presidente da** *Câmara Municipal do Porto* ...	**presidente da**
	... *Steve Jobs*, o **director-geral** da empresa, foi o ponto alto para os fãs da *Apple* ...	**director-geral**
Verbais	... *Santos Ferreira* **tiver sucesso no** *BCP* ...	**tiver sucesso no**
	... *Ministro da Defesa Nacional* **visita** *Forças Nacionais Destacadas* ...	**visita**
	Abbas **pode fingir mandar em** *Ramallah*	**pode fingir mandar em**
	... *RCM* **gostaria de ouvir o** *Concelho de Mafra* ...	**gostaria de ouvir o**
	... *Aristides Junqueira*, **que juntou as principais lideranças do** *Congresso* ...	**que juntou as principais lideranças do**
	... *Legião da Boa Vontade*, instituição educacional, cultural, beneficente e filantrópica, **foi fundada no Brasil pelo** jornalista, radialista e poeta *Alziro Zarur* ...	**foi fundada no Brasil pelo**
	... *Câmara de Santo Tirso* **tem apoiado**, incondicionalmente e desde sempre, o escutismo no concelho, criando melhores condições para que os *Agrupamentos de Escuteiros* ...	**tem apoiado**
	... *Saraiva Dias*, **vereador** substituto do presidente da autarquia, referiu ao *PÚBLICO* ...	**vereador**
	... *CICA 1* **liderada pelo** *Tenente-Coronel Carlos Azeredo* ...	**liderada pelo**
	... *Escola de Pilotagem* ajudou a formar dezenas de pilotos particulares e profissionais sendo, actualmente **dirigida pelo** *Cmdt. João Filhó* ...	**dirigida pelo**

Tabela 8.3 – Exemplos positivos da base ORG-PES.

9. Apêndice B

Os resultados das bases ORG-ORG, ORG-PES e ORG-LOCAL com validação cruzada de *10-folds* são apresentados a seguir:

ORG-ORG com validação cruzada de *10-folds*:

ORG-ORG (*10-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	38	3	49	0.42	0.69	0.52
	I-REL	0	84	197	0.29	0.53	0.38
	O	17	71	1614	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	29	6	55	0.32	0.70	0.44
	I-REL	0	86	195	0.30	0.51	0.38
	O	12	75	1615	0.94	0.86	0.90
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	38	5	47	0.42	0.71	0.53
	I-REL	0	96	185	0.34	0.50	0.40
	O	15	91	1596	0.93	0.87	0.90
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	39	5	46	0.43	0.68	0.53
	I-REL	0	94	187	0.33	0.48	0.39
	O	18	94	1590	0.93	0.87	0.90
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	32	3	55	0.35	0.71	0.47
	I-REL	0	68	213	0.24	0.58	0.34
	O	13	46	1643	0.96	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	34	2	54	0.37	0.70	0.49
	I-REL	2	121	189	0.38	0.61	0.47
	O	14	47	1641	0.96	0.86	0.90

Tabela 9.1 – Classificação BIO de ORG-ORG por conjunto de features.

Em geral, os resultados da validação cruzada com *10-folds* na base ORG-ORG, apresentou taxas inferiores à validação cruzada com *5-folds* (ver Tabela 6.3). Destaca-se na Tabela 9.2, a taxa de 46% de Precisão na configuração *F2* para descritores corretos. A configuração *F3* alcançou a melhor taxa de Precisão para os descritores parcialmente corretos (71%), e as melhores taxas de F-measure: 33% e 53% para descritores corretos e descritores parcialmente corretos, respectivamente.

ORG-ORG	Descritores Corretos				Descritores Parcialmente Corretos			
(10-folds)	#C	A	P	F	#C	A	P	F
F1=POS	23	0.25	0.41	0.31	38	0.42	0.69	0.52
F2=POS+LEX	19	0.21	**0.46**	0.29	29	0.32	0.70	0.43
F3=POS+LEX+SINT	24	0.26	0.45	**0.33**	38	0.42	**0.71**	**0.53**
F4=POS+LEX+SINT+PAD	24	0.26	0.42	0.32	**39**	**0.43**	0.68	**0.53**
F5=POS+LEX+SINT+PAD+FR	20	0.22	0.44	0.29	32	0.35	**0.71**	0.46
F6=POS+LEX+SINT+PAD+FR+SEM	21	0.23	0.44	0.30	34	0.37	**0.71**	0.48

Tabela 9.2 – Resultados de ORG-ORG por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

ORG-PES com validação cruzada de *10-folds*:

ORG-PES (*10-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	50	6	49	0.47	0.67	0.55
	I-REL	4	123	185	0.39	0.53	0.45
	O	20	99	1430	0.92	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	47	6	52	0.44	0.74	0.55
	I-REL	1	135	176	0.43	0.54	0.48
	O	15	109	1425	0.91	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	53	5	47	0.50	0.76	0.60
	I-REL	1	145	166	0.46	0.55	0.50
	O	15	109	1425	0.91	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	52	5	48	0.49	0.76	0.60
	I-REL	1	144	167	0.46	0.56	0.50
	O	15	106	1428	0.92	0.86	0.89
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	49	4	52	0.46	0.77	0.58
	I-REL	2	120	190	0.38	0.61	0.47
	O	12	71	1466	0.94	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	53	2	50	0.50	0.79	0.61
	I-REL	2	121	189	0.38	0.61	0.47
	O	12	73	1464	0.94	0.85	0.90

Tabela 9.3 – Classificação BIO de ORG-PES por conjunto de features.

O desempenho das configurações de features para a base ORG-PES com validação de *10-folds* se manteve similar à com *5-folds* (ver Tabela 6.5), porém com taxas mais baixas, conforme apresentado na Tabela 9.4). Manteve-se o ganho em Precisão na configuração *F2* para descritores corretos em relação à configuração anterior (taxa de significância de 99%). A configuração *F-semant* seguiu apresentando as melhores taxas de desempenho.

ORG-PES	Descritores Corretos				Descritores Parcialmente Corretos			
(Cross 10-folds)	#C	A	P	F	#C	A	P	F
F1=POS	31	0.29	0.41	0.34	50	0.47	0.67	0.56
F2=POS+LEX	37	0.35	0.58*	0.44	47	0.44	0.74	0.55
F3=POS+LEX+SINT	43	0.40	0.62	0.49	**53**	**0.50**	0.76	0.60
F4=POS+LEX+SINT+PAD	42	0.40	0.61	0.48	52	0.49	0.76	0.60
F5=POS+LEX+SINT+PAD+FR	40	0.38	0.63	0.47	49	0.46	0.77	0.57
F6=POS+LEX+SINT+PAD+FR+SEM	**44**	**0.41**	**0.65**	**0.51**	**53**	**0.50**	**0.79**	**0.61**

Tabela 9.4 – Resultados de ORG-PES por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

ORG-LOCAL com validação cruzada de *10-folds*:

ORG-LOCAL (*10-folds*)	Matriz de Confusão da Classificação BIO						
		B-REL	**I-REL**	**O**	A	P	F
F1=POS	**B-REL**	43	3	63	0.39	0.65	0.49
	I-REL	1	55	236	0.18	0.43	0.26
	O	22	68	1675	0.94	0.84	0.89
		B-REL	**I-REL**	**O**	A	P	F
F2=POS+LEX	**B-REL**	48	1	60	0.44	0.76	0.55
	I-REL	1	61	230	0.20	0.46	0.28
	O	14	69	1682	0.95	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F3=POS+LEX+SINT	**B-REL**	43	2	64	0.39	0.66	0.49
	I-REL	2	78	212	0.26	0.44	0.33
	O	20	97	1648	0.93	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F4=POS+LEX+SINT+PAD	**B-REL**	44	2	63	0.40	0.67	0.50
	I-REL	1	79	212	0.27	0.46	0.34
	O	20	88	1657	0.93	0.85	0.89
		B-REL	**I-REL**	**O**	A	P	F
F5=POS+LEX+SINT+PAD+FR	**B-REL**	42	2	65	0.38	0.70	0.49
	I-REL	2	68	222	0.23	0.55	0.32
	O	16	52	1697	0.96	0.85	0.90
		B-REL	**I-REL**	**O**	A	P	F
F6=POS+LEX+SINT+PAD+FR+SEM	**B-REL**	47	4	58	0.43	0.77	0.55
	I-REL	2	77	213	0.26	0.57	0.36
	O	12	54	1699	0.96	0.86	0.90

Tabela 9.5 – Classificação BIO de ORG-LOCAL por conjunto de features.

Os resultados da base ORG-LOCAL com validação cruzada de *10-folds*, apresentados na Tabela 9.6, foram similares à com *5-folds* (veja Tabela 6.7), entretanto com taxas mais baixas. A configuração *F2* apresentou ganhos em Precisão para descritores corretos e também para descritores parcialmente corretos comparado à configuração anterior (grau de significância de 97.5% e 95%, respectivamente). A configuração *F6* manteve-se apresentando as melhores taxas de Precisão e F-measure para os descritores corretos e parcialmente corretos.

ORG-LOCAL	Descritores Corretos				Descritores Parcialmente Corretos			
(10-folds)	#Corretos	A	P	F	#C	A	P	F
F1=POS	30	0.27	0.45	0.34	43	0.39	0.65	0.49
F2=POS+LEX	38	0.34	0.60*	0.44	**48**	**0.44**	0.76*	**0.55**
F3=POS+LEX+SINT	36	0.33	0.55	0.41	43	0.39	0.66	0.49
F4=POS+LEX+SINT+PAD	37	0.33	0.56	0.42	44	0.40	0.67	0.50
F5=POS+LEX+SINT+PAD+FR	38	0.34	0.63	0.44	42	0.38	0.70	0.49
F6=POS+LEX+SINT+PAD+FR+SEM	**40**	**0.38**	**0.68**	**0.49**	45	0.43	**0.77**	**0.55**

Tabela 9.6 – Resultados de ORG-LOCAL por conjunto de Features. * indica que o valor atual é estatisticamente melhor do que o valor da linha anterior.

10. Apêndice C

Dr. Denise Nauderer Hogetop
Dr. Aline Aver Vanin

A etapa de anotação manual das relações foi realizada por nós, linguistas, individualmente. A partir de uma seleção de textos de coleções douradas do Primeiro e do Segundo HAREM, coube-nos julgar qual seria o descritor da relação entre as entidades nomeadas que ocorriam numa mesma sentença. Como essas entidades já haviam sido identificadas numa etapa anterior, nossa tarefa era anotar apenas as relações entre as entidades.

A análise dos textos para a extração manual das relações foi realizada com base em aspectos linguísticos, como, por exemplo, o papel das preposições no estabelecimento de relações, o papel dos verbos, adjetivos e nomes no que diz respeito de relações verbais e não verbais. A princípio, a orientação recebida era de verificar, a partir dos referidos corpora, quais sentenças continham, ou não, uma relação. Cabe ressaltar que o processo de anotação foi inteiramente manual.

Portanto, o documento deveria ser avaliado com vistas a verificar quais sentenças continham relações entre as entidades, objetos deste estudo. A partir disso, nossa tarefa era de copiar, em cada célula do documento: (1) as entidades identificadas; (2) a categoria dessas entidades; e (3) o trecho da sentença no qual a relação está definida. Nesse caso, cada uma de nós julgava a relação estabelecida de um ponto de vista linguístico. Assim, o nível de anotação linguística deste trabalho perpassou, em grande parte, questões sintáticas e semânticas, no que diz respeito ao papel de preposições, de verbos, nomes e adjetivos em determinadas relações.

Na Tabela 10.1 é apresentado um exemplo do arquivo de entrada para a anotação manual dos descritores de anotação, e na Tabela 10.2 é apresentado o arquivo de saída com a anotação manual do descritor e a indicação de quais entidades estão envolvidas nessa relação. As informações ilustradas nas colunas correspondem ao par de entidades (EN1 e EN2), as respectivas categorias (CATEG1 e CATEG2), e por fim o Descritor.

Sentença	EN1	CATEG1	EN2	CATEG2	Descritor
No próximo sábado, <CATEG=“PESSOA">Ronaldo Lemos</EM>, diretor do <CATEG=“ORGANIZACAO">Creative Commons</EM>, irá participar de um debate com os membros do <CATEG=“PESSOA">Superflex</EM>, na <CATEG=“LOCAL»Vermelho</EM>.					

Tabela 10.1 – Exemplo de entrada para a anotação dos descritores de relação.

Sentença	EN1	CATEG1	EN2	CATEG2	Descritor
No próximo sábado, <CATEG=“PESSOA"> Ronaldo Lemos</EM>, diretor do <CATEG=“ORGANIZACAO"> Creative Commons</EM>, irá participar de um debate com os membros do <CATEG=“PESSOA"> Superflex</EM>, na <CATEG=“LOCAL»Vermelho</EM>.	Ronaldo Lemos	PES	Creative Commons	ORG	diretor do

Tabela 10.2 – Exemplo de saída para a anotação dos descritores de relação.

Foram anotadas as relações expressas entre as entidades seguindo estas denominações: Organização (ORG), Pessoa (PES) e Local (LOCAL). As seguintes relações foram verificadas: ORG-ORG; ORG-PES; ORG-LOCAL. Realizada a tarefa de anotação das relações, uma comparação dos dados foi estabelecida no intuito de verificar quaisquer discordâncias entre nós. Dessa forma, as anotações que continham diferenças no descritor ou nas entidades foram analisadas em conjunto para que chegássemos a um acordo final quanto ao tipo de relação buscada.

Destacamos que, nessa coleção dourada do HAREM, as entidades ORG-PES-LOCAL estão previamente anotadas, e essas anotações não foram questionadas em nossa avaliação. Por se tratar de uma coleção dourada, assume-se a acurácia dos dados. No entanto, verificamos algumas inconsistências. Anotações divergentes e inadequadas dos corpora dificultaram, em grande medida, a classificação pelo sistema: por exemplo, determinado local estava marcado como ‘acontecimento’, o que levava à exclusão dessa entidade na análise de sua relação com outra.