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

O RECONHECIMENTO DE ENTIDADES NOMEADAS POR MEIO DE *CONDITIONAL RANDOM FIELDS* PARA A LÍNGUA PORTUGUESA

DANIELA OLIVEIRA FERREIRA DO AMARAL

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof[a]. Dr[a] Renata Vieira

Porto Alegre
2013

FICHA CATALOGRÁFICA EMITIDA PELA BIBLIOTECA

TERMO DE APRESENTAÇÃO DA DISSERTAÇÃO
EMITIDO E ASSINADO PELA FACULDADE

DEDICATÓRIA

Ao meu esposo, ANTONIO,
e ao nosso filho, JOÃO PEDRO,
por representarem a força do amor.

O que vale na vida
não é ponto de partida
e sim a caminhada.
Caminhando e semeando,
no fim terás o que colher.
(Cora Coralina)

AGRADECIMENTOS

Devo agradecer, primeiramente, a Deus e aos meus pais, Danilo e Clenir, pela vida, pela educação e por terem me ensinado o quanto o estudo com dedicação é fundamental na vida de qualquer ser humano.

Ao meu amado esposo Antonio, meu eterno muito obrigada pelo constante apoio, carinho, companheirismo, compreensão, entre outros tantos adoráveis sentimentos sempre dispensados a mim durante toda a minha jornada acadêmica. *Amor, és o exemplo de seriedade, responsabilidade e dedicação, o qual eu procuro seguir*.

Agradeço ao maior presente que recebi, meu filho João Pedro, nascido durante o período do mestrado. *És o ser mais precioso do mundo e fonte de nossa inspiração*.

À minha irmã querida Fabrize, por toda a preocupação e apoio nunca negados a mim. *Maninha, tu és o máximo*.

À Fernanda Figueira, amiga fundamental em minha vida, por me apresentar essa renomada e sólida instituição de ensino, a Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS), em especial o programa de Pós-graduação em Ciência da Computação da Faculdade de Informática.

À minha orientadora, Renata, agradeço por acreditar no meu potencial, pela confiança depositada em mim, paciência e por todos os seus ensinamentos. *Professora Renata, és um ser humano admirável e uma profissional de grande importância no nosso meio acadêmico. Eu a admiro muito*.

Às minhas colegas e amigas que cursaram algumas disciplinas durante o primeiro ano de mestrado, Raquel, Josiane e Valéria, agradeço pela cumplicidade, amizade e pelo conhecimento compartilhado em nossa área.

Aos meus colegas do laboratório de Processamento da Linguagem Natual (PLN), em especial à Sandrinha e ao Marlo por toda a orientação e amizade. Também não posso deixar de agradecer aos queridos Lucas Hilgert, Clarissa, Roger, Larissa, Aline, Evandro. Todos vocês são talentosíssimos.

Ao meu bolsista de Iniciação Científica, Lucas Pugens, pelo profissionalismo e pela responsabilidade na realização de relevantes tarefas em conjunto para o desenvolvimento desse trabalho.

O meu eterno muito obrigada à Dona Vera, minha amada sogra e amiga sincera, por ter atitudes de uma mãe dedicada e preocupada para comigo. *Obrigada por ser uma mãe para mim.*

À minha admirável amiga de todas as horas, a melhor que alguém poderia ter, Adriana, pelas palavras de incentivo, motivação e carinho durante horas de diálogo. *Dri, aprendi e aprendo sempre contigo. És para sempre a minha amiga do coração*.

Por fim, a todos aqueles que, de alguma forma, contribuíram para a realização deste importante objetivo concretizado: a conclusão do meu mestrado.

O RECONHECIMENTO DE ENTIDADES NOMEADAS POR MEIO DE *CONDITIONAL RANDOM FIELDS* PARA A LÍNGUA PORTUGUESA

RESUMO

Muitas tarefas de Processamento da Linguagem Natural envolvem a previsão de um grande número de variáveis, as quais dependem umas das outras. Métodos de predição estruturada são, essencialmente, uma combinação de classificação e de modelagem baseada em grafo. Eles unem a competência dos métodos de classificação com a capacidade desse tipo de modelagem de reproduzir, compactamente, dados multivariados. Os métodos de classificação realizam a predição usando um grande conjunto de *features* como entrada. *Conditional Random Fields* (CRF) é um método probabilístico de predição estruturada e tem sido amplamente aplicado em diversas áreas, tais como processamento da linguagem natural, incluindo o Reconhecimento de Entidades Nomeadas (REN), visão computacional e bioinformática. Sendo assim, neste trabalho é proposta a aplicação do CRF para o REN em textos da Língua Portuguesa e, sequencialmente, avaliar o seu desempenho com base no corpus do HAREM. Finalmente, testes comparativos da abordagem determinada versus a similar da literatura foram realizados, ilustrando a competitividade e eficácia do sistema proposto.

Palavras-chave: Reconhecimento de Entidades Nomeadas, *Conditional Random Fields*, Processamento da Linguagem Natural, Língua Portuguesa.

RECOGNITION OF ENTITIES NAMED BY CONDITIONAL RANDOM FIELDS TO PORTUGUESE LANGUAGE

ABSTRACT

Many tasks in Natural Language Processing involves the provision of a large number of variables, which depend on each other. Structured prediction methods are essentially a combination of classification and modeling based on graphs. They combine the power of classification methods with the ability of this type of modeling to play compactly, multivariate data. The classification methods perform prediction using a large set of features as input. Conditional Random Fields (CRF) is a probabilistic method for predicting structured and has been widely applied in various areas such as natural language processing, including the Named Entity Recognition (NER), computer vision, and bioinformatics. Therefore, this dissertation proposes the application of CRF to NER for the Portuguese Language and to evaluate their performance based on the HAREM corpus. Finally, comparative tests of similar approaches were performed, illustrating the efficiency and competitiveness of the proposed system.

Keywords: Named Entity Recognition, Conditional Random Fields, Natural Language Processing, Portuguese Language.

LISTA DE FIGURAS

Figura 2.1: Conditional Random Field de cadeia linear 33
Figura 3.1: Exemplo do texto segmentado e etiquetado..... 54
Figura 3.2: Procedimentos na etapa de treino..... 54
Figura 3.3: Exemplo da CD do Segundo HAREM com a aplicação da notação BILOU..... 55
Figura 3.4: Exemplo de um vetor com POS e com a notação BILOU..... 56
Figura 3.5: Exemplo de duas *features* aplicadas no sistema proposto. 57
Figura 3.6: Procedimentos na etapa de teste. 58
Figura 3.7: Exemplo de uma sentença segmentada e etiquetada. 59
Figura 3.8: Desenvolvimento do sistema NERP-CRF, etapa de Treino..... 60
Figura 3.9: Desenvolvimento do sistema NERP-CRF, etapa de Teste. 60
Figura 3.10: Vetor contendo POS taggin, BILOU e a categorização das EN. 62
Figura 3.11: Vetor contendo as features para o segmento de sentença 62
Figura 3.12: Vetor que contém cada palavra do texto etiquetada pelo POS *tagging*. 63
Figura 3.13: Vetor de saída que classifica o texto com o BILOU e com as categorias. 63
Figura 3.14: As *features* destacadas receberam o valor “null”..... 64
Figura 4.1: NERP-CRF comparado com os sistemas apresentados para o ‘Teste 2’. 78
Figura 4.2: NERP-CRF comparado com os sistemas apresentados para o ‘Teste 3’. 79

LISTA DE TABELAS

Tabela 2.1: Categorias e tipos definidos conforme o Segundo HAREM. 39
Tabela 4.1: Identificação das EN por meio da notação BIO. 72
Tabela 4.2: Classificação das EN usando a notação BIO. 73
Tabela 4.3: Identificação das EN por meio da notação BILOU. 73
Tabela 4.4: Classificação das EN usando a notação BILOU. 73
Tabela 4.5: Identificação das EN no ‘Teste 2’. 74
Tabela 4.7: Identificação das EN no ‘Teste 3’. 76
Tabela 4.8: Classificação das EN no ‘Teste 3’. 77

LISTA DE SIGLAS

CRF – *Conditional Random Fields*

HAREM – Avaliação de Sistemas de Reconhecimento de Entidades Nomeadas

EI – Extração de Informação

MEMM - *Maximum Entropy Markov Model*

HMM – *Hidden Markov Model*

REM – Reconhecimento de Entidades Mencionadas

REN – Reconhecimento de Entidades Nomeadas

EM – Entidades Mencionadas

EN – Entidades Nomeadas

ReRelEM – Reconhecimento de Relações entre Entidades Mencionadas

CD – Coleção Dourada

PLN – Processamento de Linguagem Natural

REX – *Rosette Entity Extractor*

MUC-6 – *Sixth Message Understanding Conference*

RENC – *Classification and Recognition of Named Entities*

ICML – *International Conference on Machine Learning*

TR – *Template Relation*

ACE – *Automatic Content Extraction*

EDT – *Entity Detection and Tracking*

GPE – *Geographical-Political Entity*

POS – *Part-of-Speech*

NERP-CRF – *Named Entity Recognition Portuguese - Conditional Random Fields*

BILOU – Begin Inside Last Outside Unit

CoNLL – *Conference on Natural Language Learning*

BioCreAtIvE - *Critical Assessment of Information Extraction systems in Biology*

SUMÁRIO

1. INTRODUÇÃO 25
1.1 Motivação e Objetivos 27
1.2 Organização do Trabalho 28

2. FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS 29
2.1 Conditional Random Fields 30
2.2 Reconhecimento de Entidades Nomeadas 34
2.3 Conferências de Avaliação Conjunta 35
2.3.1 MUC 35
2.3.2 ACE 36
2.3.3 HAREM 37
2.4 Sistemas de Reconhecimento de Entidades Nomeadas para a Língua Portuguesa 40
2.4.1 Sistema Priberam 40
2.4.2 Sistema R3M 41
2.4.3 Sistema REMBRANDT 43
2.4.4 Sistema SEI-Geo no Segundo HAREM 44
2.4.5 Sistema CaGE 44
2.4.6 Comparação entre os sistemas 46
2.5 Reconhecimento de Entidades Nomeadas aplicando CRF 46

3. NERP-CRF 53
3.1 Modelagem do Sistema 53
3.2 Implementação 58
3.3 Avaliação 65
3.3.1 Metodologia de Avaliação 65
3.3.1.1 Cross-validation 66
3.3.1.2 SAHARA 67
3.3.2 Medidas de Avaliação 68
3.3.3 Processo de Avaliação 69

4. RESULTADOS 72
4.1 ‘Teste 1’ 72
4.2 ‘Teste 2’ 74
4.3 ‘Teste 3’ 76
4.4 Comparação com outros Sistemas 78

4.5 Análise de Erros 79

5. CONSIDERAÇÕES FINAIS 82
5.1 Conclusões 82
5.2 Contribuições Científicas 82
5.3 Trabalhos Futuros 83

REFERÊNCIAS BIBLIOGRÁFICAS 86

APÊNDICES 92

1. INTRODUÇÃO

A Extração da Informação (EI) é uma importante tarefa na mineração de texto e tem sido amplamente estudada em várias áreas de pesquisa, incluindo o processamento da linguagem natural, recuperação de informação e mineração de dados na Web. O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa primordial na área de EI, juntamente com a extração de relação entre Entidades Nomeadas (EN) [JIN12]. Segundo Nadeau [NAD07], os termos que apresentam um ou mais designadores rígidos, num determinado texto, por exemplo, substantivos próprios, tais como nomes de pessoas, organizações e entidades locais definem as EN.

Dentro desse contexto, o REN em textos tem sido amplamente estudado por meio de métodos como aprendizagem supervisionada para classificar entidades do tipo pessoa, lugar e organização em textos ou, ainda, doenças e genes nos resumos das áreas médicas e biológicas [CHI94].

Existem vários sistemas comerciais de REN [WHI08], tais como *AeroText, Rosette Entity Extractor (REX), ClearForest, Inxight, PalyAnalyst,* e *SRA NetOwl.* Tais sistemas utilizam um número significativo de regras *hand-coded*, que permitem obter um desempenho limitado, apenas aplicado a alguns de tipos de entidades sobre corpora de domínio restrito, como por exemplo, textos de notícias. Esses métodos dependem de recursos caros e extensos para a etiquetagem manual, a qual realiza a identificação das entidades.

Dentro da tarefa do REN, constata-se que a necessidade de segmentar e rotular sequências surge por diferentes problemas em vários campos da ciência. Os modelos de Markov e as gramáticas estocásticas são largamente utilizados e baseiam-se em modelos probabilísticos para resolver tais problemas. Em biologia computacional, por exemplo, esses modelos têm sido aplicados com sucesso para alinhar sequências biológicas, onde sequências homólogas, para uma conhecida família evolutiva, são identificadas e, posteriormente, é realizada uma análise da estrutura secundária de RNA [DUR98]. Em Linguística Computacional e em Ciência da Computação, os Modelos de Markov e as gramáticas estocásticas têm sido aplicados para uma ampla variedade de problemas em processamento do discurso e do texto, incluindo segmentação tópica, etiquetagem, extração de informação e desambiguidade sintática [MAN08].

Os Modelos de Markov de Máxima Entropia (MEMMs) são modelos de uma sequência probabilística condicional que atingem todas as vantagens acima [MCC00].

Nos MEMMs, cada estado inicial tem um modelo exponencial que captura as características de observação como entrada e as saídas, uma distribuição sobre os próximos estados possíveis. Estes modelos exponenciais são treinados por um método apropriado de dimensionamento iterativo no framework de máxima entropia. Resultados experimentais publicados, anteriormente, mostram que o MMEMs aumentam a abrangência e duplicam a precisão relativa para os Modelos de Markov Ocultos em tarefas de segmentação.

MEMMs e outros modelos de estados finitos não geradores, baseados em classificadores do próximo estado (*next-state classifiers*), tais como os modelos de Markov discriminativos, [BOT91], compartilham uma fraqueza conhecida como o problema do viés dos rótulos (*label bias problem*): as transições que deixam um estado competem apenas umas contra as outras, ao invés de competir contra todas as outras transições do modelo. Em termos probabilísticos, as pontuações das transições são probabilidades condicionais dos próximos estados possíveis, dado um estado corrente e a sequência de observação. Esta normalização por estado de pontuação de transição implica uma “conservação da pontuação em conjunto” [BOT91] através do qual todo o conjunto que chega a um estado deve ser distribuído entre os estados sucessores possíveis.

Neste contexto, surge o modelo denominado Conditional Random Fields (CRF), um *framework* de modelagem de sequência de dados que tem todas as vantagens dos MEMM, mas também resolve o problema do viés dos rótulos de uma maneira fundamentada. A diferença crítica entre CRF e MMEM é que o MMEM utiliza modelos exponenciais por estados para as probabilidades condicionais dos próximos estados, dado o estado atual. Já o CRF tem um modelo exponencial único para uma probabilidade conjunta de uma sequência de entrada de rótulos, dado uma sequência de observação. Portanto, as influências das diferentes características em estados distintos podem ser tratadas independentemente umas das outras [LAF01].

O CRF pode também ser entendido como um modelo de estado finito com probabilidades de transição não normalizadas. Além disso, o CRF especifica uma distribuição probabilística bem definida sobre os possíveis rótulos, preparado por uma máxima verossimilhança. CRF generaliza, facilmente, para as semelhanças das gramáticas estocásticas livres de contexto, que podem ser úteis em problemas tais como predição de estrutura secundária de RNA e o processamento de linguagem natural.

1.1 Motivação e Objetivos

O termo chamado "entidade", hoje amplamente utilizado em Processamento de Linguagem Natural, foi cunhado para a *Sixth Message Understanding Conference* (MUC-6) [GRI96]. Naquele momento, a MUC estava focada nas tarefas de Extração de Informação (IE), nas quais informações sobre atividades de empresas e relacionadas à defesa são extraídas a partir de textos não estruturados, tais como artigos de jornal.

Kripke [KRI81] afirma que a palavra “Nome”, na expressão Entidade Nomeada, tem como objetivo restringir a tarefa de que somente essas entidades, as quais podem estar relacionadas a um ou vários designadores rígidos, representam o referente. Por exemplo, a empresa automotiva criada por Henry Ford em 1903 é referida como *Ford* ou *Ford Motor Company*.

Na definição da tarefa, verifica-se a importância de se reconhecer as unidades de informação, expressões numéricas e expressões de porcentagem. Como exemplo de unidades de informação destacam-se os nomes de pessoas, organizações e nomes de locais; já as expressões numéricas são do tipo data, hora e dinheiro. Identificar as referências a estas entidades, no texto, foi reconhecido como uma das subtarefas importantes de Extração da Informação e foi chamada de *Classification and Recognition of Named Entities* (RENC) [NAD07].

Lafferty *et al.* [LAF01] determinaram um modelo matemático probabilístico que descreve dois procedimentos de treino aplicando CRF. O modelo apresenta o algoritmo de estimativa de parâmetros iterativos para CRF e, posteriormente, este é comparado com o desempenho dos modelos resultantes para os Modelos Ocultos de Markov e os Modelos de Markov de Máxima Entropia em dados sintéticos e em linguagem natural.

Os autores desse artigo expõem resultados experimentais e dados sintetizados mostrando que o CRF resolve a versão clássica do problema de viés do rótulo. Mais significativamente, eles demonstraram que o desempenho do CRF é melhor que o desempenho dos modelos HMM e MMEM quando a distribuição dos dados tem dependências de ordem superior ao modelo. Esses dados são confirmados através de resultados satisfatórios e, as vantagens reavidas dos modelos condicionais, a partir de avaliações dos HMM, MEMM e CRF com estrutura de estado idêntica às tarefas de etiquetar parte do discurso.

Os resultados da Conferência Internacional de Aprendizado de Máquina (*International Conference on Machine Learning* - ICML) no ano de 2001 [LAF01], seguido

de outros trabalhos sobre *Conditional Random Fields* [SUA11], [LI11], [LIN12], indicam que o algoritmo de CRF apresenta um dos melhores desempenhos para o REN.

Com isto, esta dissertação tem como motivação o fato de: (i) o REN ter sido pouco explorado utilizando o método de aprendizagem supervisionada CRF para a língua portuguesa; (ii) não existir proposta de REN aplicando o CRF para identificar as EN e classificá-las de acordo com as dez categorias dos textos da conferência do HAREM. O corpus do HAREM é considerado a principal referência na área de PLN, e caracteriza-se por ter um conjunto de textos anotados e validados por humanos (Coleção Dourada), o que facilita a avaliação do método em estudo; e (iii) o método de CRF pode ajudar a identificar um maior número de EN, o que poderá ser verificado por meio da comparação com outros sistemas.

O objetivo geral do trabalho é aplicar CRF para a tarefa de REN em corpus da língua portuguesa e avaliar comparativamente com outros sistemas que realizam REN, tendo como base o corpus do HAREM.

Para que o objetivo geral seja alcançado, a seguir serão apresentados os objetivos específicos:

- Aprofundar o estudo teórico sobre CRF e sobre REN;
- Verificar quais são as outras técnicas citadas na literatura para o REN;
- Realizar um estudo sobre as *features* referenciadas em trabalhos literários para a área de extração da informação, a fim de gerar um modelo de CRF;
- Aplicar o modelo de CRF gerado;
- Avaliar a técnica de CRF para o REN;
- Comparar os resultados obtidos após aplicar CRF, no corpus do HAREM, com outros métodos existentes, na literatura, que realizam REN

1.2 Organização do Trabalho

Apresentaremos, no Capítulo 2, uma revisão dos trabalhos relacionados à pesquisa proposta e a aplicabilidade do CRF no reconhecimento de Entidades Nomeadas. O Capítulo 3, trata do desenvolvimento do sistema NERP-CRF, sua descrição, a modelagem do sistema, implementação e o processo de avaliação. Os resultados obtidos com a avaliação dos testes bem como a análise de erros serão descritas no Capítulo 4. Por fim, no Capítulo 5, serão apresentadas as considerações finais e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA E TRABALHOS RELACIONADOS

Extração de informações (EI) é a tarefa de encontrar informações estruturadas a partir de textos não estruturados ou semi-estruturados. Já a tarefa de REN consiste em identificar entidades nomeadas, na sua maioria nomes próprios, a partir de textos de forma livre e classificá-las dentro de um conjunto de tipos de categorias pré-definidas, tais como pessoa, organização e localização. O REN em textos que abordam os mais variados domínios, além da extração de relações entre EN, é uma das tarefas primordiais dentro da área de EI.

Inicialmente, soluções para o REN dependem de padrões de regras trabalhadas manualmente, pois exigem experiência humana além de um trabalho intenso para a criação de tais padrões [APP99]. Adicionalmente, a criação de sistemas tem o objetivo de aprender automaticamente esses padrões de dados etiquetados [CIR01]. Os Modelos de Markov Ocultos [BIK97], modelos de Máxima Entropia [CHI03] e modelos de Markov de Máxima Entropia [BEN03] [CUR03] e [FIN05] são formalismos estatísticos de aprendizagem de máquina que realizam o REN. Outros trabalhos sobre REN utilizam modelos matemáticos probabilísticos, denominados CRF [SET04], [SUT10], [LAF01] e [CHA12], o qual é o formalismo desenvolvido nesta dissertação.

Neste capítulo serão apresentados os assuntos que se relacionam com o tema proposto. Dentro da pesquisa realizada, destacam-se fontes bastante referenciadas na literatura, além de outras atuais, descritas a seguir.

[SUT10], [LAF01] e [CHA12] apresentam um *framework* para a construção de modelos probabilísticos para segmentação e etiquetagem de dados sequenciais baseados em *Conditional Random Fields* (CRF).

O trabalho de [SUX08] descreve uma nova abordagem do reconhecimento de Entidades Nomeadas Chinesas baseado em Conditional Random Fields. Na abordagem proposta, a estrutura do modelo foi desenhada com a forma de cascata, e o resultado é passado para um modelo principal onde se aplicará, a partir desse modelo, o reconhecimento de entidades como, por exemplo, nomes de pessoas e de organizações.

Ratinov e Roth [RAT 09] investigaram a aplicação do Reconhecimento de Entidades Nomeadas a partir da necessidade de usar o conhecimento prévio e decisões não locais para a identificação de tais entidades nomeadas em um texto.

O artigo em [FRE10] apresenta a segunda edição de uma conferência que avalia sistemas, os quais aplicam REN para o português, o Segundo HAREM. Especificamente,

este trabalho aborda a trilha ReRelEN, que trata a detecção de relações semânticas entre Entidades Nomeadas.

2.1 Conditional Random Fields

A tarefa de atribuir uma sequência de rótulos para um grupo de sequências de observação surge em diferentes áreas, incluindo a bioinformática, linguística computacional e reconhecimento da fala. Por exemplo, considere a tarefa de processamento da linguagem natural de rotular as palavras constituintes de uma sentença. Nesta tarefa, cada palavra é marcada com um rótulo que indica a sua etiquetagem morfológica adequada, como por exemplo, a indicação se a palavra em foco é um artigo ou uma preposição, resultando assim em um texto anotado.

Um dos métodos mais comuns para a realização de tais tarefas de etiquetagem e de segmentação é a de empregar os Modelos de Markov Ocultos (HMM) ou o estado finito automático e probabilístico para identificar a maioria das sequências de rótulos nas palavras, mais facilmente, dada uma sentença. Os HMM são uma forma de modelos generativos, que definem um conjunto de distribuição probabilística p(X,Y) onde X e Y são variáveis aleatórias, respectivamente, classificando uma sequência de observação e suas sequências de rótulos correspondentes. A fim de definir uma distribuição conjunta desta natureza, os modelos geradores devem enumerar todas as possíveis sequências de observação. Esta é uma tarefa que, para a maioria dos domínios, é intratável, a menos que os elementos de observação sejam representados como unidades isoladas, independente de outros elementos numa sequência de observação. Mais precisamente, o elemento de observação, em algum dado instante, só pode diretamente depender do estado, ou rótulo, naquele momento. Isto é um pressuposto necessário para um conjunto de dados um pouco simples, contudo a maioria das sequências de observação de palavras é melhor representada por várias características interagindo e pela longa distância de dependência entre os elementos de observação.

Esta é uma questão de representação dentre a maioria dos problemas fundamentais quando se rotula dados sequenciais. Um modelo que suporte inferência tratável é necessário, no entanto, um modelo que represente os dados sem fazer suposições de independência injustificáveis também é desejável. Uma maneira de satisfazer ambos os critérios é utilizar um modelo que defina uma probabilidade condicional p(Y|x) sobre uma sequência de rótulos, dada uma sequência de observação particular x, ao invés de uma distribuição conjunta sobre o rótulo e as sequências de

observação. Os modelos condicionais são usados para etiquetar uma nova sequência de observação x, selecionando a sequência de rótulo y que aumente a probabilidade condicional p(y|x). A natureza condicional de tais modelos significa que nenhum esforço é desperdiçado em modelar as observações, e é livre de ter que fazer suposições de independências injustificadas sobre essas sequências. Arbitrariamente, atributos de dados de observação podem ser capturados pelo modelo, sem o modelador ter que preocupar-se sobre como esses atributos são relatados.

Conditional Random Fields (CRF), segundo Lafferty *et al.* em [LAF01], é um modelo matemático probabilístico que tem o objetivo de etiquetar e segmentar dados sequenciais, baseados numa abordagem condicional descrita no parágrafo anterior. O CRF é uma forma de modelo gráfico não direcionado que define uma única distribuição logaritmicamente linear sobre sequências de rótulos, dada uma sequência de observação particular. A vantagem primária dos modelos de CRF sobre os modelos de Markov Ocultos é a sua natureza condicional, pois resulta no abrandamento de pressupostos independentes, necessários para os modelos HMM, a fim de assegurar uma inferência tratável. Adicionalmente, os modelos de CRF evitam o problema de viés do rótulo, uma fraqueza exibida pelos Modelos de Markov de Máxima Entropia [MAC00] e outros modelos de Markov condicionais baseados em modelos gráficos direcionados. O CRFs supera ambos os modelos MEMM e HMM em número de tarefas de etiquetagem dada uma sequência de palavras [MAC00, PIN03, SHA03].

Em [LAF01], os autores definiram X como sendo uma variável aleatória sobre uma sequência de dados para serem etiquetados, Y como uma variável aleatória sobre uma sequência de etiquetas correspondentes. As sequências X e Y podem ser representadas da seguinte forma respectivamente: $X = (X_1, X_2,..., X_n)$ e $Y = (Y_1, Y_2,..., Y_n)$. Todos os Y_i componentes de Y são assumidos para variar ao longo de um alfabeto Y de rótulos finitos. Por exemplo, X pode variar mais sobre sentenças de linguagem natural e Y variar sobre os rótulos de parte do discurso daquelas sentenças, sendo Y o conjunto de possíveis rótulos de parte do discurso. As variáveis aleatórias X e Y são distribuídas conjuntamente, mas em um quadro discriminativo, foi construído um modelo condicional p(Y|X) de observações pariadas e de sequências de rótulos.

Em função das condições acima, surge a seguinte definição: seja G = (V, E) um grafo tal que $Y = (Y_v)_{v \in V}$ de maneira que Y é indexado para os vértices de G. Então (X,Y) é um *conditional random field,* em casos, nos quais, condicionadas sobre X, as variáveis aleatórias Y_v obedecerem à propriedade de Markov com relação ao grafo:

$p(\mathbf{Y}_v \mid \mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}_v \mid \mathbf{X}, \mathbf{Y}_w, w \sim v)$, onde w ~ v significa que w e v são vizinhos em G. Deste modo, o CRF é um campo aleatório, completamente, condicionado sobre a observação X.

Se o grafo G = (V, E) de Y é uma árvore (dos quais uma cadeia é o exemplo mais simples), seus subgrafos de G são as arestas e os vértices. Portanto, pelo teorema fundamental dos campos aleatórios, a distribuição conjunta sobre a sequência de rótulo Y dado a X tem a forma

$$p_\theta(\mathbf{y} \mid \mathbf{x}) \propto \qquad (1)$$
$$\exp\left(\sum_{e \in E,k} \lambda_k \, f_k(e, \mathbf{y}|_e, \mathbf{x}) + \sum_{v \in V,k} \mu_k \, g_k(v, \mathbf{y}|_v, \mathbf{x})\right),$$

onde x é uma sequência de dados, y uma sequência de rótulos e y|s é o conjunto de componentes de y associado com os vértices em um subgrafo S.

Assume-se que as *features* f_k e g_k são dadas e fixadas. Por exemplo, uma *feature* de vértice Booleano g_k pode ser verdadeira se a palavra X_i é uma letra maiúscula e a *tag* Y_i é um nome próprio.

Sejam os seguintes parâmetros representados por $\theta = (\lambda_1, \lambda_2, \ldots; \mu_1, \mu_2, \ldots)$ para os dados de treino $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^N$ com distribuição empírica $\tilde{p}(\mathbf{x}, \mathbf{y})$, então a função Objetiva $\mathcal{O}(\theta)$ de verossimilhança logarítmica é:

$$\begin{aligned} \mathcal{O}(\theta) &= \sum_{i=1}^{N} \log p_\theta(\mathbf{y}^{(i)} \mid \mathbf{x}^{(i)}) \\ &\propto \sum_{\mathbf{x},\mathbf{y}} \tilde{p}(\mathbf{x}, \mathbf{y}) \, \log p_\theta(\mathbf{y} \mid \mathbf{x}) \, . \end{aligned}$$

Embora isso englobe modelos semelhantes ao HMM, a classe dos *Conditional Random Fields* é muito mais expressiva, porque permite dependências arbitrárias sobre a sequência de observação. Além disso, as características não precisam especificar completamente um estado ou uma observação. Desse modo, espera-se que o modelo possa ser estimado a partir de menos dados de treino.

Pode-se assumir, neste caso, que as dependências de Y, condicionadas sobre X, formam uma cadeia. Para uma estrutura em cadeia, a probabilidade condicional de uma sequência de rótulos pode ser expressa, concisamente, em forma de matriz. Suponha que $p_\theta(\mathbf{Y} \mid \mathbf{X})$ é um CRF dado por (1). Para cada posição i numa sequência x de observação, é definida a variável aleatória da matriz |*Y*|x|*Y*| por $M_i(\mathbf{x}) = [M_i(y', y \mid \mathbf{x})]$ através da fórmula:

$$
\begin{aligned}
M_i(y', y \,|\, \mathrm{x}) &= \exp\left(\Lambda_i(y', y \,|\, \mathrm{x})\right) \\
\Lambda_i(y', y \,|\, \mathrm{x}) &= \textstyle\sum_k \lambda_k \, f_k(e_i, \mathbf{Y}|_{e_i} = (y', y), \mathrm{x}) + \\
&\quad \textstyle\sum_k \mu_k \, g_k(v_i, \mathbf{Y}|_{v_i} = y, \mathrm{x}) \,,
\end{aligned}
$$

onde e_i é a aresta com os rótulos $(\mathbf{Y}_{i-1}, \mathbf{Y}_i)$ e v_i é o vértice com rótulo Y_i. Em contraste com os modelos gerativos, os modelos condicionais como os CRF não necessitam enumerar sobre todas as sequências x de observações possíveis e, por conseguinte, essas matrizes podem ser calculadas diretamente, a partir de um ou de vários dados de treino ou de uma sequência x de observação de teste e de um vetor de parâmetro, que pode ser chamado de θ. Então a normalização, função de partição, $Z_\theta(\mathrm{x})$ é a entrada $(\mathsf{start}, \mathsf{stop})$ do produto dessas matrizes:

$$
Z_\theta(\mathrm{x}) \;=\; (M_1(\mathrm{x})\, M_2(\mathrm{x}) \cdots M_{n+1}(\mathrm{x}))_{\mathsf{start},\mathsf{stop}} \;. \quad (2)
$$

Ao simplificar algumas expressões, adicionam-se os estados, inicial e final, representados por: $\mathbf{Y}_0 = \mathsf{start}$ and $\mathbf{Y}_{n+1} = \mathsf{stop}$.

Usando a função dada em (2), a probabilidade condicional de uma sequência de rótulo y é escrita conforme a notação abaixo:

$$
p_\theta(\mathrm{y} \,|\, \mathrm{x}) \;=\; \frac{\prod_{i=1}^{n+1} M_i(\mathrm{y}_{i-1}, \mathrm{y}_i \,|\, \mathrm{x})}{\left(\prod_{i=1}^{n+1} M_i(\mathrm{x})\right)_{\mathsf{start},\mathsf{stop}}} \,,
$$

As dependências de Y condicionadas sobre X formam uma cadeia linear, conforme a Figura 2.1. Assim, para as formulações de cadeia linear de CRF convencional, uma cadeia de Markov de primeira ordem e unidimensional é assumida para representar as dependências entre as variáveis de etiquetas previstas, enquanto nenhuma dependência temporal é imposta entre as variáveis observadas.

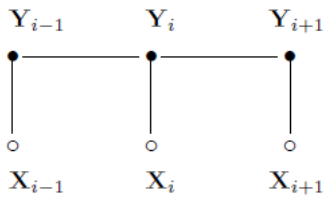

Figura 2.1: Conditional Random Field de cadeia linear: um nó aberto denota uma variável aleatória e um nó sombreado foi definido como o seu valor observado.

Fonte: Lafferty, John; McCallum, Andrew; Pereira, Fernando. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.* [LAF01]

Dando continuidade a pesquisa, tem-se como um dos propósitos deste estudo, aplicar CRF para o Reconhecimento de Entidades Nomeadas. Portanto, a próxima seção apresenta uma descrição pormenorizada sobre REN.

2.2 Reconhecimento de Entidades Nomeadas

Entidades Nomeadas (EN) compreendem-se como termos que apresentam um ou mais designadores rígidos, num determinado texto [NAD07]. Alguns dos tipos mais comuns de entidades são substantivos próprios, tais como nomes de pessoas, organizações e entidades locais; temporais como datas, tempo, dia, ano e mês; entidades numéricas, tais como medições, percentagens e valores monetários. Entidades de domínio numerosas e entidades de aplicações específicas não são consideradas como entidades nomeadas, como: peças e tipos de defeitos no setor de manufatura ou nomes de doenças e sintomas no setor da saúde [SUR09].

O Reconhecimento de Entidades Nomeadas (REN) define-se como uma tarefa cujo objetivo é identificar as entidades nomeadas bem como sua posterior classificação, atribuindo uma categoria semântica para essas entidades. Segundo Sureka *et al.* [SUR09], o Reconhecimento de Entidades Nomeadas e a posterior classificação de tais entidades é uma técnica amplamente utilizada no Processamento da Linguagem Natural (PLN) e consiste da identificação de nomes de entidades-chave presentes na forma livre de dados textuais. A entrada para o sistema de extração de EN é o texto de forma livre e a saída é um conjunto de chamadas anotações, ou seja, grupo de caracteres através de trechos do texto de entrada. A saída do sistema de extração de entidades nomeadas é, basicamente, uma representação estruturada a partir da entrada de um texto não estruturado. Em geral, o processo de REN é referido como um modelo pré-definido de enriquecimento, o qual é constituído de enchimentos e cargas. Os enchimentos do modelo representam os tipos de entidades e as cargas, os valores de *sub-string* do texto de entrada.

As três principais abordagens para extração de entidades nomeadas são: sistemas baseados em regras, sistemas baseado em aprendizado de máquina e abordagens híbridas. Sistemas baseados em regras, também conhecidos como sistemas baseados no conhecimento, consistem em definir heurísticas na forma de expressões regulares ou padrões linguísticos. Um exemplo de uma regra ou heurística pode ser a presença de palavras como “*Incorporated*”, “*Corporation*”, “*Limited*”, entre outros, indicando a presença de uma entidade do tipo Organização ou a heurística, cuja string pode ser o símbolo “@”

e terminar com um “.com”, ou “.org” ou “.edu” um endereço de e-mail. Sistema baseados em regras também fazem o uso de dicionários ou léxicos que contêm, comumente, a ocorrência de termos ou palavras “trigger”. Tais léxicos aumentam a precisão e o *recall* do sistema [NAD07] e [MAN08].

Conforme Chatzis e Demiris [CHA12], durante os últimos anos temos assistido a uma explosão de vantagens nos modelos de *Conditional Random Fields*, à medida que tais modelos conseguem alcançar uma previsão de desempenho excelente em uma variedade de cenários. Sendo assim, uma das abordagens de maior sucesso para o problema de predição de saída estruturada, com aplicações bem sucedidas, inclui o processamento de texto, a área da bioinformática e o processamento da linguagem natural.

A seguir serão apresentadas três conferências: o MUC, ACE e o HAREM, as quais tratam da avaliação conjunta de sistemas de Reconhecimento de Entidades Nomeadas.

2.3 Conferências de Avaliação Conjunta

A Avaliação Conjunta consiste de uma atividade na qual participam vários sistemas e tem como objetivo aprimorar o estado da arte da área, proporcionando pesquisas nas áreas julgadas necessárias, de acordo com a tarefa em questão. Tais sistemas são avaliados e comparados quando executam uma mesma tarefa e seus resultados são, principalmente, recursos de avaliação que serão reutilizados como testes em outras pesquisas [SAN07b].

As conferências destinadas à avaliação de sistemas inteligentes, demonstraram uma importante ajuda no avanço da área de Processamento da Linguagem Natural, pois envolvem tarefas distintas na compreensão da língua. Conferências que tratam tarefas de reconhecimento de entidades nomeadas e a identificação de relações entre estas entidades são apresentadas a seguir.

2.3.1 MUC

A conferência *Message Understanding Conference* (MUC) [MUC6] foi a primeira conferência que tratou a avaliação do Reconhecimento de Entidades Nomeadas. No ano de 1987 foi realizada a sua primeira edição e teve como objetivo o desenvolvimento de uma avaliação conjunta na área de Extração de Informação (IE).

Em 1995, ocorreu a sexta edição do MUC, onde teve início a avaliação do REN para a língua inglesa. Esta edição teve a sua peculiaridade em relação a outras edições, pois, as edições anteriores consideravam o Reconhecimento de Entidades Nomeadas

como sendo uma parte da tarefa de Extração da Informação. No *Message Understanding Conference* [MUC6a], o Reconhecimento de Entidades Nomeadas consistiu em anotar as entidades nomeadas em três tipos de categorias: Enamex, Timex e Numex, as quais são descritas a seguir.

1) A categoria Enamex [MUC7] é formada por nomes próprios definidos pelos tipos Pessoa, Organização e Local. Por exemplo, nomes de pessoa ou de família, organização empresarial, organização não governamental, nomes de locais politicamente ou geograficamente definidos, entre outros.

2) A categoria Timex é uma expressão de tempo dividida em Data e Hora. A data é uma expressão completa ou parcial na qual se refere ao ano, mês ou dia. O Time define-se por uma expressão referente ao tempo, como o horário, por exemplo.

3) O Numex é uma expressão numérica formada por expressões denominadas *Money* (expressão monetária) e *Percent* (representando a porcentagem).

A sétima edição do MUC criou a tarefa de identificação de relações entre as categorias, chamada de *Template Relation* (TR) [MEC7a]. Esta tarefa realiza a extração de fatos bem determinados em textos jornalísticos da língua inglesa. Ainda nesta versão do MUC, as relações envolvendo a categoria Organização foram determinadas como funcionário_de, produto_de e localização_de.

2.3.2 ACE

Automatic Content Extraction (ACE) foi a conferência que surgiu após o MUC-7. A ACE teve início em 1999 com um estudo piloto para a língua inglesa, cujo objetivo foi verificar quais tarefas de Extração de Informação seriam avaliadas [DOD04]. A ACE, no período de 2000 a 2001, realizou o Reconhecimento de Entidades Nomeadas por meio da identificação e da classificação de entidades e das expressões anafóricas. Tais expressões abrangeram, além de nomes próprios, descrições ou pronomes. Esse processo foi determinado para as línguas inglesa e chinesa e denominou-se *Entity Detection and Tracking* (EDT).

O sistema de reconhecimento de relações caracterizou os anos de 2002 a 2003 para o ACE, onde a referida tarefa foi chamada de *Recognition of relations*. Ainda em 2003, iniciou-se o tratamento de relações para a língua árabe e, na sequência, em 2004, houve o reconhecimento de eventos.

O EDT contemplou, além dos tipos de categorias mencionadas no MUC, os tipos *Facility* e *Geographical-Political Entity*[1](GPE). O primeiro tipo Facility expressa as categorias armas, veículos ou instalações, por exemplo, aeroporto. Já o segundo, GPE, representa a supercategoria de Organização e Local, por exemplo, país.

2.3.3 HAREM

A Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas, cuja sigla é denominada HAREM, é um evento de avaliação conjunta da língua portuguesa, com o objetivo de realizar a avaliação de sistemas reconhecedores de entidades mencionadas criadas pela Linguateca [SAN07b], [SAN09].

O HAREM utiliza o termo Entidade Mencionada [MOT07] para designar nomes próprios, os quais são referenciados em um texto. Salienta-se que a expressão Entidade Mencionada tem a mesma denominação que o termo, utilizado no ACE, designado por Entidade Nomeada (*Named Entities*).

Neste trabalho será adotada a nomenclatura Entidades Nomeadas (EN - *Named Entities*), como é proposto na conferência do ACE, o qual inclui na sua análise os substantivos comuns e sintagmas nominais relacionados aos nomes próprios identificados. Todavia, os termos EM e REN são utilizados aqui sem diferenciar EN ou EM, assim como REN de REM.

A metodologia do HAREM é formada por:

- especificar as tarefas que serão avaliadas,
- definir as diretivas de etiquetagem e
- estabelecer a criação das coleções de textos.

Entre os eventos do HAREM destacam-se: o primeiro HAREM decorrido no ano de 2004 e o Segundo HAREM, em 2008. A coleção do Primeiro HAREM [2], dentro de uma estimativa de seu tamanho, é formada por 466355 palavras, abrangendo os mais variados tipos de textos, destacando-se os: jornalísticos, literários, políticos, textos da web e textos transcritos de entrevistas. Já a Coleção Dourada do Primeiro HAREM compõe-se de 89241 palavras, sendo que dentro deste grupo houve o reconhecimento de 3851 entidades nomeadas.

---

[1] ***Assessment of Detection and Recognition of Entities and Relations Within and Across Documents.*** Automatic Content Extraction 2008. Evaluation Plan (ACE08). Ago, 2008.

[2] Disponível em http://www.linguateca.pt/

O primeiro HAREM apresenta dois aspectos fundamentais utilizados na avaliação de REN: 1) as tarefas de classificar e identificar uma expressão como entidade nomeada ligada ao seu uso no contexto, não estando dependentes, por exemplo, de dicionários, almanaques bem como ontologias; e 2) aceita-se atribuir mais de uma classificação a uma mesma entidade nomeada, caso o contexto, em que essa se encontra, não possibilite escolher uma delas somente.

A avaliação conjunta que o HAREM realiza é feita através da comparação do desempenho dos sistemas de vários grupos. Estes grupos realizam a referida avaliação utilizando um conjunto de recursos em comum e uma métrica estabelecida por meio de um consenso.

O evento do Segundo HAREM possui uma coleção composta por 1040 documentos, sendo que, dentro deste grupo, encontram-se 129 documentos constituintes da coleção Dourada (CD). Os documentos da coleção do Segundo HAREM foram selecionados respeitando as seguintes condições: 1) deveria conter igualmente, na coleção, o português de Portugal e o do Brasil; 2) nos documentos deveriam estar presentes distintos gêneros e registros textuais, 3) esta coleção deveria conter algum material já usado no Primeiro HAREM, a fim de que, posteriormente, fosse possível comparar a performance dos sistemas nesses documentos e em outras avaliações.

O Segundo HAREM manteve o modelo semântico do primeiro HAREM [SAN07a] assim como o modelo de avaliação [SAN07]. Esta segunda edição do HAREM, além de realizar uma avaliação mais justa dos sistemas, incluiu: a tarefa de reconhecer e normalizar expressões classificadas como Tempo e o reconhecimento de relações semânticas entre as entidades nomeadas, ou seja, a criação da pista de ReRelEN.

A coleção Dourada é um subconjunto da coleção do Segundo HAREM, sendo essa utilizada para tarefa de avaliação dos sistemas que tratam REN. Primeiramente, o mesmo conjunto de textos da CD foi anotado por duas anotadoras com o auxílio da ferramenta Etiquet(H)AREM. Posteriormente, as anotações foram comparadas, com o auxílio do programa Alinhador e discutidas pelas anotadoras, sendo que, em alguns casos, por toda a organização até que se chegasse a uma consensual anotação. Em outra etapa, as anotadoras analisaram diferentes textos da CD.

Após a conclusão do processo de anotação da CD como um todo, ocorreu a revisão dos textos de um modo geral de toda a CD e um revisão detalhada das EN por categoria, considerando sempre o contexto do qual faziam parte as Entidades Nomeadas.

A anotação e revisão da CD encontraram 7836 entidades nomeadas, repartidas nas várias categorias do HAREM. A categoria PESSOA foi a mais frequente na CD e sequencialmente, fizeram-se presentes as categorias LOCAL, TEMPO e ORGANIZAÇÃO.

Posteriormente, as EM sofreram uma identificação e classificação por todos os participantes do evento, onde esses obedeceram ao grupo de diretivas e usaram as categorias e os tipos conforme a Tabela 2.1 [MOT08].

Tabela 2.1: Categorias e tipos definidos conforme o Segundo HAREM.

Categoria	**Tipo**
Abstração	Disciplina, Estado, Ideia, Nome, Outro.
Acontecimento	Efemeridade, Evento, Organizado, Outro.
Coisa	Classe, MembroClasse, Objeto, Substancia, Outro.
Local	Físico, Humano, Virtual.
Obra	Arte, Plano, Reproduzida, Outro.
Organização	Administração, Empresa, Instituição, Outro.
Categoria	**Tipo**
Pessoa	Cargo, GrupoCargo, GrupoInd, GrupoMembro, Individual, Membro, Povo, Outro.
Tempo	Duração, Frequência, Genérico, TempoCalend, Outro.
Valor	Classificação, Moeda, Quantidade, Outro.
Outro	

As classificações categorizam cada uma das Entidades Nomeadas identificadas pelos sistemas. A finalidade da Coleção Dourada é avaliar os sistemas participantes por meio da comparação da Coleção Dourada original com as anotações produzidas pelos sistemas participantes. Os sistemas participantes do Segundo HAREM serão apresentados a seguir.

2.4 Sistemas de Reconhecimento de Entidades Nomeadas para a Língua Portuguesa

Nesta subseção é apresentada uma breve descrição de alguns sistemas que tratam sobre o Reconhecimento de Entidades Nomeadas para o português, bem como a extração de relações semânticas entre as ENs para esse mesmo idioma.

Os trabalhos para a língua portuguesa que realizam REN e fazem a identificação das relações entre as Entidades Nomeadas surgiram a partir do HAREM [MOT07]. Existe uma comunidade aplicada no Reconhecimento de Entidades Nomeadas e a tarefas ligadas ao português, relata Cristina Mota e colegas em [MOT07].

Dentre os sistemas que serão apresentados a seguir, apenas o sistema HENDRIX não participou do Segundo HAREM, o qual será abordado na seção 2.5. A descrição detalhada dos demais sistemas encontra-se no livro do Segundo HAREM [MOT08a].

2.4.1 Sistema Priberam

O sistema Priberam ao HAREM é baseado em um léxico com classificação morfossintática e semântica. Cada entrada do léxico, corresponde a uma ligação a um ou mais níveis de uma ontologia multilíngue [AMA04], podendo corresponder a um ou mais sentidos, os quais possuem diferentes valores morfológicos e semânticos.

Para a construção do sistema foram utilizadas regras contextuais [AMA04], as quais atribuem ou alteram valores morfológicos e semânticos a partes do texto isoladas ou a sequências de unidades. Tais regras contextuais realizam, por exemplo, a criação de: locuções por meio da combinação de sequências de palavras; categorias gramaticais e combinações de listas de palavras, chamadas de "*constantes*", formadas por categorias ou palavras únicas.

As regras para a tarefa de REN consideram as sequências de nomes próprios, separadas ou não por algumas preposições e o contexto em que as Entidades Mencionadas são encontradas. Por exemplo, uma EM "João Pedro", classificada como PESSOA, poderá ser classificada como ORGANIZAÇÃO se esta for precedida por uma expressão como "instituto*"*.

Fez-se necessário a criação de regras para a classificação de EM das categorias COISA, ABSTRAÇÃO, ACONTECIMENTO e OBRA. Já a classificação das categorias PESSOA, LOCAL, ORGANIZAÇÃO, VALOR, TEMPO já tinha sido tratada pelo sistema automático de perguntas e respostas antes da participação no Segundo HAREM.

Complementando as ferramentas necessárias para a construção do Priberam ao HAREM, os autores ainda criaram:

a) novas constantes para a classificação contextual das EM. Para tal, utilizou-se a ontologia desenvolvida pela Priberam, permitindo uma extração de nomes relacionados com os tipos e subtipos a serem implementados de uma maneira mais detalhada, e

b) um filtro que determinasse as correspondências entre as categorias e valores originais do sistema e os do HAREM. Este filtro consulta um ficheiro XML de fácil modificação para que, quando for preciso, o texto seja etiquetado com novas categorias e valores semânticos.

A Priberam cumpriu seus objetivos para conferência do Segundo HAREM, uma vez que tratou da identificação e da classificação das Entidades Nomeadas, quer a nível de correção sintática, quer a nível de sistemas de perguntas e respostas ou ainda para motores de busca. Em função da afirmação anterior, constatou-se que os resultados foram animadores, pois o sistema Priberam identificou corretamente 72,29% das Entidades Mencionadas, considerando como referência a Coleção Dourada do Segundo HAREM.

2.4.2 Sistema R3M

O sistema R3M realiza o REN para as categorias pessoas, organizações e locais. A opção por essas três categorias deve-se ao fato de que essas, de uma forma geral, têm sido estudadas mais amplamente dentro da área de extração da informação e porque os desenvolvedores do R3M não tiveram disponibilidade de dedicar mais tempo a esse sistema. Mesmo assim, o R3M foi projetado de modo que permita estender-se ao reconhecimento de outras categorias, assim como incluir o reconhecimento de relações de EM. Esse sistema é uma reimplementação do sistema criado por Mota [MOT08h], apresentando várias melhorias.

O R3M aplica aprendizagem semi-supervisionada, utilizando um algoritmo de co-training para inferir regras de classificação [COL99]. A escolha do algoritmo de *co-training* deve-se ao fato de que este tem grande probabilidade de obter bons resultados de

classificação que se aproximam dos 80% de a*ccuracy,* usando um número muito reduzido de exemplos previamente anotados.

Principais características do R3M:

- sistema modular sequencial, separado em duas fases: fase de identificação de entidades mencionadas e de classificação;

- etapa de treino a fim de aprender regras de classificação com base num algoritmo de co-treino;

- etapa de teste que usa as regras aprendidas para classificar entidades em novos textos, produzindo um texto final anotado. Além disso, as duas fases acima possuem módulos de identificação de entidades, contextos e extração de *features*.

O módulo de identificação tem a função de reconhecer candidatos a entidades e o contexto em que este se encontra em textos não anotados, tanto numa fase de treino como numa fase de teste. Como resultado, o referido módulo produz uma lista de pares formados por entidade e contexto.

Na fase de detecção do contexto da EM, os candidatos a Entidades Mencionadas são identificados junto do seu respectivo contexto e são definidos por um grupo pequeno de regras pertencentes a este contexto. Para esta etapa, faz-se necessário rotular as sentenças por meio do treinamento do etiquetador morfossintático do Jet, baseado nos textos do Floresta Sintática [AFO02].

Já a extração de características faz a análise da lista de pares entidade-contexto e cria uma nova lista. As características da entidade consideradas são: a entidade propriamente dita; cada constituinte individualmente, com exceção dos elementos de ligação; a entidade possui somente letras maiúsculas e por fim, o comprimento da entidade. Para esta última característica condiciona-se que entidades com mais de cinco constituintes fiquem todas de comprimento seis.

O módulo de classificação rotula os pares de vectores de características alcançados pelo módulo de extração de características. Tal módulo utiliza um conjunto de regras que são concluídas por um algoritmo de *co-training*. Foi empregada a categoria OUTRA, embora esta não exista no grupo de categorias da avaliação, cuja finalidade é guardar as entidades que não pertencem a nenhuma das categorias: PESSOA, ORGANIZAÇÃO e LOCAL.

O módulo de propagação produz a anotação final do texto e será aplicado quando ocorrer a fase de teste. Ele tem como objetivo reconhecer as entidades que não estão nos contextos relacionados com as regras de detecção do contexto da EM, citadas

anteriormente, mas que podem ser idênticas às entidades já reconhecidas pelo sistema e que têm uma classificação associada a ele. Esse processo faz com que aumente a abrangência do sistema, pois permite a classificação de entidades que não foram classificadas pelo módulo de classificação, devido à falta de contexto. A precisão, contudo, pode ser diminuída, porque o módulo de propagação se limita, apenas, a escolher a classificação mais frequente.

Os autores alcançaram sucesso ao aplicar a estratégia proposta para o problema de Reconhecimento de Entidades Nomeadas em texto da língua portuguesa, uma vez que obtiveram um anotador de entidades em texto e não apenas um classificador de listas de entidades.

2.4.3 Sistema REMBRANDT

O sistema chamado REMBRANDT - Reconhecimento de ENs Baseado em Relações e Análise Detalhada do Texto consiste em um sistema que reconhece todo tipo de entidades nomeadas e detecta as relações entre entidades para textos da língua portuguesa [MOT08d], justificando a sua participação na trilha de ReRelEN, no Segundo HAREM. Este sistema utiliza a Wikipédia como base de conhecimento a fim de classificar as Entidades Nomeadas, além de um conjunto de regras gramaticais para extrair o seu significado por meio de indicações internas e externas delas. Tais regras são compreendidas como padrões, os quais indicam se há Entidades Nomeadas nas sentenças.

O REMBRANDT surgiu da necessidade de se criar um sistema de marcação de textos que indique as EN relacionadas a locais geográficos de uma forma semântica, como por exemplo, países, rios, universidades, monumentos ou sede de organizações. Um dos obstáculos encontrados para o desenvolvimento eficiente desta ferramenta é a desambiguação de sentidos, pois os nomes geográficos podem ser aplicados em vários contextos, entre eles, nomes de pessoas, entidades geográficas de tipos diferentes, por exemplo, Cuba significa um país e uma cidade portuguesa. A base para a criação do REMBRANDT foi parte do sistema PALAVRAS_REN [BIC06], o qual identifica EN baseando-se no analisador morfossintático PALAVRAS [BIC03] para criar regras que exploram sinais de EM nos textos. O funcionamento do REMBRANDT divide-se em três fases primordiais: 1) O reconhecimento de expressões numéricas e geração de candidatas a EN; 2) Classificação de EN e 3) Repescagem de EN sem classificação.

Os documentos são trabalhados um a um com processos de anotação sucessivos até a sua versão final, de modo que as EM detectadas adquirem um histórico de todas as suas alterações, desde a sua primeira identificação no texto até a sua última alteração.

Em relação aos demais sistemas, o REMBRANDT obteve o segundo lugar na sua participação no Segundo HAREM, obtendo os melhores resultados para as categorias: Pessoa, Local, Valor, Organização e Obra.

2.4.4 Sistema SEI-Geo no Segundo HAREM

O sistema SEI-Geo, participante também da trilha de ReRelEN do Segundo HAREM, tem o objetivo de fazer o Reconhecimento de Entidades Mencionadas classificando a categoria Local e suas relações [MOT08e].

Dentre as características que compõem o SEI-Geo destacam-se:

- incorporação na arquitetura global do sistema *GKBGeographic Knowlegde Base*, o qual estabelece o gerenciamento de conhecimento geográfico [3];

- possui dois módulos básicos: o extrator e anotador de informações geográficas e o integrador de conhecimento geográfico;

-trabalha com as Geo-ontologias, que exploram as relações entre locais identificados em textos a partir de relações presentes na ontologia.

Destaca-se que para o bom desenvolvimento do SEI-Geo, o domínio Organização ajudou significativamente, no reconhecimento de relações de Entidades Mencionadas, pois, nos textos, Locais estão localizados próximos a Organizações.

Quanto aos resultados obtidos pelo sistema, o SEI–Geo alcançou o melhor resultado comparado aos demais sistemas participantes no Segundo HAREM, salientando-se na identificação da relação de Inclusão.

2.4.5 Sistema CaGE

O sistema CaGE trata do problema do reconhecimento e desambiguação de nomes de locais, pois esta é uma tarefa muito importante na geo-codificação de documentos textuais [MAR09]. O objetivo principal do sistema CaGE é atribuir a área geográfica e o âmbito temporal aos documentos de modo geral, combinando a informação diferente extraída do texto.

[3] Disponível em: http://mchaves.wikidot.com/publicacoes

O CaGe participou do Primeiro HAREM e no Mini-HAREM com o objetivo de avaliar o seu desempenho em ambientes selecionados para o REM, classificando-as com a categoria LOCAL [SAN07].

Já no Segundo HAREM, esse sistema avaliou o processo de REM nas categorias PESSOA, ORGANIZAÇÃO e TEMPO, além do reconhecimento e classificação da categoria LOCAL considerando os tipos e subtipos dessas entidades.

O CaGe caracteriza-se por ser um método híbrido utilizando dicionários e regras de desambiguação. As referências geográficas são, muitas vezes, ambíguas com relação às entidades de outras categorias, por exemplo, a entidade Mariana que refere-se ao nome de uma Localidade, pode também indicar o nome de uma Pessoa. A solução proposta pelo CaGe para o problema de REM está no uso de métodos de aprendizagem automática [MCC03], porém para as tarefas de desambiguação completa de entidades geográficas, faz-se necessário o uso de um almanaque geográfico. Isso porque as referências devem estar relacionadas a uma representação única para o conceito geográfico, por exemplo, coordenadas de latitude e longitude ou identificadores no almanaque geográfico.

Um dicionário, contendo nomes de entidades, que trata exceções para entidades geográficas é outra ferramenta que auxilia o funcionamento do CaGE. É utilizado como complemento do sistema em questão um almanaque mais específico para desambiguação completa das EM, correspondendo a locais ou a períodos temporais, o qual faz parte do projeto DIGMAP.

Quatro etapas resumem uma sequência de operações de processamento que compõem o algoritmo do sistema: 1) Identificação inicial das EM; 2) classificação das entidades mencionadas e tratamento da ambiguidade; 3) desambiguação completa de entidades geográficas e temporais; 4) atribuição de âmbitos geográficos e temporais aos documentos.

O resultado das quatro etapas, anteriores, mostra que o sistema CaGE além de reconhecer e classificar entidades mencionadas em textos, desambigua as entidades correspondentes à referências geográficas ou temporais.

Conclui-se que este sistema obteve resultados moderados, embora os dicionários, aqui apresentados, tenham abrangido dois milhões de nomes diferentes, as regras e deduções abordadas pelo CaGE necessitam de alguma melhoria.

2.4.6 Comparação entre os sistemas

Cláudia Freitas e colegas em [FRE10] concluíram que, no processo de avaliação dos sistemas apresentados, apenas os sistemas Priberam e REMBRANDT_4 reconheceram o conjunto completo das categorias, tipos e subtipos.

Entre os sistemas que utilizaram da Coleção Dourada do Segundo HAREM, apenas o R3M adotou uma abordagem de aprendizagem de máquina, especificamente, o *co-training*. Os outros sistemas basearam-se em regras manualmente codificadas em combinação com recursos externos como dicionários, *gazetteers* [SAR06] e ontologias. Dois deles, o REMBRANDT e o REMMA fizeram uso da enciclopédia Wikipedia para o Português, de diferentes maneiras. Isso evidencia que a comunidade dedicada a REN em Português não adotou técnicas de aprendizado de máquina, ao contrário da situação para o Inglês.

Em geral, o melhor desempenho, no Segundo HAREM, foi aquele obtido pelo sistema Priberam, atuação muito próxima ao melhor funcionamento de REMBRANDT. Isto quer dizer que o primeiro utiliza uma ontologia multilíngue combinada com regras contextuais léxico-semânticas, enquanto o segundo explora a Wikipedia como fonte de conhecimento, associado com regras gramaticais que descrevem evidências internas e externas sobre as entidades nomeadas.

A comparação entre os demais sistemas participantes do Segundo HAREM não é tão simples, porque eles participaram em cenários seletivos diferentes, como, por exemplo, o sistema SEI-Geo, que aplicou somente Inclusão para a extração de Entidades Nomeadas. A avaliação por cenários seletivos só fornece uma avaliação completamente justa no caso em que o cenário de avaliação está contido nos cenários de participação, caso contrário, os sistemas que correspondem exatamente ao cenário de avaliação podem ter uma ligeira vantagem [FRE10].

Existem ainda outros sistemas que tratam de REN para o português, tais como o REMMA [MOT08f] e o Reconhecimento de Entidades Nomeadas com o XIP [MOT08g].

2.5 Reconhecimento de Entidades Nomeadas aplicando CRF

Suxiang [SUX08] apresenta uma nova abordagem para o reconhecimento de entidades nomeadas chinesas baseadas no modelo CRF, o qual possui a forma de cascata. Para o cumprimento deste propósito foi realizada a extração de uma entidade candidata do tipo pessoa e outra do tipo local. A entidade candidata será inserida dentro de um modelo estatístico para decidir se é algum tipo de entidade ou não, uma vez que a

ambiguidade de segmentação da palavra na língua chinesa sempre existe. Quando há o reconhecimento de nome próprio, alguns padrões são propostos para o tipo de modelo diferente de ambiguidade de segmentação**,** e algumas etiquetas são usados para expressar regras específicas de caracteres chineses em nomes de pessoas.

Conforme [SUX08], sete padrões para o reconhecimento de nome de pessoa foram elaborados da seguinte forma: as duas primeiras regras não são ambíguas, enquanto as outras modelam algumas possíveis ambiguidades em nome de pessoa Chinesa causadas por segmentadores de palavras. Serão apresentados três dos sete padrões para demonstrar que cada idioma tem a sua particularidade, utilizando exemplos da língua portuguesa. Os demais padrões encontram-se em [SUX08].

1º) O padrão BCD foi usado para modelar um nome de pessoa, o qual é formado por três nomes chineses. B é o sobrenome da pessoa, C é o primeiro nome dado e D é o último nome dado. Por exemplo: Leon(B) João(C) Pedro(D).

2º) O padrão BD foi usado para modelar um nome de pessoa, que é formado por dois nomes. Por exemplo: Maria(B) Antônia(D).

3º) O padrão BCH é usado para modelar um caso ambíguo no qual o último nome dado e o seu próximo caractere podem formar palavras diferentes. Por exemplo, Antônio (B) Luis (C) Vicentin o (H). Isto quer dizer que os caracteres anteriores, com exceção do último, formam um nome de pessoa, mas o último caractere é uma palavra. No exemplo dado o último caractere é um artigo e, portanto, não pertence ao último nome próprio. (Ex.: Antonio Luis Vicentin o convidou para estudar.)

Foram coletados nomes de pessoas candidatas baseados nos padrões de *feature* acima. Além disso, os caracteres ou palavras anteriores e posteriores do nome de pessoa candidato também foram utilizados.

Neste trabalho, foi implementado o reconhecimento do nome pessoal chinês e traduziu-se o nome pessoal separadamente. Estes dois tipos de nomes são muito diferentes em um texto chinês, mas, às vezes, deve-se ter atenção especial para distingui-los. Alguns nomes pessoais traduzidos sempre incluem sobrenomes chineses, que são pistas importantes para o modelo de reconhecer nomes chineses de pessoas e nomes estrangeiros de pessoas. Neste caso, um pedaço do nome de pessoa traduzido pode, muitas vezes, ser reconhecido como um nome da pessoa chinês.

O modelo procura caracteres chineses que sejam anteriores e posteriores, no contexto, a fim de encontrar alguns outros caracteres chineses que sejam nome pessoal chinês ou o nome pessoal estrangeiro. De acordo com os resultados coletados, o modelo

irá escolher o modelo pessoal chinês ou modelo pessoal de tradução para identificar o nome pessoal candidato. Experiências mostram que esse método é promissor, o *recall* e a precisão tiveram melhoras.

Foi determinada uma função de confiança para uma sequência de caracteres, para ajudar o modelo a estimar a probabilidade de uma determinada palavra ser um nome de pessoa. Por exemplo, seja f_{1F} a probabilidade de C_i , f_{iM} a probabilidade de C_1 , f_{nE} a probabilidade de C_n. Então a função de confiança é:

$$K(w, PERSON\ ) = f_{1F} + \sum_{2<=i<=n-1} f_{iM} + f_{nE}$$

Esta função é incluída no quadro do CRF como uma *feature*. Neste modelo, o reconhecimento de nomes de locais é semelhante a nomes de pessoas. A diferença entre eles é a direção da busca quando se coleta uma entidade candidata. Os modelos de CRF e a função de confiança são também utilizados para reconhecer o nome de um local. A função de confiança para nome de locais é representada pela equação:

$$K(w, location) = f_{1F} + \sum_{2<=i<=n-1} f_{iM} + f_{nE}$$

Foram estabelecidas, por exemplo, *features* para nomes de pessoa Chinês, nome de locais.

As *features* designadas para nomes de pessoa Chinês foram: informação do contexto, sobrenome, o primeiro nome, o último nome, o contexto semântico, a probabilidade de um sobrenome, a probabilidade de ocorrer um determinado nome e a função de confiança.

Já as *features* dos nomes de locais eram: a informação do contexto, a probabilidade de um caractere ocorrer e a função de confiança.

O método para o reconhecimento de organização é diferente para o reconhecimento de pessoa e de localização. Estabeleceu-se um modelo para reconhecimento da organização, de modo que a estrutura para esse apresenta-se na forma de cascata. Os resultados para o reconhecimento de local e de pessoa são repassados ao modelo, onde supõem-se que a decisão para o reconhecimento de nomes de organizações seja complicado.

O modelo pró-processado segmentará o texto chinês em uma sequência de palavras. O único passo em relação à língua chinesa é a segmentação da palavra. A segmentação é um estágio especial para as línguas orientais, por exemplo, as línguas Chinesa, Japonesa e Coreana. Isso porque aquelas línguas são formadas por uma

sequência de caracteres sem os delimitadores da palavra. Nesse processo de segmentação, nomes de pessoas e nomes de locais serão reconhecido antes do reconhecimento de uma organização numa sequência de palavras. Esse passo é muito importante e a arquitetura em cascata mostra a ordem de reconhecimento de diferentes tipos de entidades, por exemplo, um nome de organização, se a local não tiver sido reconhecido com antecedência, esta organização não será reconhecida com sucesso. Por conseguinte, os reconhecimentos do nome de pessoa e do nome de uma localização, estão em um estado muito importante.

Foram, finalmente, extraídas algumas *features* importantes para gerar o vetor de *features* e enviá-los a um classificador. Foram estabelecidos quatro rótulos BILO, que significam:

B: a primeira palavra da organização

I: a palavra do meio da organização

L: a última palavra da organização

O: uma palavra independente

Para realizar a tarefa de classificação do rótulo BILO, utilizou-se o CRF que pertence a uma *feature* baseada em aprendizado de máquina. A seleção das *features* foram apresentadas da seguinte forma:

1) *Feature* inicial;

2) *Feature* da palavra entidade, ou seja, se a palavra é um entidade do tipo pessoa, local ou organização, o valor da *feature* é 1, caso contrário é 0;

3) A palavra tem o rótulo B. Se a palavra satisfaz esta *feature,* o valor é 1;

4) A palavra tem rótulo I. Se a palavra satisfaz esta *feature,* o valor é 1;

5) A *feature* de probabilidade: a probabilidade da palavra vir a ser uma organização;

6) *Feature* de sufixo: a *feature* pode ser utilizada para decidir se a palavra tem um sufixo que antecede uma palavra do tipo organização;

7) *Feature* de bigrama, por exemplo, $W_0\, W_{+1}$ e

8) *Feature* de trigrama, como $W_0\, W_{+1}\, W_{+2}$.

Conforme apresentado, o autor propôs o reconhecimento de entidades nomeadas chinesas aplicando o modelo Conditional Random Fields utilizando uma estrutura em cascata. Ao mesmo tempo, foi estabelecida uma ordem de combinação baseada em regras e método estatístico, onde as funções de *feature* probabilística são usadas em vez das funções de *features* binárias. Foram exploradas várias novas *features* e os resultados

mostram que o modelo de CRF proposto, combinado com os novos elementos acima trouxeram melhoras significativas.

Batista, em [BAT10], elaborou um sistema, cujo acrônimo é *Hendrix - Entity Name Desambiguator and Recognizer for Information Extraction* - com o propósito de extrair entidades geográficas de documentos em português e produzir o seu resumo geográfico. O processo dividiu-se em três partes:

1ª) Reconhecer Entidades Geográficas em um documento: utilizando um modelo condicional (CRF), a fim de extrair de documentos nomes de entidades com significado geográfico, como por exemplo, nomes de ruas, rios, serras, entre outros;

2ª) Desambiguar significados geográficos: definir que significado possuem as entidades geográficas, eliminando nomes idênticos aos extraídos dos textos. Para cumprir este propósito, utilizou-se uma base de conhecimento externa, a Geo-Net-PT [CHA05];

3ª) Geração de um resumo geográfico: criar uma lista de entidades geográficas descoberta em uma base de conhecimento externa, por exemplo, em uma ontologia. O resumo geográfico pode ser utilizado em outras aplicações, sendo representados por identificadores de conceitos associados a uma ontologia.

Três módulos principais compõem o HENDRIX: primeiro, um módulo baseado no CRF, implementado pela ferramenta *Minorthird* [COH04], para extrair nomes de entidades geográficas, um segundo módulo, denominado PAREDES, o qual foi criado para análise e referenciação dos nomes das entidades encontradas pelo *Minorthird* e um terceiro, o PAGE, que faz a extração de EM em um grande corpora junto com o HENDRIX.

As Coleções Douradas do HAREM I e do Mini-HAREM foram os recursos utilizados para criar o modelo baseado em CRF para a obtenção dos nomes de entidades geográficas e, posteriormente, possibilitar a comparação do modelo de reconhecimento de entidades geográficas com outros sistemas existentes. Tais entidades foram classificadas como LOCAL e ainda, receberam outra categorização em subtipos: FÍSICO, HUMANO ou VIRTUAL.

Antes de gerar o CRF, foi realizada a etapa de etiquetagem, com a anotação de POS das entidades envolvidas. Os termos etiquetados foram ainda classificados em quatro categorias: *Begin* (termo inicial de uma entidade que será extraída), *End* (termo final de uma entidade que será extraída), *Continue* (termo que faz parte de uma entidade a ser extraída e que não é o inicial nem o final), *Unique* (um único termo que constitui a entidade a ser extraída) e *Neg* (o termo não se enquadra em nenhuma das categorias anteriores).

Foi possível comparar as métricas de Precisão, Abrangência e Medida-F com outros sistemas que tiveram uma participação na avalição seletiva apenas na categoria LOCAL no Segundo HAREM. Verificou-se que sistemas como REMBRANDT, SEI-Geo e SeRELep tiveram desempenho superior comparado com o modelo de CRF. Os resultados sugerem que gerando melhores funções de *features*, na fase de aprendizagem, seja por meio da Coleção Dourada ou através da codificação manual, os resultados poderão melhorar no que se refere à Precisão e Abrangência.

No período de teste, a Coleção Dourada, foi modificada de modo que mantivesse, apenas, as anotações para entidades geográficas. No entanto, os testes mostraram que, por exemplo, quando o modelo extrai a entidade “Portugal”, ele a identifica sempre como uma entidade geográfica, quando este pode se referir a um termo não geográfico (como, no texto, fazer alusão ao governo de Portugal), elevando assim o número de falsos positivos.

O HENDRIX participou do evento de avaliação de sistemas de perguntas e respostas, O GikiCLEF [SAN09], na edição de 2009. Seu objetivo é fazer a avaliação de sistemas que utilizam a Wikipedia para buscar documentos que contém a resposta a uma determinada pergunta ou uma informação necessária. Um único modelo de CRF foi desenvolvido para a participação neste evento, a fim de reconhecer além de lugares, organizações, eventos e pessoas. Compuseram a fase de treino do modelo as CD do Primeiro HAREM e do evento do Mini-HAREM. Já a fase de teste foi formada, somente, pela CD do Segundo HAREM, alcançando 64% de Precisão e 45% de Abrangência.

Em relação aos resultados, percebe-se que o desempenho para a categoria LOCAL diminuiu e muitas entidades foram corretamente identificadas, porém classificadas com a categoria errada. Conforme [BAT10], conclui-se, então, que deveria ter sido treinado um modelo independente para cada uma das categorias e, consequentemente, originado resultados mais satisfatórios.

3. NERP-CRF

O capítulo descreve todo o sistema proposto desde o pré-processamento dos textos, o modelo gerado pelo CRF para o reconhecimento das entidades nomeadas até o método de avaliação empregado. O objetivo é realizar REN aplicando o método CRF e, sequencialmente, fazer uma avaliação do seu desempenho com base no corpus do HAREM. A descrição do processo de avaliação, bem como a sua finalidade, serão abordadas na Seção 3.3.

3.1 Modelagem do Sistema

O NERP-CRF é o nome atribuído para o sistema desenvolvido com o propósito de realizar duas funções: a identificação de ENs e a classificação dessas com base nas dez categorias do HAREM: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro. Esse sistema teve como base o trabalho de Marlo Souza [SOU12].

A elaboração do modelo consiste em duas etapas: treino e teste. Dessa forma, o corpus é dividido em um conjunto de textos para treino (Figura 3.2) e um conjunto de textos para teste. O corpus trabalhado nesse processo refere-se às Coleções Douradas do Primeiro e do Segundo HAREM descrito na Seção 2.3.3. Estes corpora foram escolhidos, primeiramente, por serem a principal referência na área, sendo utilizados pela maioria dos trabalhos relacionados ao REN, e devido ao fato deles disponibilizarem um conjunto de textos anotados e validados por humanos (Coleção Dourada), o que facilita a avaliação do método em estudo.

Os textos, utilizados como entrada, estão no formato XML com a marcação das entidades e sofreram dois procedimentos, os quais pertencem ao pré-processamento do sistema: primeiro, a etiquetagem de cada palavra por meio do Part-of-Speech (POS) tagging [SCH94] e segundo, a segmentação em sentenças a fim de que a complexidade seja menor ao aplicar o algoritmo de CRF nos textos de entrada. O exemplo, de acordo com a sentença retirada da CD do Segundo HAREM (Figura 3.1), ilustra esses procedimentos iniciais:

“Os EUA ganharam um interesse...”

Os<art> <EM EUA<prop> /EM> ganharam <v-fin> um <art> interesse<n>...

Figura 3.1: Exemplo do texto segmentado e etiquetado.

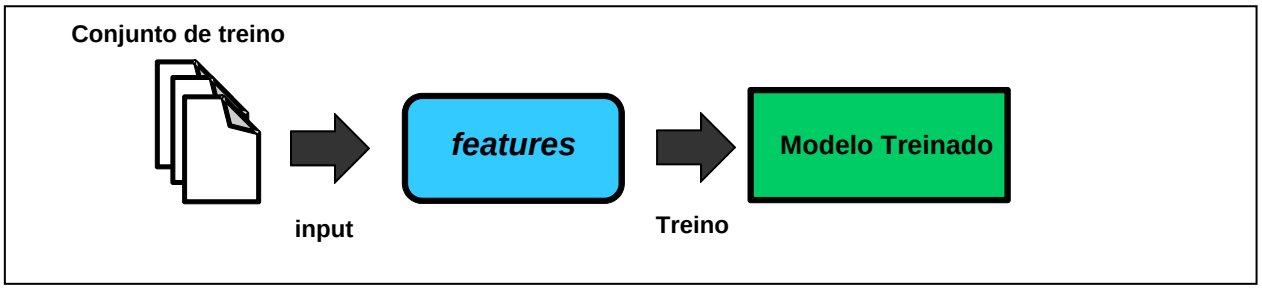

Figura 3.2: Procedimentos na etapa de treino.

Após a conclusão da etiquetagem POS e da segmentação das sentenças, determinou-se como as EN seriam identificadas. Para tal, foi feito um estudo de duas notações citadas na literatura [RAT09]: BIO e BILOU. A primeira possui o seguinte significado: B (Begin) significa a primeira palavra da EN; I (Inside) uma ou mais palavras que se localizam entre as entidades e O (Outside) a palavra não é uma EN. Já a segunda notação, tem a mesma descrição do BIO, acrescentando-se as seguintes particularidades: L (Last) a última palavra reconhecida como EN e U (Unit) quando a EN for uma única palavra. Salienta-se que I (Inside), na notação BILOU, encontra-se entre *Begin* e *Last*.

Optou-se por utilizar, para o presente trabalho, a notação BILOU por dois motivos: (i) Testes aplicados sob a CD do Segundo HAREM, empregando ambas notações, demonstraram que a notação BILOU supera a BIO, conforme os resultados apresentados na Seção 4. Resultados (Tabelas 4.1 a 4.4). Isso porque o BILOU facilita o processo de classificação feito pelo sistema desenvolvido por possuir mais duas identificações: L(Last) e U(Unit); e (ii) Ratinov e Roth em [RAT09] também fizeram testes com as duas notações, concluindo também com os seus resultados obtidos que, apesar do formalismo BIO ser amplamente adotado, o BILOU o supera significativamente.

Logo, pelas duas justificativas apresentadas no parágrafo anterior, as EN do corpus de treino receberam a notação BILOU a qual pode ser exemplificada de acordo com a Figura 3.3.

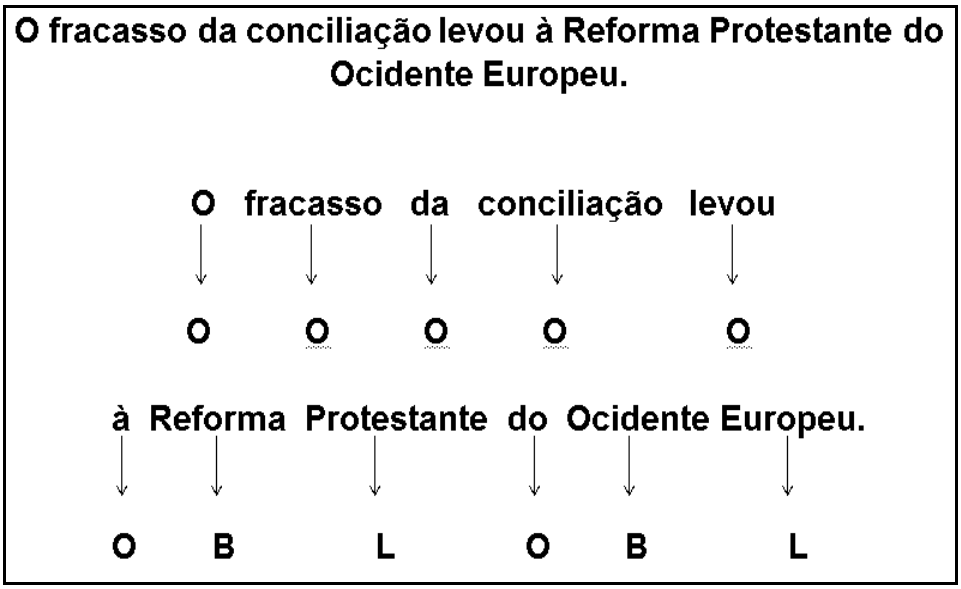

Figura 3.3: Exemplo retirado da CD do Segundo HAREM com a aplicação da notação BILOU.

Após a identificação das EN por meio do BILOU, foi gerado o vetor de *features.* Tal vetor corresponde aos dados de entrada que serão aplicados ao sistema de aprendizado do CRF. As *features* têm o objetivo de caracterizar todas as palavras do corpus escolhido para este processo, direcionando o CRF na identificação e na classificação das ENs. A seguir, a lista das *features* criadas:

1) ‘tag’: a etiqueta *POS tagging* de cada palavra de acordo com a sua classe gramatical;

2) ‘word’: a própria palavra, ignorando letras maiúsculas e minúsculas;

3) ‘prevW’: a palavra anterior, ignorando letras maiúsculas e minúsculas;

4) ‘prevT’: a classe gramatical da palavra anterior;

5) ‘prevCap’: se a palavra anterior for totalmente formada por letras minúsculas, formada por letras minúsculas e maiúsculas ou totalmente formada por letras maiúsculas;

6) ‘prev2W’: igual a *feature* 3, porém considerando a palavra que está na posição p-2;

7) ‘prev2T’: o mesmo que a *feature* 4, considerando a palavra que está na posição p-2;

8) ‘prev2Cap’: igual a *feature* 5, porém considerando a palavra que está na posição p-2;

9) ‘nextW’: a palavra subsequente àquela que está sendo analisada, ignorando maiúsculas e minúsculas;

10) ‘nextT’: A classe gramatical da palavra subsequente àquela que está sendo analisada;

11) ‘nextCap’: o mesmo que a *feature* 5, levando em consideração a palavra subsequente àquela que está sendo analisada;

12) ‘next2W’, ‘next2T’, ‘next2Cap’: semelhante as *features* 3,4 e 5, mas para a palavra que se encontra na posição p + 2;

13) ‘cap’: o mesmo que a *feature* 5, mas para a palavra atual que está sendo analisada;

14) ‘ini’: se a palavra iniciar com letra maiúscula, minúscula ou símbolos;

15) ‘simb’: Caso a palavra seja composta por símbolos, dígitos ou letras.

Dois vetores são considerados como entrada para o CRF: primeiro, o vetor contendo a etiquetagem POS, as categorias estabelecidas pela Conferência do HAREM e a notação BILOU (Figura 3.4) e segundo, o vetor de *features*. Para um melhor entendimento das *features* apresentadas anteriormente, a Figura 3.5 ilustra a 1ª e a 3ª *features,* para mesma sentença anterior. No final deste trabalho, localizam-se todas as *features* aplicadas no CRF (Apêndice A) e um exemplo de vetor completo de entrada para o sistema com todas essas *features* (Apêndice B).

Os EUA ganharam um interesse...

Vetor com POS e com o BILOU:

[(('Os', 'art'), 'O'), (('EUA', 'prop'), 'U- ORGANIZAÇÃO'),
(('ganharam', 'v-fin'), 'O'),

Figura 3.4: Exemplo de um vetor com POS e com a notação BILOU.

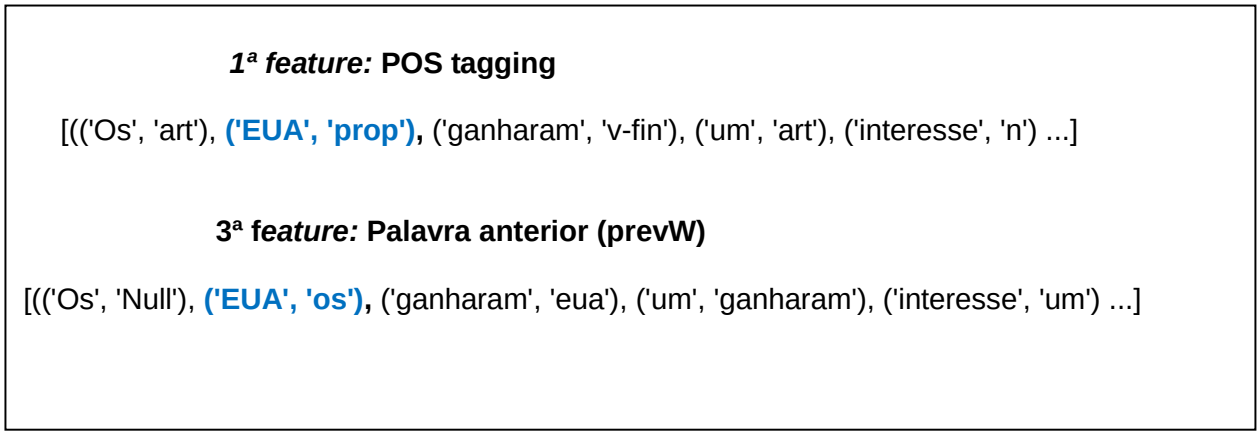

Figura 3.5: Exemplo de duas *features* aplicadas no sistema proposto.

Uma matriz é gerada, a partir do vetor de *features,* dado como entrada do sistema, para produzir o modelo CRF. Esse modelo corresponde a uma matriz de pesos que indica os pesos sobre o valor de cada *feature,* cujo objetivo é informar, conforme a Figura 3.5, quantas palavras, por exemplo, serão classificadas como U e são artigos na posição p-1. Baseado na condição anterior, o algoritmo resulta em um valor de peso para o atributo p-1. Consequentemente, toda vez que uma palavra for etiquetada como um artigo e estiver na posição p-1, o algoritmo atribuirá a ela, por exemplo, um peso no valor de 0,1. Os pesos formadores da matriz do CRF estão entre 0 e 1, pois valores probabilísticos ficam compreendidos dentro desse intervalo numérico.

A etapa de teste utiliza o mesmo vetor de *features* da etapa de treino como vetor de entrada. O modelo de CRF gerado, na etapa de treino, é aplicado no corpus de teste, a fim de que se possa avaliar o seu desempenho, de acordo com a Figura 3.6.

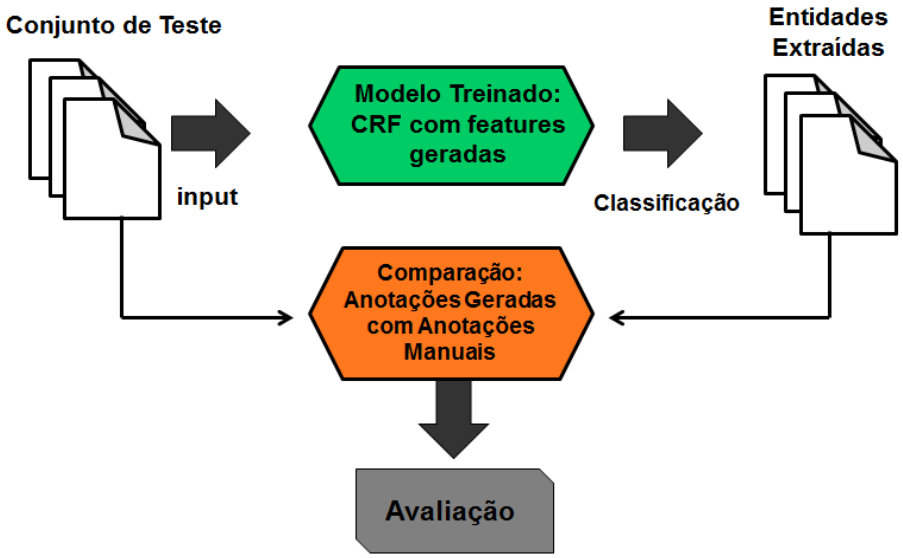

Figura 3.6: Procedimentos na etapa de teste.

De acordo com a sentença-exemplo: “Os EUA ganharam um...”, o CRF seleciona o conjunto de etiquetas de maior probabilidade para cada sentença e gera a seguinte saída final do sistema:

['O-', 'U-ORGANIZAÇÃO', 'O-', 'O-', ...]

A Seção seguinte descreverá a implementação para a modelagem proposta. Para melhor compreensão de como o NERP-CRF foi executado, serão apresentados exemplos de entradas e saídas em cada um deles, durante o cumprimento do sistema.

3.2 Implementação

A implementação é caracterizada pelo uso de ferramentas com a finalidade de: 1) segmentar os textos em sentenças, para que a complexidade seja menor ao aplicar o algoritmo de CRF nos textos de entrada; 2) etiquetar as sentenças através do *Part-of-Speech* a fim de identificar morfologicamente cada palavra dos textos, auxiliando o CRF na classificação das EN; 3) criar as *features*; e 4) gerar o modelo de CRF conforme explicado na Seção 2.1.

O pré-processamento do sistema NERP-CRF é formado pela segmentação dos textos e etiquetagem (*POS tagging*) das palavras (Figura 3.7). Para o desenvolvimento dos procedimentos anteriores foi utilizada a biblioteca OpenNLP[4], implementada na

[4] Disponível em http://opennlp.apache.org/

linguagem de programação Java, pois ela possui vários recursos bem estruturados para PLN, como esses aplicados no pré-processamento.

```
Os<art> <EM CATEG="ORGANIZAÇÃO"> EUA<prop> </EM> ganharam<v-fin> um<art>
interesse<n> acrescido<v-pcp> pelas<adv> armas<n> não<adv> letais<adj> após<prp> a<art>
sua<pron-det> desastrosa<adj> missão<n> pacificadora<adj> na<adj> <EM CATEG="LOCAL">
Somália<prop> </EM>
```

Figura 3.7: Exemplo de uma sentença segmentada e etiquetada.

O sistema NERP-CRF (Figuras 3.8 e 3.9) tem como entrada o corpus pré-processado e foi desenvolvido em *Python*, uma vez que essa linguagem de programação apresenta sintaxe clara, concisa e elegante, facilitando a manutenção do código.

Na etapa de treino, o NERP-CRF transforma o corpus pré-processado em: um vetor e uma função de tradução. Tal vetor é formado pela etiquetagem POS *tagging* e pela notação BILOU. A função de tradução utiliza os dados extraídos do vetor de entrada para criar o vetor de *features* o qual vai caracterizar as sentenças formadoras desse vetor. A biblioteca NLTK, escrita na linguagem Python, foi utilizada para criar o vetor POS + BILOU e a função de tradução (vetor de *features*). Já a biblioteca *Mallet* (versão 0.4), escrita na linguagem Java, foi utilizada na implementação do CRF, que irá gerar o modelo de CRF baseado nas features determinadas como uma das entradas do NERP-CRF. Optou-se por trabalhar com o
Mallet, porque esse possui recursos que facilitam a extração de informação e a criação de aplicações de aprendizado de máquina para textos [SUT09].

Na etapa de teste, um conjunto de textos é enviado ao NERP-CRF. O referido sistema cria o vetor POS e a função de tradução; envia esses vetores para o modelo de CRF gerado que, por sua vez, treina e classifica as EN do corpus trabalhado. Por fim são apresentadas aos usuários do sistema as EN extraídas e as métricas precisão e abrangência.

Em suma, o NERP-CRF permite que as ENs dos textos de entrada sejam categorizadas e que os resultados finais sejam expostos, de modo que se possam comparar os textos anotados manualmente e os textos anotados pelo sistema.

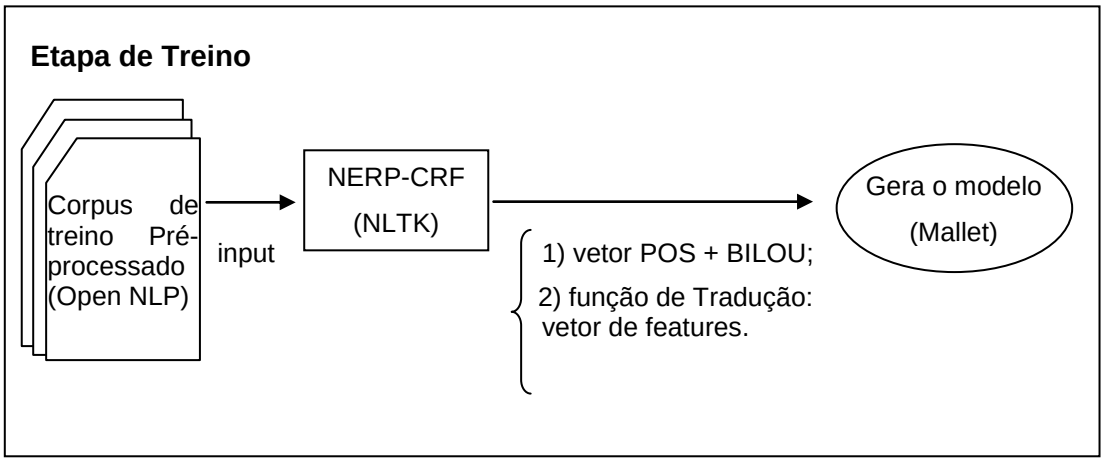

Figura 3.8: Desenvolvimento do sistema NERP-CRF, etapa de Treino.

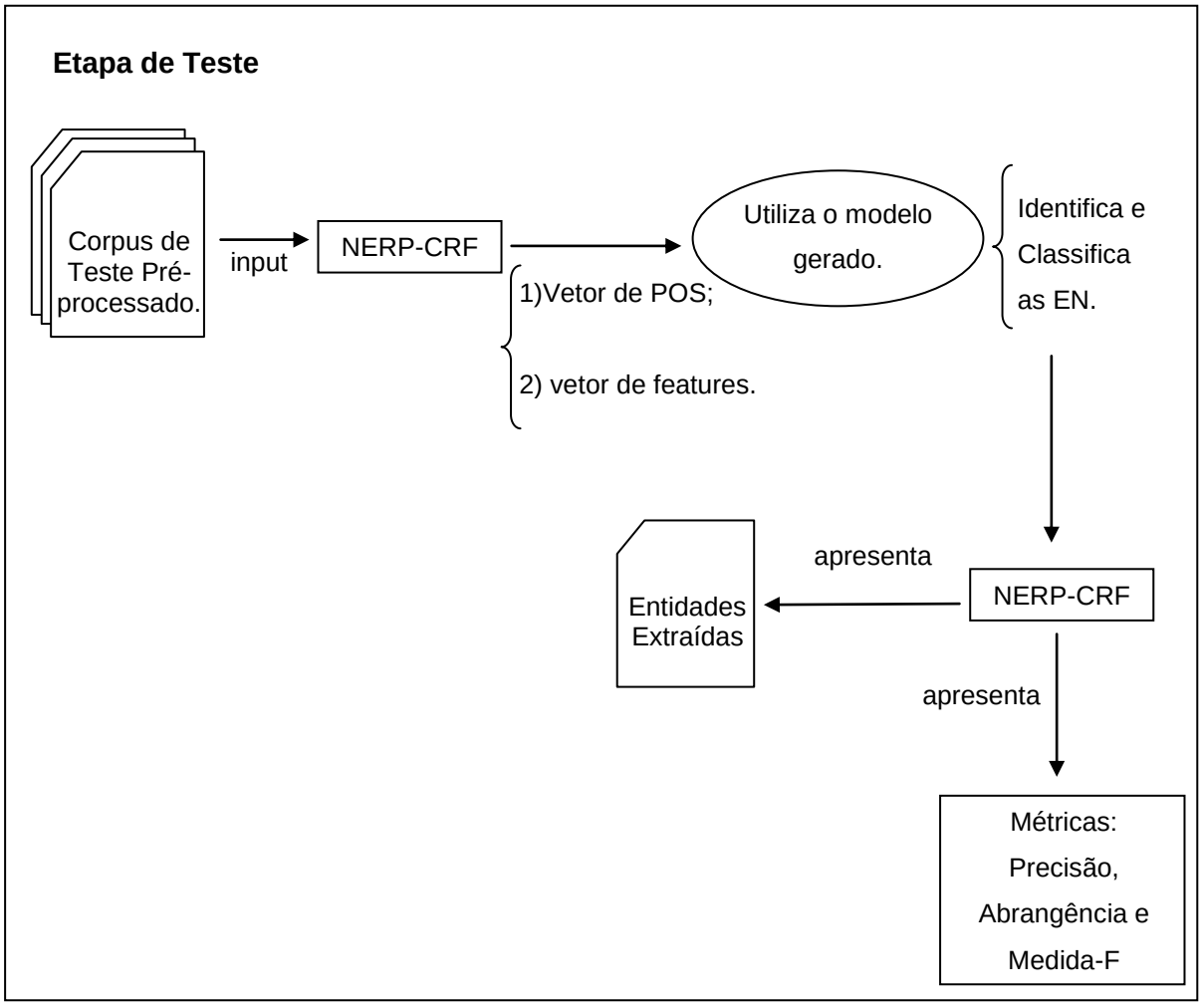

Figura 3.9: Desenvolvimento do sistema NERP-CRF, etapa de Teste.

Com base na implementação descrita, três testes diferentes foram estabelecidos para aplicar o CRF. Obedeceu-se a mesma modelagem do sistema, conforme a seção 3.2, para os três, os quais foram chamados de Teste 1, 2 e 3.

A sentença “Os EUA ganharam um interesse acrescido pelas armas não letais após a sua desastrosa missão pacificadora na Somália”, exemplifica as entradas e as saídas aplicadas na implementação desse trabalho bem como os procedimentos realizados, a seguir pormenorizados.

Na etapa de treino são utilizados dois vetores:

1) o primeiro vetor é constituído de:

- etiquetagem POS *tagging* de cada palavra do texto,
- notação BILOU em cada palavra do texto e
- a categorização das EN identificadas pelo BILOU.

Esse experimento empregou cinco das dez categorias consideradas pela Conferência do HAREM: Pessoa, Local, Tempo, Organização e Obra. Inicialmente, optou-se por trabalhar somente com as cinco categorias anteriores, pois essas são as mais frequentes encontradas nas CD do HAREM [MOT08c] e além disso, tinha-se o propósito de se verificar o comportamento do modelo de CRF proposto para o corpus em questão.

2) Já o segundo vetor é formado por um conjunto de *features*, as quais foram determinadas para esse trabalho na Seção 3.2. A entrada para o CRF, nessa etapa, para esse sistema, consiste:

- do primeiro vetor mencionado anteriormente e
- de uma função de tradução.

Na etapa de treino, os dois vetores de entrada, Figuras 3.10 e 3.11, utilizados para gerar o modelo CRF, são apresentados a seguir. O vetor de *features* completo encontra-se no final dessa dissertação (Apêndice C).

(('Os', 'art'), 'O-'), (('EUA', 'prop'), 'U-ORGANIZAÇÃO'), (('ganharam', 'v-fin'), 'O-'), (('um', 'art'), 'O-'), (('interesse', 'n'), 'O-'), (('acrescido', 'v-pcp'), 'O-'), (('pelas', 'adv'), 'O-'), (('armas', 'n'), 'O-'), (('não', 'adv'), 'O-'), (('letais', 'adj'), 'O-'), (('após', 'prp'), 'O-'), (('a', 'art'), 'O-'), (('sua', 'pron-det'), 'O-'), (('desastrosa', 'adj'), 'O-'), (('missão', 'n'), 'O-'), (('pacificadora', 'adj'), 'O-'), (('na', 'adj'), 'O-'), (('Somália', 'prop'), 'U-LOCAL')

Figura 3.10: Vetor contendo a etiquetagem POS taggin, a notação BILOU e a categorização das EN.

{'nextCap': 'max', 'cap': 'maxmin', 'word': 'os', 'next2W': 'ganharam', 'next2T': 'v-fin', 'tag': 'art', 'nextT': 'prop', 'simb': 'alfa', 'nextW': 'eua', 'next2Cap': 'min', 'ini': 'max'}

{'nextCap': 'min', 'cap': 'max', 'word': 'eua', 'prevCap': 'maxmin', 'next2W': 'um', 'next2T': 'art', 'prevT': 'art', 'prevW': 'os', 'nextT': 'v-fin', 'simb': 'alfa', 'nextW': 'ganharam', 'next2Cap': 'min', 'tag': 'prop', 'ini': 'max'}

{'nextW': 'um', 'nextCap': 'min', 'cap': 'min', 'word': 'ganharam', 'prevCap': 'max', 'next2W': 'interesse', 'next2T': 'n', 'prevT': 'prop', 'prevW': 'eua', 'nextT': 'art', 'simb': 'alfa', 'tag': 'v-fin', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'maxmin', 'prev2W': 'os', 'prev2T': 'art'}

{'nextW': 'interesse', 'nextCap': 'min', 'cap': 'min', 'word': 'um', 'prevCap': 'min', 'next2W': 'acrescido', 'next2T': 'v-pcp', 'prevT': 'v-fin', 'prevW': 'ganharam', 'nextT': 'n', 'simb': 'alfa', 'tag': 'art', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'max', 'prev2W': 'eua', 'prev2T': 'prop'}

{'nextW': 'acrescido', 'nextCap': 'min', 'cap': 'min', 'word': 'interesse', 'prevCap': 'min', 'next2W':'pelas', 'next2T': 'adv', 'prevT': 'art', 'prevW': 'um', 'nextT': 'v-pcp', 'simb': 'alfa', 'tag': 'n', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'ganharam', 'prev2T': 'v-fin'}

Figura 3.11: Vetor contendo as features para o segmento de sentença “Os EUA ganharam um interesse...”

Após a introdução dos vetores de entrada, no sistema, foi gerado o modelo de CRF de acordo com as features apresentadas pela Seção 3.2.

Na etapa de teste, o vetor de entrada é formado pelos:

vetor de etiquetagem de cada palavra do texto por meio do POS taggin (Figura 3.12) e

pelo mesmo vetor de features aplicado, na etapa de treino, de acordo com a Figura 3.11.

Já o vetor de saída da etapa de teste possui o seguinte formato (Figura 3.13):

{('Os', 'art'), ('EUA', 'prop'), ('ganharam', 'v-fin'), ('um', 'art'), ('interesse', 'n'), ('acrescido', 'v-pcp'), ('pelas', 'adv'), ('armas', 'n'), ('não', 'adv'), ('letais', 'adj'), ('após', 'prp'), ('a', 'art'), ('sua', 'pron-det'), ('desastrosa', 'adj'), ('missão', 'n'), ('pacificadora', 'adj'), ('na', 'adj'), ('Somália', 'prop')}

Figura 3.12: Vetor que contém cada palavra do texto etiquetada pelo POS *tagging*.

'O-', 'U-ORGANIZACAO', 'O-', 'O-', 'O-', 'O-', 'O-', 'O-', 'O-', 'O-',
'O-', 'O-', 'O-', 'O-', 'O-', 'O-', 'O-', 'U-LOCAL'

Figura 3.13: Vetor de saída que classifica o texto com a notação BILOU e com as categorias.

O “Teste 1” utilizou a CD do Segundo HAREM, composta por 129 textos, e gera um modelo de CRF que faz a classificação de cinco categorias: Acontecimento, Local, Pessoa, Obra e Organização. O intuito de realizar um experimento com essas distinções deve-se ao fato de verificar qual a melhor notação a ser utilizada aplicando o CRF: BIO ou BILOU.

O “Teste 2” também utilizou a CD do Segundo HAREM para treinar e testar o modelo de CRF, o qual faz a classificação de dez categorias: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro. A finalidade de executar esse experimento foi para realizar a avaliação dos resultados obtidos pelo CRF com os outros sistemas participantes do Segundo HAREM, podendo assim comparar tais resultados.

Já o “Teste 3” caracteriza-se por trabalhar com a CD do Primeiro HAREM para treino, na qual abrange 129 textos e a CD do Segundo HAREM para teste formada por mais 129 textos. O novo corpus recebe a classificação do CRF abordando as dez categorias do HAREM, citadas no “Experimento 2”. Essa terceira estrutura foi arquitetada com o objetivo de verificar o desempenho do CRF em um maior número de textos.

Os vetores de entrada e saída, tanto nas etapas de treino quanto nas etapas de teste para os “Experimentos 2 e 3” possuem a mesma estrutura do “Experimento 1” exemplificado anteriormente. As duas diferenças dos dois últimos experimentos em relação ao primeiro é que:

1) o CRF para esses textos considerou todas as categorias do HAREM para classificar cada EN e

2) as features referentes as posições vazias, na sentença, de acordo com a janela considerada, em relação a posições anteriores e posteriores à palavra em questão, receberam o valor nulo. Por exemplo, considerando a sentença “Os EUA ganharam um interesse...”, se for considerada duas posições anteriores em relação a palavra EUA, os valores de suas features serão nulos, uma vez que não há nenhuma palavra duas posições anteriores a ela na sentença. Logo, o vetor de features para a palavra EUA, o qual aborda essa particularidade, possui o formato apresentado na Figura 3.14:

{'nextW': 'ganharam', 'nextCap': 'min', 'next2T': 'art', **'word': 'eua'**, 'prevCap': 'maxmin', 'next2W': 'um', 'cap': 'max', 'prevT': 'art', 'prevW': 'os', 'nextT': 'v-fin', 'simb': 'alfa', 'tag': 'prop', 'ini': 'max', **'prev2Cap': 'null'**, **'prev2W': 'null'**, 'next2Cap': 'min', **'prev2T': 'null'**}

Figura 3.14: As *features* destacadas receberam o valor “null”. Particularidade essa aplicada nos experimentos 2 e 3.

Resumidamente, a implementação é apresentada contendo as suas características mais relevantes. NERP-CRF é o sistema desenvolvido em *Python* neste trabalho de mestrado. A implementação que envolveu todo o processo proposto é formada por uma combinação de ferramentas associadas a uma implementação mais específica desenvolvida nessa pesquisa. O modelo do NERP-CRF foi gerado empregando as seguintes ferramentas: a) a biblioteca OpenNLP[5] em Java, para a etiquetagem e divisão dos textos em sentença, utilizando a técnica do *Part-of-Speech (POS) tagging* [SCH94] e notação BILOU na identificação das ENs, na etapa do pré-processamento; b) a biblioteca NLTK, escrita na linguagem Python, foi utilizada para criar dois vetores de entrada para o sistema; c) a biblioteca Mallet, escrita em Java, para a implementação do modelo.

A aplicação do sistema é dividida em duas etapas: treino e teste. Na etapa de treino, a biblioteca NLTK transformou o corpus pré-processado em dois vetores considerados como entrada para o sistema. O primeiro contendo a etiquetagem POS, as categorias estabelecidas pela Conferência do HAREM e a notação BILOU. O segundo

[5] Disponível em http://opennlp.apache.org/

definindo as *features.* Este vetor foi gerado com o objetivo de caracterizar todas as palavras do corpus a fim de orientar o NERP-CRF na identificação e na classificação das ENs. A biblioteca *Mallet* (versão 0.4), foi utilizada na implementação do NERP-CRF gerando o modelo de CRF baseado nas *features* determinadas como uma das entradas do sistema. Na etapa de teste, a biblioteca *Mallet* foi aplicada com o objetivo de treinar o modelo, a partir de um outro conjunto de textos diferente do utilizado na etapa de treino. Novamente nessa etapa o sistema cria o vetor POS e juntamente com o vetor de *features* identifica e classifica as ENs.

3.3 Avaliação

A avaliação tem por objetivo comparar o NERP-CRF com os outros sistemas que fizeram REN utilizando o corpus da CD do Segundo HAREM. Para alcançarmos uma avaliação precisa, estudou-se de que maneira os resultados seriam avaliados e que ferramenta seria utilizada para gerar os mesmos. A avalição foi feita quantitativamente após o desenvolvimento do sistema NERP-CRF, a qual restringiu-se à tarefa de reconhecimento e classificação das EN. A metodologia de avaliação descreve esses estudos feitos bem como as medidas utilizadas para avaliar o nosso sistema. O processo de avaliação aborda de que maneira a metodologia pesquisada foi aplicada nos três experimentos concluídos.

3.3.1 Metodologia de Avaliação

A elaboração do processo de avaliação implicou no conhecimento do *Cross-Validation,* do SAHARA, bem como do estudo do corpus de referência HAREM. A seguir, cada um deles é descrito de acordo com a ordem de aplicabilidade nesse trabalho.

Os corpora utilizados neste trabalho são os do Primeiro e do Segundo HAREM, apresentados na seção 2.3.3. Estes corpora foram escolhidos, primeiramente, por ser a principal referência na área, empregados pela maioria dos trabalhos relacionados ao REN, e pelo fato do HAREM estar disponível aos usuários, caracterizando-se por ser um conjunto de textos anotados e validados por humanos (Coleção Dourada), o que facilita a avaliação do método em estudo. Nos três testes realizados, utilizou-se a Coleção Dourada do Segundo HAREM, como o corpus de teste, para validar o modelo de CRF gerado, uma vez que se optou por avaliar esses textos por meio da ferramenta SAHARA[6],

[6] Disponível em: http://www.linguateca.pt

a qual é executada em textos que possuem o formato, apenas, do Segundo HAREM [MOT08b].

Posteriormente, optou-se por trabalhar com o corpus do Primeiro HAREM para treino e o do Segundo HAREM para teste, pois assim o CRF pode ser aplicado em um maior número de textos: de 129 alterou-se para 258 textos na sua totalidade (129 da CD do Primeiro HAREM mais 129 CD do Segundo HAREM). Para ambos os experimentos, a modelagem do sistema foi a mesma, conforme explanado na seção 3.1.

3.3.1.1 Cross-validation

Cross-validation foi a técnica empregada para validar os dados obtidos a partir do modelo treinado no corpus do HAREM. Segundo [ARL09], *Cross-validation* ou Validação-cruzada é uma técnica ou um estimador de validação de dados que realiza a média de várias estimativas de maneira segura, ou seja, os dados que foram validados são autênticos, eficazes e correspondem a diferentes divisões dos textos. Kohavi em [KOH95] afirma que a estimativa da validação cruzada é um número randômico que depende da divisão dos dados de entrada em *folds*, ou seja, iterações. A Validação-cruzada completa é a média de todas as possibilidades para a escolha de instâncias m/k por m, mas ela é geralmente muito cara. Isso significa que apesar de apresentar uma investigação completa sobre a variação do modelo em relação aos dados utilizados, este método possui um alto custo computacional, sendo indicado para situações onde poucos dados estão disponíveis. O *Cross-validation* caracteriza-se por ser um estimador intermediário entre *Holdout* e do *Leave-one-out*, descritos posteriormente.

O método *Holdout* [KOH95] tem a função de dividir o conjunto total de dados em dois subconjuntos: um para treinamento e outro para teste. O subconjunto de treinamento faz a estimativa ou cálculo dos parâmetros, já o de teste realiza a validação dos dados. A proporção usual para divisão dos dados é dois terços para treinamento e um terço dos dados para teste, contudo o conjunto de dados pode ser separado em quantidades de proporções diferentes ou iguais. Depois de ocorrido o particionamento, o modelo é gerado, os dados de teste são aplicados e é estimado o erro de predição. Indica-se a utilização desse método quando for empregada uma grande quantidade de dados, pois se o conjunto de dados for pequeno, o erro estimado na predição poderá ter como resultado muita variação.

Existem alguns casos especiais de *Cross-validation* como: o método *K-fold* e o *Leave-one-out.* O método *K-fold* [KOH95] fundamenta-se em dividir o conjunto de dados

total em x subconjuntos de tamanhos iguais. Após esse particionamento um subconjunto é utilizado para treino e o restante para teste, calculando por fim, a acurácia do modelo gerado. É realizado n vezes esse processo, chamado de *folds* ou iterações, de modo que se combinem todos os subconjuntos criados, alternado os de treino com o de teste. No momento em se findam as iterações, calcula-se a taxa de erro '*err(h)'* de um classificador *'h'* encontrados por meio da equação:

$$err(h) = 1/n \sum_{i=1}^{n} || y_i \# h(x_i) ||$$

Dado 'n' o número de atributos, a taxa de erro compara o texto original classificado com a etiqueta atribuída pelo sistema classificador criado. O operador ||E|| retornará 1 se a expressão E for considerada verdadeira, caso contrário retornará zero. Dessa forma, obtém-se a medida mais confiável sobre a competência do modelo de representar o processo gerador dos dados.

Já o método *Leave-one-out* envolve o uso de uma amostra original onde, a partir dela ocorre a divisão dos dados em uma amostra de validação e amostras de treinamento. Esse processo é repetido de modo que a cada observação na amostra, os dados de validação são utilizados uma única vez. É o mesmo procedimento que ocorre no método *K-fold* com k sendo igual ao número de observações na amostra original. Dada uma amostra de tamanho '*n'*, o erro na amostragem consiste na soma dos erros em cada iteração divido por '*n'*. *Leave-one-out* é computacionalmente caro porque requer muitas repetições de treinamento e é usado, com frequência, em amostras pequenas.

3.3.1.2 SAHARA

Além da técnica de validação *Cross Validation* [ARL10], também foi utilizado o SAHARA, ferramenta que automaticamente avalia os resultados dos textos da CD do Segundo HAREM. Através do SAHARA é possível fazer a comparação entre os sistemas participantes do Segundo HAREM com outros sistemas que realizam REN, desde que estes trabalhem com textos que estejam no mesmo formato dos textos da CD do Segundo HAREM, condição imposta pelo SAHARA para se utilizar essa ferramenta. Esse avaliador está disponível na Web pelo site da Linguateca[7] e facilita a avaliação dos sistemas que fazem o REN, pois o usuário não necessita executar comandos próprios dos

[7] http://www.linguateca.pt/

programas de avaliação para obter os resultados referentes ao desempenho do sistema a ser avaliado.

O processo de avaliação feito pelo SAHARA é formado por três fases:

- validação do sistema a ser avaliado de acordo com o formato do Segundo HAREM;

- configuração do tipo de avaliação, isto é, a escolha dos cenários, o modo de avaliação e as coleções que serão trabalhadas e

- a exibição dos resultados formada por tabelas e gráficos, os quais identificam o desempenho do sistema.

Findada a apresentação dos resultados, o SAHARA faz a verificação da existência de resultados oficiais os quais possam ser comparáveis com a avalição executada, ou seja, se os sistemas participantes do Segundo HAREM foram avaliados de acordo com o mesmo cenário e modo de avaliação utilizando a mesma CD. Os resultados oficiais, caso esses sejam considerados, são exibidos no formato de um gráfico-resumo o qual apresenta também a melhor saída para os três melhores sistemas.

3.3.2 Medidas de Avaliação

Nesta sub-seção serão apresentadas duas avaliações do modelo gerado pelo CRF, utilizando a CD do Segundo HAREM: primeiro os resultados obtidos pelo *Cross Validation* e, segundo os valores oriundos do SAHARA. Os dados repassados aos dois avaliadores resultam nas métricas de Precisão, Abrangência e Medida-F. De acordo com as diretivas do SAHARA, as medidas de avaliação obedecem as seguintes condições mencionadas a seguir.

A Precisão corresponde à medida da qualidade do sistema em termos de resposta e mede a proporção de respostas corretas dadas todas as respostas fornecidas pelo sistema, ou seja:

$$\text{Precisão} = \frac{\text{Total de EN corretamente classificadas pelo NERP-CRF}}{\text{Total de EN classificadas pelo CRF}}$$

A Abrangência mede a quantidade de EN classificadas corretamente em relação ao universo das EN identificadas pela Coleção Dourada (CD) do Segundo HAREM.

$$\text{Abrangência} = \frac{\text{Total de EN corretamente classificadas pelo NERP-CRF}}{\text{Total de EN classificadas pela CD}}$$

A Medida F combina as métricas de precisão e de abrangência de acordo com a fórmula:

$$\text{Medida-F} = \frac{2 * \text{Precisão} * \text{Abrangência}}{(\text{Precisão} + \text{Abrangência})}$$

3.3.3 Processo de Avaliação

O processo de avaliação determinado é caracterizado pelo uso de uma técnica de validação e de uma ferramenta, a seguir apresentadas:

- A avaliação do desempenho do modelo treinado para o "testes 2 utilizou a técnica de *Cross Validation* [ARL10], com cinco repetições (*5 – fold cross validation).* Trabalhou-se com 5 *folds* porque foi empregado uma pequena quantidade de textos, 129, para os testes iniciais. Dado o conjunto de textos da CD do Segundo HAREM, utilizou-se a cada *fold,* 80% do conjunto de textos para treino e 20% para teste, de modo que a cada repetição do *Cross Validation,* não se empregasse o mesmo conjunto de teste das *folds* anteriores e assim, não reduzisse, significativamente, o número de casos para teste. A CD é um subconjunto da coleção do Segundo HAREM, organizado pela Linguateca[8]. Tal subconjunto é formado por 129 documentos e seus textos foram anotados por humanos. O "teste 3" considerou um *fold Cross-validation* para que também fosse possível utilizar o SAHARA como ferramenta de avaliação. Para isso aplicou-se a CD do Primeiro HAREM como conjunto de treino e a CD do Segundo HAREM para ser o conjunto de teste, uma vez que o SAHARA exige que, para se trabalhar com essa ferramenta, os textos estejam no formato desse conjunto de textos utilizados para teste.

- A ferramenta SAHARA[9] [MOT08] foi empregada para fazer a comparação dos resultados obtidos pelo NERP-CRF com os sistemas participantes do Segundo HAREM.

[8] Disponível em http://www.linguateca.pt/harem/

[9] Disponível em http://www.linguateca.pt/harem/

Como o principal objetivo dessa dissertação é realizar a avaliação do NERP-CRF, o SAHARA é o sistema de avaliação adequado para essa finalidade. Logo os três testes, detalhados na Seção 3.2, utilizaram a ferramenta mencionada e a CD do Segundo HAREM como corpus de teste.

4. RESULTADOS

Os resultados apresentados pelo NERP-CRF identificam cada EN por meio da notação BILOU e as classificam considerando o corpus das CD do Primeiro e do Seguno HAREM. A difícil missão de identificar possíveis falhas nos procedimentos pode ser feita de forma automatizada por meio de técnicas de PLN como técnicas de aprendizado de máquina. Aplicações que se beneficiam de tal suporte podem ser aplicadas em textos dos mais diversos domínios. Seguindo a metodologia adotada, verifica-se que o sistema desenvolvido apresentou os melhores resultados de Precisão quando comparado com outros sistemas, os quais adotaram os mesmos recursos. Os resultados estão organizados e serão apresentados de acordo com os três testes pormenorizados na Seção 3.2, os quais podem ser sintetizados da seguinte forma:

O ‘Teste 1’ utilizou a CD do Segundo HAREM, cujo objetivo é definir qual a notação que será utilizada para gerar o modelo de CRF: BIO ou BILOU.

O ‘Teste 2’ também utilizou a CD do Segundo HAREM para treinar e testar o modelo de CRF, o qual faz a classificação de dez categorias: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Já o ‘Teste 3’ caracteriza-se por trabalhar com a CD do Primeiro HAREM para treino e a CD do Segundo HAREM para teste. O novo corpus recebe a classificação do CRF abordando as dez categorias, citadas no ‘Teste 2’.

4.1 ‘Teste 1’

Os primeiros resultados para esse trabalho foram gerados com o objetivo de verificar qual a melhor notação a ser utilizada pelo NERP-CRF. A Tabela 4.1 apresenta os resultados da identificação das EN por meio da notação BIO. Já a Tabela 4.2 classificada cada EN por meio da mesma notação.

Tabela 4.1: Identificação das EN por meio da notação BIO.

	B	I	O	Rec	Prec	F-Measure
B	5800	393	1044	0.80	0.86	0.83
I	520	6053	1686	0.73	0.89	0.80
0	374	295	68517	0.99	0.96	0.98

Tabela 4.2: Classificação das EN usando a notação BIO.

CATEGORIAS	PESSOA	ACONTECIMENTO	LOCAL	OBRA	ORGANIZAÇÃO	OUTRO	Rec	Prec	F-Measure
PESSOA	**2832**	**15**	**206**	**84**	**189**	**519**	**0.74**	**0.64**	**0.68**
ACONTECIMENTO	**118**	**233**	**89**	**72**	**68**	**224**	**0.29**	**0.66**	**0.40**
LOCAL	**333**	**32**	**1111**	**63**	**178**	**251**	**0.56**	**0.58**	**0.57**
OBRA	**297**	**23**	**117**	**405**	**173**	**600**	**0.25**	**0.44**	**0.32**
ORGANIZAÇÃO	**343**	**26**	**224**	**79**	**814**	**328**	**0.45**	**0.48**	**0.47**
OUTRA	**520**	**24**	**177**	**209**	**259**	**73447**	**0.98**	**0.97**	**0.98**

A Tabela 4.3 exibe os valores de identificação de cada EN utilizando a notação BILOU e a apresentação dos resultados por categorias, considerando a mesma notação, é descrita pela Tabela 4.4.

Tabela 4.3: Identificação das EN por meio da notação BILOU.

	B	I	L	O	U	Rec	Prec	F-Measure
B	**3012**	**186**	**28**	**706**	**142**	**0.74**	**0.82**	**0.78**
I	**295**	**2445**	**202**	**1207**	**36**	**0.58**	**0.82**	**0.68**
L	**24**	**142**	**3182**	**515**	**211**	**0.78**	**0.87**	**0.82**
O	**205**	**155**	**122**	**68566**	**138**	**0.99**	**0.96**	**0.97**
U	**113**	**38**	**116**	**365**	**2531**	**0.80**	**0.83**	**0.81**

Tabela 4.4: Classificação das EN usando a notação BILOU.

CATEGORIAS	PESSOA	ACONTECIMENTO	LOCAL	OBRA	ORGANIZAÇÃO	OUTRA	Rec	Prec	F-Measure
PESSOA	**2764**	**19**	**209**	**86**	**208**	**559**	**0.84**	**0.70**	**0.77**
ACONTECIMENTO	**129**	**184**	**103**	**82**	**65**	**241**	**0.33**	**0.63**	**0.43**
LOCAL	**371**	**35**	**1074**	**69**	**168**	**251**	**0.63**	**0.62**	**0.62**
OBRA	**292**	**26**	**115**	**410**	**172**	**600**	**0.40**	**0.56**	**0.47**
ORGANIZACAO	**374**	**26**	**242**	**80**	**736**	**356**	**0.50**	**0.54**	**0.52**
OUTRA	**531**	**30**	**179**	**212**	**254**	**73430**	**60.89**	**36.59**	**45.71**

Devido aos melhores resultados obtidos na tabela de categorias com o BILOU (Tabela 4.4), essa foi a notação adotada para os próximos testes. Acreditamos que por

essa notação ter uma maior granularidade, ela facilita o processo de classificação feito pelo NERP-CRF, por possuir mais duas identificações: L (Last) e U (Unit).

4.2 ‘Teste 2’

A técnica de Cross Validation avalia a classificação BILOU de cada palavra do texto e a categorização das EN, apresentadas pelas Tabelas de Confusão a seguir. O Teste 2 foi executado sobre o corpus da CD do Segundo HAREM, contendo 129 textos, incluindo 670.610 palavras. Esse procedimento resultou em 7.610 EN identificadas pelo NERP-CRF num valor máximo de 17.767 EN identificadas por humanos nessa mesma CD.

De acordo com a Tabela 4.5 (Tabela de Confusão da classificação BILOU), observa-se que os valores de F-Measure aproximam 80% para as categorias B L e U. O menor F foi para I, que deve obedecer à condição de que a palavra esteja localizada entre B (Begin) e L (Last). Contudo, como esta situação é menos frequente, houve poucos exemplos para treino do CRF para essa categoria. A categoria O (Outside) possui alto F porque a maioria das palavras do texto recebe esse tipo de notação.

De acordo com Tabela 4.6, a categoria Tempo foi a que obteve a melhor Precisão, 83,99% e também um bom resultado de Abrangência, 68,05%. Consequentemente, foi o melhor resultado de F-Measure classificado por esse sistema, alcançando 75,18%. Pode-se constatar, nessa mesma Tabela, que 372 EN foram classificadas como Pessoa, ao passo que deveriam ser classificadas pelo NERP-CRF como Organização. A explicação para esse fato foi a falta de contexto existente no corpus, a qual não auxiliou o NERP-CRF na classificação correta da categoria Organização.

Tabela 4.5: Identificação das EN no ‘Teste 2’.

	B	**I**	**L**	**O**	**U**	**Rec**	**Prec**	**F-Measure**
B	3041	69	33	664	167	75%	83%	79%
I	92	496	207	**1154**	36	60%	83%	**69**%
L	28	42	3166	520	218	77%	86%	82%
O	171	162	131	68590	132	99%	96%	**98**%
U	**15**	26	21	361	2540	80%	82%	81%

Tabela 4.6: Classificação das EM do NERP-CRF no ‘Teste 2’.

CATEGORIAS	PESSOA	ACONTECIMENTO	LOCAL	OBRA	ORGANIZAÇÃO	TEMPO	COISA	ABSTRAÇÃO	VALOR	OUTRA	Rec	Prec	F-Measure
PESSOA	**2764**	30	277	103	233	6	22	27	0	4	**71,89%**	**61,57%**	**66,33%**
ACONTECIMENTO	134	**183**	131	103	78	41	12	9	3	0	**22,76%**	**50,83%**	**31,44%**
LOCAL	362	37	**126**	82	177	16	8	9	0	1	**57,22%**	**52,06%**	**54,51%**
OBRA	312	34	132	**460**	178	25	17	24	5	2	**28,48%**	**40,71%**	**33,52%**
ORGANIZACAO	**372**	29	257	98	**788**	6	14	21	1	1	**43,44%**	**44,75%**	**44,08%**
TEMPO	14	7	11	14	0	**2266**	0	2	51	1	**68,05%**	**83,99%**	**75,18%**
COISA	136	7	78	61	75	10	**41**	19	2	3	**7,36%**	**26,80%**	**11,55%**
ABSTRAÇÃO	217	9	90	106	120	6	11	**25**	1	1	**3,65%**	**16,45%**	**5,97%**
VALOR	1	1	**0**	4	8	111	2	2	**363**	0	**54,42%**	**78,23%**	**64,19%**
OUTRA	51	9	19	24	31	7	11	4	1	**10**	**4,74%**	**43,49%**	**8,55%**

4.3 ‘Teste 3’

O ‘Teste 3’ teve como base de treino a CD do Primeiro HAREM e como base de validação o corpus do Segundo HAREM. Os dois conjuntos somam 258 textos e aproximadamente 804.179 palavras.

A Tabela 4.7, sobre a classificação BILOU, mostra que os valores de F-Measure ficam em torno de 65% e 70%. Como esperado, os valores ficam um pouco abaixo do Teste 2, baseado em C*ross Validation* sobre uma única base.

Quanto a classificação das EN (Tabela 4.8), no ‘Teste 3’, pode-se constatar que a categoria Valor obteve o melhor resultado de F-Measure, 66%. Já o pior resultado classificado pelo NERP-CRF foi para a categoria Coisa com F-Measure de 3%. A diferença dos resultados originados pelas métricas entre os ‘Testes 2 e 3’ deve-se à diferença da distribuição das categorias da Primeira para a Segunda CD do HAREM. Algumas categorias ocorreram mais vezes na primeira CD e menos na segunda CD, influenciando no treinamento do CRF.

Tabela 4.7: Identificação das EN no ‘Teste 3’.

	B	**I**	**L**	**O**	**U**	**Rec**	**Prec**	**F-Measure**
B	2356	105	27	1386	200	57,83%	74,86%	65,25%
I	464	1704	182	1782	53	40,72%	83,90%	**54,83%**
L	34	111	2500	829	600	61,36%	86,87%	71,92%
O	123	91	67	68722	183	99,33%	93,77%	**96,47%**
U	170	20	102	570	2301	72,75%	68,95%	70,80%

Tabela 4.8: Classificação das EN no ‘Teste 3’.

CATEGORIAS	PESSOA	ACONTECIMENTO	LOCAL	OBRA	ORGANIZAÇÃO	TEMPO	COISA	ABSTRAÇÃO	VALOR	OUTRA	Rec	Prec	F-Measure
PESSOA	2425	9	345	49	**378**	5	8	82	3	0	**63%**	**63%**	**63%**
ACONTECIMENTO	122	69	11	52	161	31	0	0	16	0	**9%**	**45%**	**14%**
LOCAL	291	8	1017	25	323	4	2	48	2	0	**52%**	**51%**	**51%**
OBRA	257	27	112	243	304	18	8	47	8	0	**15%**	**43%**	**22%**
ORGANIZACAO	255	8	198	38	996	1	1	51	1	0	**55%**	**39%**	**45%**
TEMPO	5	0	15	8	4	1019	0	2	133	0	**31%**	**87%**	**45%**
COISA	159	6	52	31	113	6	9	13	7	0	**2%**	**25%**	**3%**
ABSTRAÇÃO	187	4	91	46	145	2	1	48	2	0	**7%**	**14%**	**9%**
VALOR	1	0	2	1	5	32	0	0	449	0	**67%**	**65%**	**66%**
OUTRA	59	4	12	12	46	3	1	8	2	5	**2%**	**100%**	**5%**

4.4 Comparação com outros Sistemas

A comparação dos resultados do NERP-CRF com os sistemas que participaram da Conferência do Segundo HAREM foram obtidos por meio do SAHARA, o qual determinou as métricas Precisão, Abrangência e Medida-F a cada um deles nas tarefas de reconhecimento e classificação de EN. O NERP-CRF, no 'Teste 2', apresentou os melhores resultados para as medidas de Precisão e Medida-F em relação aos outros sistemas (Figura 4.1).

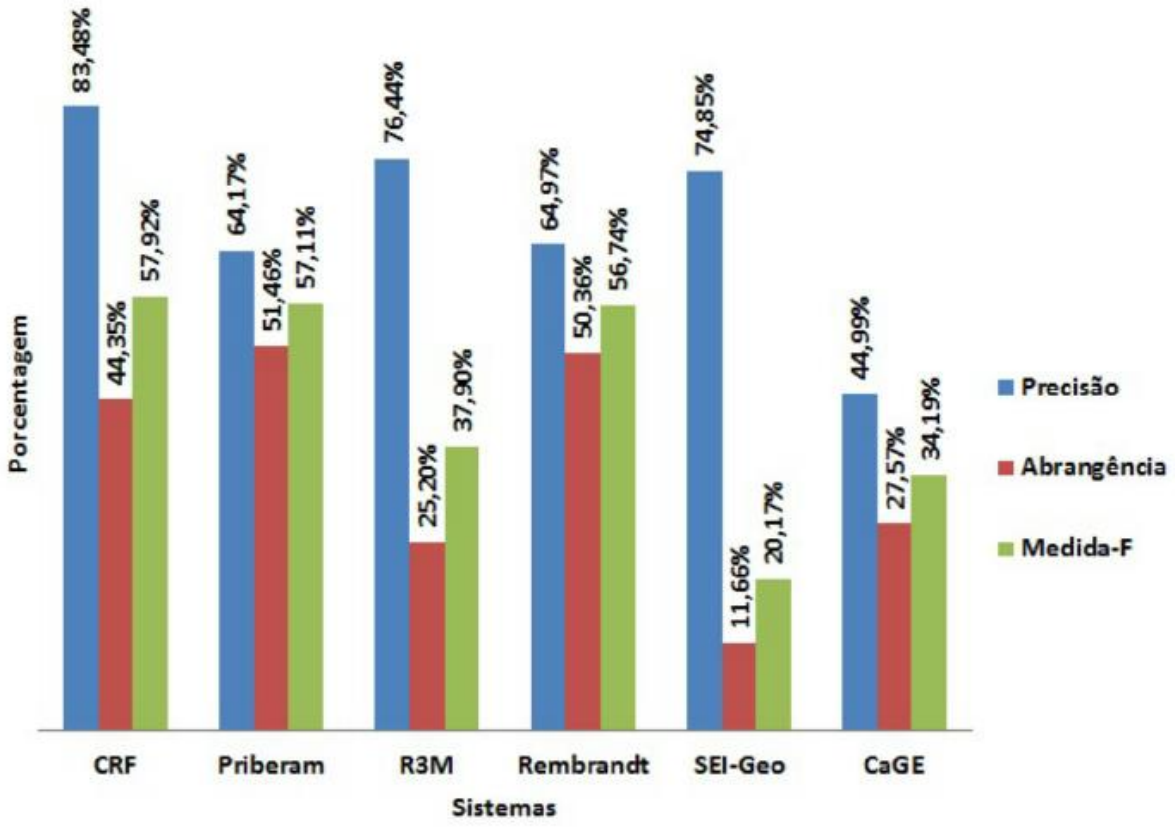

Figura 4.1: NERP-CRF comparado com os sistemas apresentados para o 'Teste 2'.

O 'Teste 3' caracterizou a Precisão de 80,77% como o melhor resultado do NERP-CRF (Figura 4.2). A Medida-F ocupou a terceira posição em relação aos sistemas em comparação, 48,43%. Essa última métrica não alcançou a melhor posição como no 'Teste 2' devido a uma baixa Abrangência de classificação, 33,74%.

A desigualdade dos resultados entre os dois testes ocorreu, principalmente, por dois motivos: a mudança do corpus de treino e de validação além do número reduzido de exemplos para determinadas categorias, por exemplo, Coisa, Abstração. Isso faz com que

o CRF treine menos com essas categorias e gere um modelo menos abrangente para elas. Neste cenário, consideram-se os nossos resultados muito positivos, principalmente no que tange ao valor de Precisão alcançado pelo NERP-CRF.

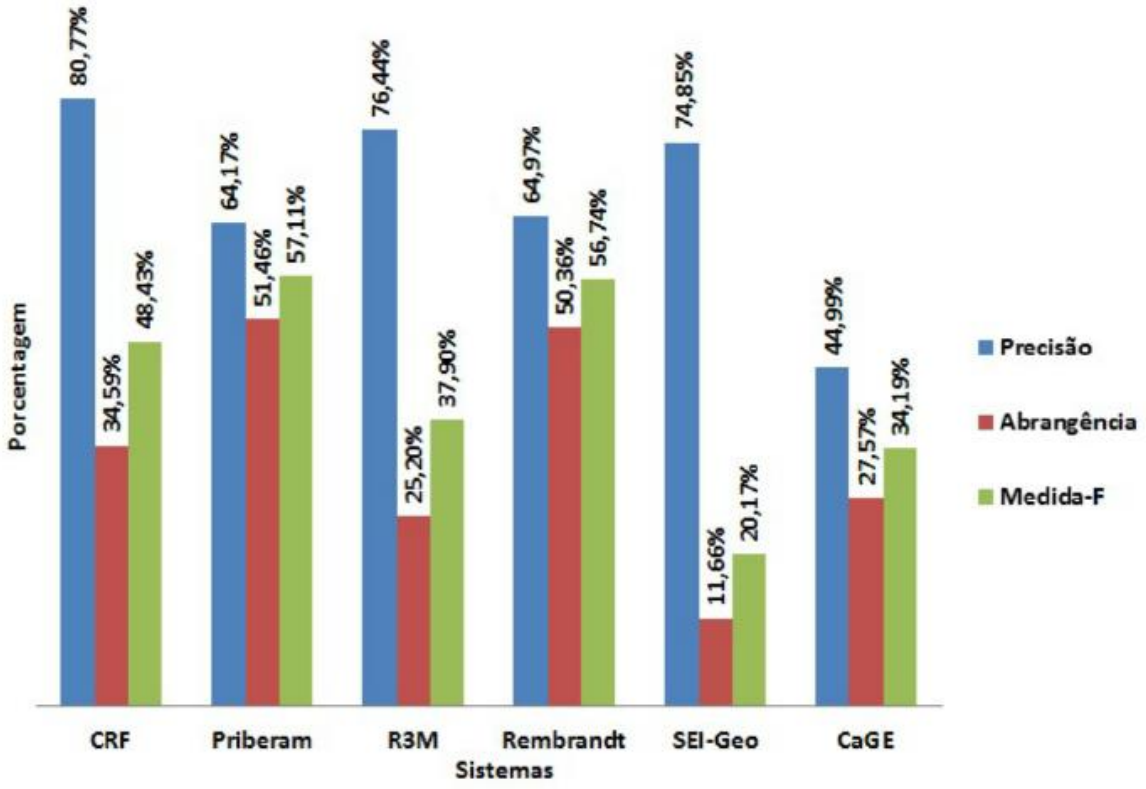

Figura 4.2: NERP-CRF comparado com os sistemas apresentados para o 'Teste 3'.

4.5 Análise de Erros

Com base em uma análise nos textos utilizados como entrada para testar o NERP-CRF, constata-se que o sistema, tanto para o 'Teste 2' quanto para o 'Teste 3', não identificou determinadas EN ou não as identificou corretamente pelos seguintes motivos:

- má formatação de alguns textos, como por exemplo, falta de pontuação;

- erros referentes a anotação do *POS tagger*;

- erros de delimitação da entidade, isto é, a EN *"Ministério da Cultura"*, por exemplo, foi marcada pela CD como B I L, no entanto o CRF a identificou como B I U;

Os erros de classificação das EN ocorreram pelos seguintes motivos

- a não identificação correta das EN pelo BILOU e, consequentemente, a classificação errada dessa mesma entidade;

- EN representadas por siglas;

- palavra estrangeira;
- a categoria Tempo é difícil de classificar, pois esse tipo de EN pode não iniciar com letra maiúscula e assim, fica mais difícil para o sistema aprender essa particularidade. Além disso, a referida categoria segue um padrão bem rígido de sintaxe como <um número> de <outro número>, indicando data, ou até mesmo outras palavras indicativas de tempo como desde, enquanto e quando.
- pouco contexto para classificar corretamente certas EN, por exemplo, a categoria Abstração tem muito pouca exemplificação na CD e são ENs que não seguem padrão algum como o caso de categoria Tempo que segue uma sintaxe própria; e
-as preposições, por exemplo “de”, são comuns tanto fora como dentro de uma EN, e muitas vezes o classificador não a identifica com I.

As três sentenças seguintes apresentam algumas situações dos erros ocorridos de acordo com os motivos já discriminados. Relatam-se exemplos mais detalhados e completos, referentes a análise de erros, no anexo IV.

Sentença 1:
“Avanços na área de radares e de comunicação **Radio** <prop,B-PESSOA> **Frequency** <prop,L-PESSOA> continuaram através das décadas de 50 e 60.”

Erro de classificação, pois o NERP-CRF deveria ter classificado a entidade *Radio Frequency* como Coisa e não como Pessoa.

Sentença 2:
“O parto ocorreu no quarto andar esquerdo do nº 4 do Largo de São Carlos, em frente da ópera de Lisboa, **Teatro** <prop,B-OBRA> **de** <prp,I-LOCAL> **São** <prop,B-PESSOA> **Carlos** <prop,L-PESSOA> .”

Erro de classificação, ou seja, o sistema classificaria corretamente a EN grifada se a considerasse como Local.

Sentença 3:

Às<adv,O-OUT> **três**<num,O-OUT> horas<n,I-TEMPO> e<conj-c,I-TEMPO> **vinte**<num,O-OUT> **minutos**<n,O-OUT> da<v-pcp,I-TEMPO> tarde<adv,I-TEMPO> de<prp,B-TEMPO> 13<num,I-TEMPO> de<prp,I-TEMPO> Junho<n,I-TEMPO> de<prp,I-TEMPO> 1888<num,L-TEMPO> nascia em Lisboa, capital portuguesa, Fernando Pessoa.

O NERP-CRF não identificou parte da EN em negrito e consequentemente, não a classificou corretamente.

5. CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as conclusões e as contribuições científicas alcançadas nessa dissertação. Bem como, serão apresentados também os trabalhos futuros os quais poderão complementar essa dissertação de mestrado no processo de identificação e classificação de EN por meio do CRF.

5.1 Conclusões

O principal objetivo dessa dissertação foi a aplicação do CRF para a tarefa de REN em corpus da língua portuguesa e a avaliação comparativa com outros sistemas que realizam REN, tendo como base o corpus do HAREM. Para isso, efetuou-se, inicialmente, um estudo teórico sobre CRF para REN. Após a conclusão desta fundamentação teórica, apresentou-se o modelo gerado pela técnica de aprendizagem automática, CRF, bem como os testes executados aplicando esse modelo em textos do Português.

A literatura tem apresentado a aplicação do formalismo matemático probabilístico denominado CRF para essa tarefa. Tal formalismo vem crescendo em importância, por ser um modelo gráfico não direcionado que define uma única distribuição logaritmicamente linear, sobre sequências de etiquetas, dada uma sequência de observação particular. Adicionalmente, o CRF evita o problema de viés dos rótulos, uma fraqueza exibida pelos MEMM e outros modelos de Markov condicionais baseados em modelos gráficos direcionados, onde os vértices são os estados e as arestas são as probabilidades de transição entre esses estados.

CRF oferece uma combinação única de propriedades: modelos treinados para etiquetar e segmentar sequências; combinação por arbitrariedade, características de observação aglomeradas, decodificação e treinamento eficiente baseado em programação dinâmica e estimativa de parâmetro garantida para encontrar o ótimo global. Sua principal limitação corrente é a lenta convergência do algoritmo de treino em relação aos MEMMs, por exemplo, para que o treino sobre os dados completamente observados seja muito eficiente [LAF01] [RAT09]. Na próxima seção destacamos as principais contribuições obtidas pela pesquisa.

5.2 Contribuições Científicas

O NERP-CRF foi o sistema desenvolvido neste trabalho para realizar duas funções: a identificação de ENs e a classificação dessas com base nas dez categorias do HAREM:

Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro.

Dois testes foram realizados. Um dos testes utilizou a CD do Segundo HAREM para treino e teste, obtendo Precisão de 83,48% e Medida-F de 57,92%. Tais resultados são os melhores quando comparados com os outros sistemas participantes do Segundo HAREM.

O outro empregou a CD do Primeiro HAREM para treinar o modelo de CRF e a CD do Segundo HAREM para testar o mesmo modelo gerado. Nesse caso as métricas obtidas foram: 80,77% de Precisão e 48,43% de Medida-F. A Precisão também foi o melhor resultado quando comparado com os outros sistemas. Já a Medida-F apresentou o terceiro melhor resultado, ficando abaixo dos sistemas Priberam e Rembrandt, que apresentaram maior abrangência.

De acordo com os dois testes desenvolvidos neste trabalho, verificou-se que o CRF é um modelo que produziu o efeito significativo esperado com base nos excelentes resultados apresentados, face à concorrência com os outros sistemas os quais ele foi avaliado.

O objetivo foi alcançado e o modelo proposto, baseado em CRF bem como no conjunto de *features* estabelecidas, gerou um sistema eficaz, competitivo, sendo ainda passível de fácil adaptação e modificação. Esse sistema obteve resultados melhores quando comparados com sistemas avaliados no mesmo corpus, apresentando a melhor pontuação de Precisão, até agora, para o conjunto de dados do corpus do HAREM.

Pode-se citar, dentre as contribuições científicas dessa dissertação, o processo de identificação e classificação de EN por meio do método supervisionado denominado CRF para o corpus do HAREM. Até então, não há nenhum trabalho que apresente exatamente essa proposta para o referido corpora. Uma vez que os métodos encontrados, na literatura, aplicaram nos textos do HAREM heurísticas para identificar e classificar as EN considerando as dez categorias estabelecidas por essa conferência.

5.3 Trabalhos Futuros

Os trabalhos futuros dessa dissertação determinam-se em duas abordagens de pesquisa: algoritmos de indução de *features* e classificação de EN consideradas ambíguas.

Um aspecto atraente do CRF é que esse pode implementar, eficientemente, a seleção de *features* e de algoritmos de indução de *features.* Isto quer dizer que ao invés

de especificar antecipadamente quais *features* serão utilizadas, pode-se iniciar a partir de regras que geram *features* e avaliam o benefício dessas geradas automaticamente sobre os dados [LAF01]. Em particular, os algoritmos de indução de *features* apresentados em [PIE97] podem ser aplicados para adaptar-se à técnicas de programação dinâmica de CRF.

Outra abordagem de pesquisa futura é a classificação correta de uma mesma EN apresentada de formas diferentes, por exemplo: a EN *Pontifícia Universidade Católica do Rio Grande do Sul* pode receber a mesma classificação ou ser categorizada como Organização e Local dependendo do contexto na qual essa entidade está inserida. Isso implica que o REN é caracterizado por tornar as decisões interdependentes complexas, as quais exigem grande quantidade de conhecimento prévio e a aplicação de decisões não locais para essa EN receber classificações diferentes. Outra situação que pode ocorrer é quando as EN Pontifícia Universidade Católica do Rio Grande do Sul e PUCRS são a mesma entidade e, portanto, devem receber a mesma classificação. As soluções para a correta categorização de EN nesse caso pode ser a aplicabilidade de recursos externos como, por exemplo, Correferência [MUC7b] [BLA98] [LEE11] e o emprego de *Gazetters* [RAT09].

Além disso, outros trabalhos futuros relevantes podem ser feitos, os quais incluem: qual modelo usar para inferência sequencial, como representar *chunks* em textos e quais algoritmos de inferência utilizar. Assim, será muito provável que se resolva o problema da ambigüidade entre EN minimizando o processo de anotação semântica exaustiva e aumentando a abrangência dos sistemas que utilizam CRF para a classificação de EN [CHA11].

REFERÊNCIAS BIBLIOGRÁFICAS

[AFO02] Afonso, S.; Bick, E.; Haber, R.; Santos, D. “Floresta sintática: um treebank para o português”. In: XVII Encontro Nacional da Associação Portuguesa de Linguística (APL), 2002, pp. 533–545.

[AMA04] Amaral, C.; Figueira, H.; Mendes, A.; Mendes, P.; Pinto, C. “A workbench for developing natural language processing tools”. In: 1st Workshop on International Proofing Tools and Language Technologies, Patras, Greece, July 1-2, 2004.

[APP99] Appelt, D. E.; Hobbs, J. R.; Bear, J.; Israel, D.; Tyson, M. “FASTUS: A finite-state processor for information extraction from real-world text”. *Journal Computational Linguistic,* vol. 25, Jun 1999, pp. 237-265.

[ARL10] Arlot, S.; Celisse, A. “A survey of cross-validation procedures for model selection”. Statistics Surveys, vol. 04, 2010, pp. 4, 40.

[BAT10] Batista, D. S.; Silva, M. J.; Couto, F.; Behera, B. “Geographic Signatures for Semantic Retrieval”. In: 6th Workshop on Geographic Information Retrieval, 2010, pp.18-19.

[BEN03] Bender, O.; Och, F. J.; Ney, H. “Maximum entropy models for named entity recognition”. In: 7th Conference on Natural Language Learning, vol. 04, 2003, pp. 148-151.

[BIC03] Bick, E. “Multi-level REN for Portuguese in a CG framework”. In: 6th International Conference on Computational Processing of the Portuguese Language, PROPOR, 2003, pp. 118-125.

[BIC06] Bick, E. “Functional Aspects on Portuguese REN”. In: 7th International Conference on Computational Processing of the Portuguese Language, PROPOR, 2006, pp. 80-89.

[BIK97] Bikel, D. M.; Miller, S.; Schwartz, R.; Weischedel, R. “Nymble: a high-performance learning name-finder”. In: 5th Conference on Applied Natural Language Processing, 1997, pp. 194–201.

[BIR09] Bird, S.; Loper E.; Klein, E. “Natural Language Processing with Python”. O'Reilly Media Inc., 2009, pp. 504.

[BLA98] Black, W. J.; Rinaldi, F.; Mowatt, D. “FACILE: Description of the NE System Used for MUC-7.” In: 7th Message Understanding Conference (MUC-7), 1998.

[BOT91] Bottou, L.; Gallinari, P. “A framework for the cooperation of learning algorithms”. In: Advances in Neural Information Processing Systems, vol. 3, D. Touretzky and R. Lippmann, Eds. Denver, CO: Morgan Kaufmann, 1991.

[CHA05] Chaves, M. S.; Silva, M. J.; Martins, B. “A geographic knowledge base for Semantic Web applications”. In: 20º Simpósio Brasileiro de Banco de Dados (SBBD), 2005, pp. 40–54.

[CHA11] Eric C.; Michel G.; Benoit O. “Automatic Semantic Web annotation of named entities”. In: Canadian Conference on AI, 2011, pp. 74-85.

[CHA12] Chatzis, S. P.; Demiris, Y. “The echo state conditional random field model for sequential data modeling”. *International Journal of Expert Systems with Applications*, 2012.

[CHI94] Chinchor N.; Hirschman L.; Lewis D. “Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). Computational Linguistics, 1994, pp. 409–449.

[CHI03] Chieu, H. L.; Ng, H. T. “Named entity recognition with a maximum entropy approach”. In: 7th Conference on Natural Language Learning, 2003, pp. 160–163.

[CIR01] Ciravegna, F. “Adaptive information extraction from text by rule induction and generalisation”. In: 17th International Joint Conference on Artificial Intelligence, vol. 2, 2001, pp. 1251–1256.

[COL99] Collins, M.; Singer, Y. “Unsupervised models for named entity classification”. In: Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999, pp.100–110.

[COH04] William W. C. “Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data”. Capturado em: http:// minorthird.sourceforge.net, 2004.

[CUR03] Curran, J. R.; Clark, S. “Language independent REN using a maximum entropy tagger”. In: 7th Conference on Natural Language Learning, 2003, pp. 164-167.

[DOD04] Doddington, G.; Mitchell, A.; Przybocki, M.; Ramshaw, L.; Strassel, S.; Weischedel, R. “The Automatic Content Extraction (ACE) program: Tasks, data, and evaluation”. In: 4th International Conference on Language Resources and Evaluation – LREC, 2004, pp. 837–840.

[DUR98] Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. “Biological sequence analysis: Probabilistic models of proteins and nucleic acids”. Cambridge University Press, 1998.

[FIN05] Finkel, J.; Dingare, S.; Manning, C. D.; Nissim, M.; Alex, B.; Grover, C. “Exploring the boundaries: gene and protein identification in biomedical text”. BMC Bioinformatics, 2005, 6 (Suppl 1): S5.

[FRE10] Freitas, C.; Mota, C.; Santos, D.; Oliveira, H. G.; Carvalho, P. “Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese”. In: 7th International Conference on Language Resources and Evaluation (LREC), 2010. European Language Resources Association (ELRA), Valletta.

[GRI96] Grishman, R.; Sundheim, B. “Message Understanding Conference - 6: A Brief History”. In: 16th International Conference on Computational Linguistics, 1996, pp. 466–471.

[HU08] Hu, H.; Zhang, H. “Chinese Named Entity Recognition with CRFs: Two Levels”. In: International Conference on Computational Intelligence and Security, 2008, pp. 1-6.

[JIA12] Jiang, J. “Information extraction from text”. In: Mining Text Data. Cap. 2, 2012, pp. 11–41.

[JUN 12] Jung, J. J. “Online named entity recognition method for microtexts in social networking services: A case study of twitter”. *In: International Jornal Expert Systems with Applications*, Elsevier, vol. 39, 2012, pp. 8066–8070.

[KRI81] Kripke, S. A. “Naming and necessity”. Cambridge, MA: Harvard University Press, 1981.

[KOH95] Kohavi, R. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. In: International joint Conference on artificial intelligence, 1995, pp. 1137–1143.

[LAF01] Lafferty, J.; McCallum, A.; Pereira, F. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In: 18th International Conference on Machine Learning (ICML), 2001, pp. 282-289.

[LEE11] Lee, H.; Peirsman, Y.; Chang, A.; Chambers, N.; Surdeanu, M.; Jurafsky, D. “Stanford’s Multi-Pass Sieve Conference Resolution System at the CoNLL-2011 Shared Task”. In: 15th Conference on Computational Natural Language Learning, 2011, pp. 28-34.

[LI11] Lishuang L.; Degen H.; Dan L. “Recognizing Chinese Person Names based on Hybrid Models”. Advanced Intelligence, vol. 3, Number 2, July 2011, pp.219-228.

[MAN08] Manguinhas, H. M. A.; Martins, B. E. G.; Borbinha, J. “A geo-temporal Web gazetteer service integrating data from multiple sources”. *IEEE International Conference on Digital Information Management*, Nov 2008, pp. 146-153.

[MANN08] Mann, G. S.; Mccallum, A. “Generalized expectation criteria for semi-supervised learning of conditional random fields”. In: Human Language Technology e Association of Computational Linguistics (HLT/ACL), 2008, pp.870-878.

[MAR09] Martins, B. “Geographically aware Web text mining”, Tese de Doutorado, Faculdade de Ciências, Universidade de Lisboa, 2009, 155-157p.

[MCC00] McCallum, A.; Freitag, D.; Pereira, F. “Maximum entropy Markov models for information extraction and segmentation”. In: International Conference on Machine Learning, 2000, pp. 591-598.

[MCC03] McCallum, A.; Li, W. “Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons”. In: 7th Conference on Computational Natural Language Learning, 2003, pp. 188–191.

[MAN08] Mansouri, A.; Affendey, L. S.; Mamat, A. “Named Entity Recognition Approaches”. *IJCSNS International Journal of Computer Science and Network Securit*y, vol.8, February 2008, pp. 339-344.

[MAN99] Manning, C. D.; Schütze, H. “Foundations of statistical natural language processing”. Cambridge Massachusetts: MIT Press, 1999.

[MOT07] Mota, C.; Santos, D.; Ranchhod, E. “Avaliação de Reconhecimento de Entidades Mencionadas: Princípio de HAREM”, cap. 14, 2007, 161–176p.

[MOT08] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”, 2008, 277-286p.

[MOT08a] Mota, C.; Diana, S. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, ISBN: 978-989-20-1656-6 2008.

[MOT08b] Mota, C.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 347-354p.

[MOT08c] Mota, C.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 11-31p.

[MOT08d] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 195-211p.

[MOT08e] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 231-245p.

[MOT08f] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 213-229p.

[MOT08g] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 261-274p.

[MOT08h] Mota, C. M.; Santos, D. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Linguateca, 2008, 181-193p.

[MUC6] Coreference task definition. In *Proceedings of 6th Message Understanding Conference - MUC-6*, 1995. Capturado em: http://cs.nyu.edu/faculty/grishman/muc6.html, Maio 2012.

[MUC6a] MUC -6. “The 6th in a series of Message Understanding Conferences”. Capturado em: http://cs.nyu.edu/cs/faculty/grishman/muc6.html, Maio 2012.

[MUC7] MUC-7. “MUC-7 Named Entity Task Definition”. Capturado em: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/ne_task.html, Junho 2012.

[MUC7a] MUC-7. “Overview of MUC-7/MET-2”. Capturado em: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html, Junho 2012.

[MUC7b] MUC-7. “MUC-7 Coreference Task Definition”. Capturado em: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html, Junho 2012.

[MUN12] Munkhdalai, T.; Li, M.; Kim, T.; Namsrai, O.; Jeong, S.; Shin, J.; Ryu, K. H. “Bio Named Entity Recognition based on Co-training Algorithm”. In: 26th International Conference on Advanced Information Networking and Applications Workshops (AINA), IEEE, 2012, pp. 857-862.

[NAD07] Nadeau, D.; Sekine, S. “A survey of named entity recognition and classification”. *Journal Linguisticae Investigationes, National Research Council*, vol. 30, 2007, pp. 3-26.

[PIE97] Pietra, D. S.; Pietra, D. V.; Lafferty, J. “Inducing features of random fields”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, 1997, pp. 380–393.

[PIN03] Pinto, D.; McCallum, A.; Wei, X.; Croft, W. B. “Table extraction using conditional random fields”. In: 26th Annual International Conference on Research and Development in Informaion Retrieval ACM SIGIR, 2003, pp. 235-242.

[RAT09] Ratinov, L.; Roth, D. “Design Challenges and Misconceptions in Named Entity Recognition”. In: 13th Conference on Computational Natural Language Learning, CONLL, 2009, pp. 147-155.

[SAN07] Santos, D.; Cardoso, N. “Reconhecimento de Entidades Mencionadas em Português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, 2007, 245–282p.

[SAN07a] Santos, D.; Cardoso, N. “Reconhecimento de Entidades Mencionadas em Português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, 2007, 43–57p.

[SAN07b] Santos, D.; Cardoso, N. “Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área, 2008, 1-16p.

[SAN09] Santos, D. “Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva”. Linguamática, 2009, 25-59p.

[SAN09] Santos, D.; Cabral, L. M. “GikiCLEF: Crosscultural issues in an international setting: asking non-english-centered questions to wikipedia”. Cross Language Evaluation Forum: Working notes for CLEF 2009.

[SAR06] Sarmento, L.; Pinto, A. S.; Cabral, L. “REPENTINO – A wide-scope gazetteer for entity recognition in Portuguese”. In: 7th International Workshop of Computational Processing of the Portuguese Language, 2006, pp. 31–40.

[SCH94] Schmid, H. “Probabilistic part-of-speech tagging using decision trees”. In: International Conference on New Methods in Language Processing, 1994, pp. 44-49.

[SET04] Settles, B. “Biomedical named entity recognition using conditional random fields and rich feature sets”. In: 4th International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, 2004, pp. 104–107.

[SHA03] Sha, F.; Pereira, F. “Shallow parsing with conditional random fields”. In: 3rd Proceedings of Human Language Technology, NAACL 2003, pp. 134-141.

[SOU12] Souza, M. “Mineração de opinião aplicada a mídias sociais”. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2012.

[SUX08] Suxiang, Z. “Based Cascaded Conditional Random Fields Model for Chinese Named Entity Recognition”. In: 9th International Conference on Signal Processing Proceedings (ICSP), 2008, pp.1573 – 1577.

[SUA11] Suakkaphong, N.; Zhang, Z.; Chen, H. “Disease Named Entity Recognition Using Semisupervised Learning and Conditional Random Fields”. *Journal of the American Society for Information Science and Technology*, vol. 62, April 2011, pp. 727-737.

[SUR09] Ashish S.; Pranav, P. M.; Kishore, I. V. “Polarity Classification of Subjective Words Using Common-Sense Knowledge-Base”. In: 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, 2009, pp. 486-493.

[SUT09] Sutton, C.; Mccallum, A. “Piecewise training for structured prediction”. *Journal Machine Learning Arquive*, vol. 77, ed. 2-3, 2009, pp. 165-194.

[WHI08] Casey W.; Alex K.; Nemanja P. “Web-Scale Named Entity Recognition”. In: 17th Conference on Information and Knowledge Management, 2008, pp. 123-132.

APÊNDICES

Apêndice A: exemplo de todas as *features,* aplicadas no CRF, de acordo com a sentença retirada da CD do Segundo HAREM, conforme a tabela abaixo.

“Os EUA ganharam um interesse...”

Features	**Os**	**EUA**	**ganharam**	**um**	**interesse**
1: ‘tag’	**‘art’**	**‘prop’**	**‘v’**	**‘art’**	**‘v’**
2: ‘word’	**‘Os’**	**‘EUA’**	**‘ganharam’**	**‘um’**	**‘interesse’**
3: ‘prevW’	**-**	**‘Os’**	**‘EUA’**	**‘ganharam’**	**‘um’**
4: ‘prevT’	**-**	**‘art’**	**‘prop’**	**‘v’**	**‘art’**
5: ‘prevCap’	**-**	**‘maxmin’**	**‘max’**	**‘min’**	**‘min’**
6: ‘prev2W’	**-**	**-**	**‘Os’**	**‘EUA’**	**‘ganharam’**
7: ‘prev2T’	**-**	**-**	**‘art’**	**‘prop’**	**‘v’**
8: *‘prev2Cap’*	**-**	**-**	**‘maxmin’**	**‘max’**	**‘min’**
9: ‘nextW’	**‘EUA’**	**‘ganharam’**	**‘um’**	**‘interesse’**	**...**
10: ‘nextT’	**‘prop’**	**‘v’**	**‘art’**	**‘v’**	**...**
11: ‘nextCap’	**‘ma’**	**‘min’**	**‘min’**	**‘min’**	**...**
12: ‘next2W’	**‘ganharam’**	**‘um’**	**‘interesse’**	**...**	**...**
13: ‘next2T’	**‘v’**	**‘art’**	**‘v’**	**...**	**...**
14: ‘next2Cap’	**‘min’**	**‘min’**	**‘min’**	**...**	**...**
15: ‘cap’	**‘maxmin’**	**‘max’**	**‘min’**	**‘min’**	**‘min’**
16: ‘ini’	**‘max’**	**‘max’**	**‘min’**	**‘min’**	**‘min’**
17: ‘simb’	**‘alfa’**	**‘alfa’**	**‘alfa’**	**‘alfa’**	**‘alfa’**

Apêndice B: exemplo de vetor completo de entrada para o sistema com todas essas *features*. Texto hub-30518.txt (iteração 1).

{História<prop,O-> do<v-pcp,O-> RFID<prop,O-> A<art,O-> tecnologia<n,O-> de<prp,O-> RFID<prop,U-PESSOA> tem<v-fin,O-> suas<pron-det,O-> raízes<n,O-> nos<pron-pers,O-> sistemas<n,O-> de<prp,O-> radares<n,O-> utilizados<v-pcp,O-> na<v-fin,O-> Segunda<prp,B-ACONTECIMENTO> Guerra<prop,I-ACONTECIMENTO> Mundial<adj,L-ACONTECIMENTO> .}

Apêndice C: exemplo de um vetor de features aplicadas no Experimento 1 de acordo com a sentença **“Os EUA ganharam um interesse acrescido pelas armas não letais após a sua desastrosa missão pacificadora na Somália”.**

{'nextCap': 'max', 'cap': 'maxmin', 'word': 'os', 'next2W': 'ganharam', 'next2T': 'v-fin', 'tag': 'art', 'nextT': 'prop', 'simb': 'alfa', 'nextW': 'eua', 'next2Cap': 'min', 'ini': 'max'}

{'nextCap': 'min', 'cap': 'max', 'word': 'eua', 'prevCap': 'maxmin', 'next2W': 'um', 'next2T': 'art', 'prevT': 'art', 'prevW': 'os', 'nextT': 'v-fin', 'simb': 'alfa', 'nextW': 'ganharam', 'next2Cap': 'min', 'tag': 'prop', 'ini': 'max'}

{'nextW': 'um', 'nextCap': 'min', 'cap': 'min', 'word': 'ganharam', 'prevCap': 'max', 'next2W': 'interesse', 'next2T': 'n', 'prevT': 'prop', 'prevW': 'eua', 'nextT': 'art', 'simb': 'alfa', 'tag': 'v-fin', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'maxmin', 'prev2W': 'os', 'prev2T': 'art'}

{'nextW': 'interesse', 'nextCap': 'min', 'cap': 'min', 'word': 'um', 'prevCap': 'min', 'next2W': 'acrescido', 'next2T': 'v-pcp', 'prevT': 'v-fin', 'prevW': 'ganharam', 'nextT': 'n', 'simb': 'alfa', 'tag': 'art', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'max', 'prev2W': 'eua', 'prev2T': 'prop'}

{'nextW': 'acrescido', 'nextCap': 'min', 'cap': 'min', 'word': 'interesse', 'prevCap': 'min', 'next2W': 'pelas', 'next2T': 'adv', 'prevT': 'art', 'prevW': 'um', 'nextT': 'v-pcp', 'simb': 'alfa', 'tag': 'n', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'ganharam', 'prev2T': 'v-fin'}

{'nextW': 'pelas', 'nextCap': 'min', 'cap': 'min', 'word': 'acrescido', 'prevCap': 'min', 'next2W': 'armas', 'next2T': 'n', 'prevT': 'n', 'prevW': 'interesse', 'nextT': 'adv', 'simb': 'alfa', 'tag': 'v-pcp', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'um', 'prev2T': 'art'}

{'nextW': 'armas', 'nextCap': 'min', 'cap': 'min', 'word': 'pelas', 'prevCap': 'min', 'next2W': 'n\xc3\xa3o', 'next2T': 'adv', 'prevT': 'v-pcp', 'prevW': 'acrescido', 'nextT': 'n', 'simb': 'alfa', 'tag': 'adv', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'interesse', 'prev2T': 'n'}

{'nextW': 'n\xc3\xa3o', 'nextCap': 'min', 'cap': 'min', 'word': 'armas', 'prevCap': 'min', 'next2W': 'letais', 'next2T': 'adj', 'prevT': 'adv', 'prevW': 'pelas', 'nextT': 'adv', 'simb': 'alfa', 'tag': 'n', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'acrescido', 'prev2T': 'v-pcp'}

{'nextW': 'letais', 'nextCap': 'min', 'cap': 'min', 'word': 'n\xc3\xa3o', 'prevCap': 'min', 'next2W': 'ap\xc3\xb3s', 'next2T': 'prp', 'prevT': 'n', 'prevW': 'armas', 'nextT': 'adj', 'simb': 'simb', 'tag': 'adv', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'pelas', 'prev2T': 'adv'}

{'nextW': 'ap\xc3\xb3s', 'nextCap': 'min', 'cap': 'min', 'word': 'letais', 'prevCap': 'min', 'next2W': 'a', 'next2T': 'art', 'prevT': 'adv', 'prevW': 'n\xc3\xa3o', 'nextT': 'prp', 'simb': 'alfa', 'tag': 'adj', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'armas', 'prev2T': 'n'}

{'nextW': 'a', 'nextCap': 'min', 'cap': 'min', 'word': 'ap\xc3\xb3s', 'prevCap': 'min', 'next2W': 'sua', 'next2T': 'pron-det', 'prevT': 'adj', 'prevW': 'letais', 'nextT': 'art', 'simb': 'simb', 'tag': 'prp', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'n\xc3\xa3o', 'prev2T': 'adv'}

{'nextW': 'sua', 'nextCap': 'min', 'cap': 'min', 'word': 'a', 'prevCap': 'min', 'next2W': 'desastrosa', 'next2T': 'adj', 'prevT': 'prp', 'prevW': 'ap\xc3\xb3s', 'nextT': 'pron-det', 'simb': 'alfa', 'tag': 'art', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'letais', 'prev2T': 'adj'}

{'nextW': 'desastrosa', 'nextCap': 'min', 'cap': 'min', 'word': 'sua', 'prevCap': 'min', 'next2W': 'miss\xc3\xa3o', 'next2T': 'n', 'prevT': 'art', 'prevW': 'a', 'nextT': 'adj', 'simb': 'alfa', 'tag': 'pron-det', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'ap\xc3\xb3s', 'prev2T': 'prp'}

{'nextW': 'miss\xc3\xa3o', 'nextCap': 'min', 'cap': 'min', 'word': 'desastrosa', 'prevCap': 'min', 'next2W': 'pacificadora', 'next2T': 'adj', 'prevT': 'pron-det', 'prevW': 'sua', 'nextT': 'n', 'simb': 'alfa', 'tag': 'adj', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'a', 'prev2T': 'art'}

{'nextW': 'pacificadora', 'nextCap': 'min', 'cap': 'min', 'word': 'miss\xc3\xa3o', 'prevCap': 'min', 'next2W': 'na', 'next2T': 'adj', 'prevT': 'adj', 'prevW': 'desastrosa', 'nextT': 'adj', 'simb': 'simb', 'tag': 'n', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'sua', 'prev2T': 'pron-det'}

{'nextW': 'na', 'nextCap': 'min', 'cap': 'min', 'word': 'pacificadora', 'prevCap': 'min', 'next2W': 'som\xc3\xa1lia', 'next2T': 'prop', 'prevT': 'n', 'prevW': 'miss\xc3\xa3o', 'nextT': 'adj', 'simb': 'alfa', 'tag': 'adj', 'next2Cap': 'maxmin', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'desastrosa', 'prev2T': 'adj'}

{'nextW': 'som\xc3\xa1lia', 'nextCap': 'maxmin', 'cap': 'min', 'word': 'na', 'prevCap': 'min', 'next2W': ',', 'next2T': 'punc', 'prevT': 'adj', 'prevW': 'pacificadora', 'nextT': 'prop', 'simb': 'alfa', tag': 'adj', 'next2Cap': 'min', 'ini': 'min', 'prev2Cap': 'min', 'prev2W': 'miss\xc3\xa3o', 'prev2T': 'n'}

{'nextW': ',', 'nextCap': 'min', 'cap': 'maxmin', 'word': 'som\xc3\xa1lia', 'prevCap': 'min', 'next2W': 'em', 'next2T': 'prp', 'prevT': 'adj', 'prevW': 'na', 'nextT': 'punc', 'simb': 'simb', 'tag': 'prop', 'next2Cap': 'min', 'ini': 'max', 'prev2Cap': 'min', 'prev2W': 'pacificadora', 'prev2T': 'adj'}

Apêndice D: Análise de erros de classificação pelo NERP-CRF.

Classificação pelo NERP-CRF	Classificação pelo Segundo HAREM	Ident.	Justificativa do Erro
RFID<prop,L-ORGANIZACAO>	RFID<prop,U-COISA>	errado	Identificação pelo BILOU: L ao invés de U
em<prp,O-OUT>	em<prp,B-TEMPO>	errado	Identificação do BILOU
1937<num,U-TEMPO>	1937<num,L-TEMPO>	errado	
IFF<prop,U-ORGANIZACAO>	IFF<prop,U-COISA>	errado	Erro de classificação: sigla é algo difícil de categorizar.
Identify<prop,B-PESSOA>	Identify<prop,B-COISA>	errado	Identificação BILOU e erro de classificação: palavra estrangeira.
Friend<prop,L-PESSOA>	Friend<prop,I-COISA>	errado	
or<adj,O-OUT>	or<adj,I-COISA>	errado	
Foe<prop,L-ABSTRACCAO>	Foe<prop,L-COISA>	errado	
Friendly<prop,U-LOCAL>	Friendly<prop,U-ABSTRACCAO>	errado	Palavra estrangeira.
RF<prop,U-ORGANIZACAO>	RF<prop,U-COISA>	errado	Sigla.
Radio<prop,B-PESSOA>	Radio<prop,B-COISA>	errado	Palavra estrangeira
Frequency<prop,L-PESSOA>	Frequency<prop,L-COISA>	errado	
através<adv,O-OUT>	através<adv,B-TEMPO>	errado	BILOU: U ao invés de B (palavra através). Classificação: a categoria Tempo é difícil de classificar.
das<v-pcp,O-OUT>	das<v-pcp,I-TEMPO>	errado	
décadas<v-pcp,O-OUT>	décadas<v-pcp,I-TEMPO>	errado	
de<prp,O-OUT>	de<prp,I-TEMPO>	errado	
50<num,U-TEMPO>	50<num,I-TEMPO>	errado	
e<conj-c,I-VALOR>	e<conj-c,I-TEMPO>	errado	
60<num,L-TEMPO>	60<num,L-TEMPO>	certo	
João<prop,U-PESSOA>	João<prop,U-LOCAL>	errado	Pouca indicação de local na janela analisada, ou seja, pouco contexto.
até<prp,O-OUT>	até<prp,B-TEMPO>	errado	BILOU: B e L ao invés de O, respectivamente. Classificação: a categoria Tempo é difícil de classificar.
hoje<adv,O-OUT>	hoje<adv,L-TEMPO>		

Classificação pelo NERP-CRF	Classificação pelo Segundo HAREM	Ident.	Justificativa do Erro
Às<adv,O-OUT>	Às<adv,B-TEMPO>	errado	Identificação pelo BILOU e erro de Classificação: a categoria Tempo é difícil de classificar.
três<num,O-OUT>	três<num,I-TEMPO>		
horas<n,O-OUT>	horas<n,I-TEMPO>		
e<conj-c,O-OUT>	e<conj-c,I-TEMPO>		
vinte<num,O-OUT>	vinte<num,I-TEMPO>		
minutos<n,L-VALOR>	minutos<n,I-TEMPO>		
da<v-pcp,O-OUT>	da<v-pcp,I-TEMPO>		
tarde<adv,O-OUT>	tarde<adv,L-TEMPO>		
Teatro<prop,B-LOCAL>	Teatro<prop,B-LOCAL>	certo	Erro de Classificação: a preposição (de) faz parte dessa EN.
de<prp,I-ORGANIZACAO>	de<prp,I-LOCAL>	errado	
São<prop,I-ORGANIZACAO>	São<prop,I-LOCAL>	errado	
Carlos<prop,L-LOCAL>	Carlos<prop,L-LOCAL>	certo	
Diário<n,O-OUT>	Diário<n,B-ORGANIZACAO>	errado	Identificação pelo BILOU e erro de Classificação: pouca indicação de local na janela analisada, ou seja, pouco contexto.
de<prp,I-OBRA>	de<prp,I-ORGANIZACAO>	errado	
Notícias<n,L-OBRA>	Notícias<n,L-ORGANIZACAO>	errado	
Joaquim<prop,B-PESSOA>	Joaquim<prop,B-PESSOA>	certo	Erro de Classificação a preposição ‘de’ foi o motivo do erro.
de<prp,I-LOCAL>	de<prp,I-PESSOA>	errado	
Seabra<prop,I-ORGANIZACAO>	Seabra<prop,I-PESSOA>	errado	
Pessoa<prop,L-PESSOA>	Pessoa<prop,L-PESSOA>	certo	
38<num,U-TEMPO>	38<num,U-VALOR>	errado	Erro de Classificação: pouca indicação de local na janela analisada, ou seja, pouco contexto.
D<prop,B-PESSOA>	D<prop,B-PESSOA>	certo	Identificação pelo BILOU: o ponto indicando abreviatura de uma palavra pareceu mostrar que havia uma nova EN.
.<punc,I-PESSOA>	.<punc,I-PESSOA>	certo	
Maria<prop,B-PESSOA>	Maria<prop,I-PESSOA>	errado	
Magdalena<prop,I-PESSOA>	Magdalena<prop,I-PESSOA>	certo	
Pinheiro<prop,I-PESSOA>	Pinheiro<prop,I-PESSOA>	certo	
Nogueira<prop,I-PESSOA>	Nogueira<prop,I-PESSOA>	certo	
Pessoa<prop,L-PESSOA>	Pessoa<prop,L-PESSOA>	certo	
Ilha<prop,B-ORGANIZACAO>	Ilha<prop,B-LOCAL>	errado	Erro de classificação: contexto fraco.

Classificação pelo NERP-CRF	Classificação pelo Segundo HAREM	Ident.	Justificativa do Erro
Emília<prop,L-ABSTRACCAO>	Emília<prop,U-PESSOA>	errado	Erro de identificação BILOU e de classificação: contexto fraco.
em<prp,O-OUT>	em<prp,B-TEMPO>	errado	Identificação pelo BILOU e erro de Classificação: a categoria Tempo é difícil de classificar.
21<num,B-TEMPO>	21<num,I-TEMPO>	errado	
de<prp,I-TEMPO>	de<prp,I-TEMPO>	certo	
Julho<n,L-TEMPO>	Julho<n,L-TEMPO>	certo	
Igreja<v-fin,O-OUT>	Igreja<v-fin,B-LOCAL>	errado	Identificação do BILOU.
dos<n,O-OUT>	dos<n,I-LOCAL>	errado	
Mártires<prop,U-PESSOA>	Mártires<prop,L-LOCAL>	errado	
GeRENal<prop,B-PESSOA>	GeRENal<prop,B-PESSOA>	certo	Erro de identificação do BILOU, mas classificou certo.
Chaby<prop,I-PESSOA>	Chaby<prop,L-PESSOA>	errado	
Fernando<prop,B-PESSOA>	Fernando<prop,B-PESSOA>	certo	Erro de Identificação do BILOU para preposição “de” (esse faz parte da EM) e erro classificação.
de<prp,O-OUT>	de<prp,I-PESSOA>	errado	
Bulhões<prop,L-LOCAL>	Bulhões<prop,L-PESSOA>	errado	
dia<n,O-OUT>	dia<n,U-TEMPO>	errado	Erro de identificação do BILOU e de classificação da categoria Tempo.(A categoria Tempo é difícil de classificar, pois algumas palavras que indicam Tempo não iniciam com letra maiúscula. Logo não são entendidas pelo sistema como EN.)
do<num,O-OUT>	do<num,B-TEMPO>	errado	Classificação: EN de Tempo difícil de classificar. (Ocorreram muito erros desse tipo como erro de classificação da categoria Tempo.)
dia<n,O-OUT>	dia<n,L-TEMPO>		
Chevalier<prop,B-PESSOA>	Chevalier<prop,B-PESSOA>	certo	Erro de Indentificação BILOU e pouca indicação de local na janela analisada, ou seja, pouco contexto para classificar corretamente a EN.
de<prp,I-LOCAL>	de<prp,I-PESSOA>	errado	
Pas<prop,L-LOCAL>	Pas<prop,L-PESSOA>	errado	
Fernando<prop,B-PESSOA>	Fernando<prop,B-PESSOA>	certo	Erro de Indentificação BILOU.
Pessoa<adv,O-OUT>	Pessoa<adv,L-PESSOA>	errado	

Classificação pelo NERP-CRF	Classificação pelo Segundo HAREM	Ident.	Justificativa do Erro
Adolfo<prop,B-PESSOA>	Adolfo<prop,B-PESSOA>	certo	Erro de Indentificação BILOU.
Casais<prop,B-PESSOA>	Casais<prop,I-PESSOA>	errado	
Monteiro<prop,L-PESSOA>	Monteiro<prop,L-PESSOA>	certo	
À<adv,O-OUT>	À<adv,B-OUTRO>	errado	Identificação e classificação:EN estranha.
Minha<prop,B-PESSOA>	Minha<prop,I-OUTRO>	errado	
Querida<v-pcp,I-PESSOA>	Querida<v-pcp,I-OUTRO>	errado	
Mamãe<prop,L-ABSTRACCAO>	Mamãe<prop,L-OUTRO>	errado	
Portugal<prop,U-ORGANIZACAO>	Portugal<prop,U-LOCAL>	errado	Classificação: nome de país. (alguns nomes de países são confusos de classificar corretamente. O contexto deveria oferecer mais indicativos sobre a EN para que ela seja classificada corretamente.)
Pessoa<prop,U-ORGANIZACAO>	Pessoa<prop,U-PESSOA>	errado	Erro de classificação: contexto fraco.
Manuel<prop,B-PESSOA>	Manuel<prop,B-PESSOA>	certo	Erro de Identificação do BILOU e erro classificação.
Gualdino<prop,L-PESSOA>	Gualdino<prop,I-PESSOA>	errado	
da<v-pcp,I-LOCAL>	da<v-pcp,I-PESSOA>	errado	
Cunha<prop,L-LOCAL>	Cunha<prop,L-PESSOA>	errado	
Hawarden<prop,L-PESSOA>	Hawarden<prop,B-COISA>	errado	Entidades Nomeadas de nomes estrangeiros.
Castle<prop,L-PESSOA>	Castle<prop,L-COISA>		
O<art,O-OUT>	O<art,B-OBRA>	errado	Erro de classificação: contexto fraco.
Corvo<prop,U-ORGANIZACAO>	Corvo<prop,L-OBRA>	errado	
Antinous<prop,U-LOCAL>	Antinous<prop,U-OBRA>	errado	Erro de classificação: contexto fraco e difícil.
Em<prp,O-OUT>	Em<prp,B-TEMPO>	errado	Identificação pelo BILOU e erro de Classificação: a categoria Tempo é difícil de classificar.
1899<num,U-TEMPO>	1899<num,L-TEMPO>	errado	
Durban<prop,B-LOCAL>	Durban<prop,B-ORGANIZACAO>	errado	Erro de classificação: palavra estrangeira.
High<prop,I-LOCAL>	High<prop,I-ORGANIZACAO>		
School<prop,L-PESSOA>	School<prop,L-ORGANIZACAO>		

Classificação pelo NERP-CRF	Classificação pelo Segundo HAREM	Ident.	Justificativa do Erro
Alexander<prop,B-PESSOA>	Alexander<prop,B-ABSTRACCAO>	errado	Erro de classificação: contexto fraco e EN difícil de classificar.
Search<prop,L-PESSOA>	Search<prop,L-ABSTRACCAO>	errado	
Ilha<prop,B-PESSOA>	Ilha<prop,B-LOCAL>	errado	Erro de classificação.
Terceira<prop,L-PESSOA>	Terceira<prop,L-LOCAL>	errado	
<punc,O-OUT>	Paciência<punc,U-PESSOA>	errado	Erro de identificação e classificação: contexto fraco e EN difícil de classificar.
Na<prop,O-OUT>	Na<prop,B-TEMPO>	errado	Classificação: EN de Tempo difícil de classificar.
mesma<pron-det,O-OUT>	mesma<pron-det,I-TEMPO>		
época<n,O-OUT>	época<n,L-TEMPO>		