Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Estudo Comparativo de diferentes classificadores baseados em aprendizagem de máquina para o processo de Reconhecimento de Entidades Nomeadas

Jadson da Silva Santos

Feira de Santana

2017

Universidade Estadual de Feira de Santana

Programa de Pós-Graduação em Computação Aplicada

Jadson da Silva Santos

Estudo Comparativo de diferentes classificadores baseados em aprendizagem de máquina para o processo de Reconhecimento de Entidades Nomeadas

Dissertação apresentada à Universidade Estadual de Feira de Santana como parte dos requisitos para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. João B. Rocha-Junior

Feira de Santana

2017

Ficha Catalográfica - Biblioteca Central Julieta Carteado

S235e Santos, Jadson da Silva
Estudo comparativo de diferentes classificadores baseados em aprendizagem de máquina para o processo de Reconhecimento de Entidades Nomeadas / Jadson da Silva Santos. Feira de Santana – 2017.
81 f.: il.

Orientador: João B. Rocha-Júnior.
Dissertação (mestrado) - Universidade Estadual de Feira de Santana, Programa de Pós-Graduação em Computação Aplicada, 2017.

1. Reconhecimento de Entidades Nomeadas (REN). 2. Processamento de linguagem natural (Computação). I. Rocha- Júnior, João B., orient. II. Universidade Estadual de Feira de Santana. III. Título.

CDU: 004.91

Abstract

The Named Entity Recognition (NER) process is the task of identifying relevant terms in texts and assigning them a label. Such words can reference names of people, organizations, and places. The variety of techniques that can be used in the named entity recognition process is large. The techniques can be classified into three distinct approaches: rule-based, machine learning and hybrid. Concerning to the machine learning approaches, several factors may influence its accuracy, including the selected classifier, the set of features extracted from the terms, the characteristics of the textual bases, and the number of entity labels. In this work, we compared classifiers that use machine learning applied to the NER task. The comparative study includes classifiers based on CRF (Conditional Random Fields), MEMM (Maximum Entropy Markov Model) and HMM (Hidden Markov Model), which are compared in two corpora in Portuguese derived from WikiNer, and HAREM, and two corporas in English derived from CoNLL-03 and WikiNer. The comparison of the classifiers shows that the CRF is superior to the other classifiers, both with Portuguese and English texts. This study also includes the comparison of the individual and joint contribution of features, including contextual features, besides the comparison of the NER per named entity labels, between classifiers and corpora.

Keywords: Named Entity Recognition, Machine Learning, Information Extraction, Natural Linguage Process.

Resumo

O processo de Reconhecimento de Entidades Nomeadas (REN) é a tarefa de identificar termos relevantes em textos e atribuí-los um rótulo. Tais palavras podem referenciar nomes de pessoas, organizações e locais. A variedade de técnicas que podem ser usadas no processo de reconhecimento de entidades nomeadas é grande. As técnicas podem ser classificadas em três abordagens distintas: baseadas em regras, baseadas em aprendizagem de máquina e híbridas. No que diz respeito as abordagens de aprendizagem de máquina, diversos fatores podem influenciar sua exatidão, incluindo o classificador selecionado, o conjunto de *features* extraídas dos termos, as características das bases textuais e o número de rótulos de entidades. Neste trabalho, comparamos classificadores que utilizam aprendizagem de máquina aplicadas a tarefa do REN. O estudo comparativo inclui classificadores baseados no CRF (Condicional Random Fields), MEMM (Maximum Entropy Markov Model) e HMM (Hidden Markov Model), os quais são comparados em dois corporas em português derivados do WikiNer, e HAREM, e dois corporas em inglês derivados doCoNLL-03 e WikiNer. A comparação dos classificadores demonstra que o CRF é superior aos demais classificadores, tanto com textos em português, quanto inglês. Este estudo também inclui a comparação da contribuição, individual e em conjunto de *features*, incluindo *features* de contexto, além da comparação do REN por rótulos de entidades nomeadas, entre os classificadores e os corpora.

Palavras-chave: Reconhecimento de Entidades Nomeadas, Aprendizagem de Máquina, Extração de Informação, Processamento de Linguagem Natural.

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Estadual de Feira de Santana (UEFS) como requisito parcial para obtenção do grau de Mestre em Computação Aplicada.

A dissertação foi desenvolvida dentro do Programa de Pós-Graduação em Computação Aplicada (PGCA) tendo como orientador o Dr. **João B. Rocha-Junior**.

Esta pesquisa foi parcialmente financiada pela Fapesb.

Agradecimentos

Agradeço em primeiro lugar a Deus que me sustenta e me permitiu chegar nesta etapa;

Ao professor Dr. João B. Rocha-Junior, por todo o acompanhamento e orientação que foram cruciais para o desenvolvimento da pesquisa;

A todos os professores do PGCA, que contribuíram bastante durante o perıodo do mestrado;

Ao professor Dr. Hugo Saba, que foi o responsável por me incentivar a ingressar no programa de Pos Graduação, além de sempre me oferecer apoio e orientação;

À minha família, por me apoiar desde o primeiro momento. Em especial a minha mãe que lutou a vida inteira para me proporcionar a melhor criação possível, sempre me incentivando a estudar, e de onde ela estiver eu sei que está feliz por mim. Agradeço também ao apoio de minha irmã Tâmara;

Agradeço ao apoio dos meus amigos, em especial Josualdo Dias, Felipe Andrade e Alexandre Câmara;

Agradeço também aos profs. Drs. Eduardo Jorge e Rodrigo Calumby por se disponibilizarem a analisar este trabalho.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Abrebiações

Abreviação Descrição

ACE Automatic Content Extraction
CoNLL Conference on Natural Language Learning
CRF Condicional Random Field
DT Decision Tree
EDT Entity Detection and Tracking
EI Extração de Informação (Information Extraction)
EE Event Extraction
FP False Positive
FN False Negative
FSVM Fuzzy Suport Vector Machine
HAREM Avaliação Conjunta para Sistemas de Reconhecimento de Entidades Mencionadas
HMM Hidden Markov Model
IUPAC International Union of Pure and Applied Chemistry
KNN K-Nearest Neighbors
LDC Linguistic Data Consortium
MEMM Maximum Entropy Markov Model
MBL Memory Based Learner
MUC Conference Understanding Message
NERD Named Entity Recognition and Disambiguation
NERSSEAL NER on South and South East Asian Languages
NLPBA Natural Language Processing in Biomedicine and its Applications
POS Part-of-Speech
RDC Relation Detection and Characterization
RE Relation Extraction
REN Reconhecimento de Entidades Nomeadas (Named Entity Recognition)
SVM Support Vector Machine
TP True Positive

Lista de Símbolos

Símbolo Descrição

Capítulo 1

Introdução

O uso constante da tecnologia no cotidiano das pessoas proporciona a geração de grande quantidade de documentos contendo dados textuais. Técnicas que possibilitem o processamento desses dados são utilizadas como soluções automáticas para obter informações relevantes. As informações que são geradas no formato de linguagem natural são caracterizadas como não estruturadas, significando que não possuem uma estrutura que permita a extração de informações diretamente. Textos não estruturados precisam passar por etapas de pré-processamento que possibilitem estruturar seus dados.

A linguagem natural é formada pela junção de termos (palavras) no intuito de gerar uma mensagem que carrega um significado. Entretanto cada termo de um idioma pode possuir mais de um significado, e a forma de unir palavras em uma sentença é muito complexa, devido à grande diversidade de possibilidades. Dessa maneira, analisar esse tipo de informação de forma automática não é uma tarefa simples.

Um conceito relacionado ao processamento de dados dispostos em formato de linguagem natural é o de Extração de Informação (EI) [Cowie e Lehnert 1996]. A EI visa identificar informações relevantes de um documento, ou de uma coleção, com textos em linguagem natural. A identificação destas informações relevantes facilita a manipulação e análise dos documentos de forma automática, por meio de um conjunto de padrões e regras de extração.

Termos relevantes em um texto podem estar relacionados à pessoa que está sendo citada, ao local que é comentado, à ocorrência de nomes de empresas ou instituições, entre outros exemplos. Com a extração desses termos é possível obter uma informação da mensagem que está sendo transmitida no texto, contribuindo para a identificação dos dados relevantes.

O Reconhecimento de Entidades Nomeadas (REN) é uma tarefa que faz parte dos objetivos da área de EI. Consiste na identificação e classificação de entidades nomeadas que possuem um fator de representação significante num texto. Entidades nomeadas podem ser classificadas como Pessoas, Organizações, Locais, Genes, entre

outras [Fersini et al. 2014]. Ou seja, termos que contribuem para a geração da informação. A identificação desses termos possibilita o desenvolvimento de processos que visam, justamente, a extração da informação gerada.

Quando se trata de técnicas utilizadas para o REN, três tipos de abordagens distintas são conhecidas: sistemas baseados em regras, sistemas baseados em aprendizado de máquina e métodos híbridos [Liu et al. 2011]. No caso dos sistemas baseados em regras, conhecidos também como sistemas baseados no conhecimento, busca-se utilizar heurísticas no formato de padrões linguísticos ou expressões. Os sistemas baseados em aprendizado de máquina utilizam técnicas e algoritmos para permitir que o computador aprenda como identificar as entidades. Enquanto os métodos híbridos utilizam as duas técnicas para realizar o REN.

Um exemplo que ilustra a ocorrência de entidades nomeadas em uma sentença é apresentado na Figura 1.1. O REN não apenas identifica os termos que são entidades, mas também categoriza esses termos em rótulos de entidades. Neste exemplo, o termo Carlos é classificado como uma pessoa, enquanto o termo Madrid é classificado como uma cidade, um local.

<Pessoa>**Carlos**</Pessoa> viajou para <Local>**Madrid**</Local>.

Figura 1.1: Exemplo de REN em uma sentença.

As técnicas de reconhecimento de entidades nomeadas buscam adquirir características dos termos, que possibilitem identificá-los e classificá-los como entidades nomeadas. Em abordagens que fazem uso de aprendizagem de máquina, essas características também conhecidas como *features*, são utilizadas como entrada para os classificadores de entidades nomeadas [Tkachenko e Simanovsky 2012].

Muitas abordagens já foram desenvolvidas para a tarefa de identificar e classificar as entidades nomeadas de bases textuais [Nadeau e Sekine 2007]. A análise de algumas dessas técnicas e modelos, assim como o estudo comparativo dessas abordagens é relevante para auxiliar a compreensão de distintas técnicas de REN, expondo suas características e particularidades. Tais avaliações também facilitam a escolha de qual técnica, ou conjunto de técnicas utilizar em diferentes cenários e necessidades.

1.1 Objetivo

Os classificadores que fazem parte deste estudo comparativo são baseados em aprendizagem de máquina. São eles HMM (*Hidden Markov Model*), MEMM (*Maximum Entropy Markov Model*) e CRF (*Condicional Random Field*). O objetivo deste trabalho é: comparar esses classificadores, inclusive quando submetidos a textos em português; estudar a contribuição das *features*, individualmente e em conjunto; observar diferenças no REN com as distintas bases de dados; e as relações de acurácia

entre os diferentes rótulos de entidades nomeadas. Neste estudo, as técnicas são comparadas levando-se em consideração os mesmos parâmetros de entrada, bases de dados e rótulos.

1.1.1 Objetivos Específicos

- Colocar os classificadores CRF, MEMM e HMM em um mesmo framework que possibilite compará-los;
- Definir o conjunto de dados de entrada para os classificadores;
- Identificar as melhores métricas para comparar a acurácia dos classificadores;
- Fazer um estudo comparativo dos classificadores, observando diversos parâmetros de entrada.

1.2 Justificativa e Relevância

Diferentes técnicas para extrair e classificar entidades nomeadas são conhecidas, utilizando sistema baseados em regras, sistema baseado em aprendizagem de máquina e sistema híbrido. Um estudo comparativo dos classificadores em um mesmo ambiente de avaliação, permite verificar o desempenho de cada um, em diferentes cenários.

Dada a diversidade de classificadores usados para o REN, é importante conhecer não apenas as características de cada uma possui, mas também como se comportam quando são testadas em diferentes condições. O estudo comparativo de classificadores de REN preocupa-se em obter informações sobre seus respectivos resultados, por meio da utilização de diferentes testes, em bases de dados e categorias distintas de entidades nomeadas. Tudo isso utilizando um mesmo framework para avaliação.

O avanço das pesquisas no Reconhecimento de Entidades Nomeadas está relacionado com a diversidade de técnicas que são desenvolvidas. Estudos comparativos de classificadores de REN são usados para auxiliar o entendimento dessas diferentes abordagens, disponibilizando informações a respeito da eficácia de cada um, tornando-se crucial para ajudar na escolha do classificador mais adequado.

A relevância dessa proposta de pesquisa está associada à obtenção de informações sobre técnicas de REN. Devido a diversidade de técnicas apresentadas nos trabalhos de REN publicados nos últimos anos, identificar as características dessas técnicas e compará-las possibilita o entendimento do avanço e direcionamento que essa área de pesquisa tem alcançado.

A comparação realizada nesta dissertação apresenta fatores diferenciais de relevância em: 1) definir a metodologia para comparar os classificadores em diferentes cenários, 2) estudar o impacto da seleção de *features* de entrada e suas combinações

nos classificadores e 3) avaliar o desempenho dos classificadores para corpora em português.

A definição da metodologia aplicada implica na utilização de distintos corpora. Os corpora selecionados para treinar e testar os classificadores possuem características distintas, que permitem observar como cada classificador se comporta, quando há mais e menos dados para as fases de treinamento e teste. Além de comparar a eficácia dos classificadores por rótulo de entidade nomeada, relacionando os resultados dos testes com as características dos corpora.

As *features* são avaliadas, quanto a sua contribuição para melhorar o REN, individualmente e em conjunto. Dado que as *features*, são utilizadas como dados de entrada, permitindo que os classificadores associem um termo a um rótulo de entidade nomeada, é relevante estudar suas contribuições, e como suas combinações afetam o REN.

Pesquisas envolvendo o REN são bem estabelecidas, assim como existem estudos que realizam comparações do REN ([Nadeau e Sekine 2007] [Abacha e Zweigenbaum 2011]). Entretanto a maioria destes trabalhos fazem uso de textos em inglês. O estudo comparativo realizado neste trabalho conta com dois corpora com textos em português e dois corpora com textos em inglês. Logo, há a contribuição da comparação de classificadores, baseados em aprendizagem de máquina, para o REN em português.

Como o REN pode ser empregado de formas diversas, atendendo a distintas aplicações, a comparação realizada nesta dissertação não está associada a um domínio específico. As definições estabelecidas para comparar os classificadores, buscam trazer resultados que possam beneficiar futuros trabalhos de REN, de forma generalizada.

Muitas abordagens existentes na literatura utilizam um modelo híbrido que engloba diversas técnicas. Nestes trabalhos, o objetivo principal é otimizar um classificador para um problema específico. Diferente destes estudos, o nosso objetivo é comparar diferentes classificadores baseados em aprendizagem de máquina em diferentes cenários e não otimizar um classificador para um caso específico. Sendo assim, nenhum classificar híbrido é avaliado.

1.3 Aplicações

O REN, como tarefa de EI, pode ser aplicado em diversas abordagens que têm como objetivo permitir o processamento de documentos textuais não estruturados. Dada a quantidade de informações que podem ser tratadas de dados no formato de linguagem natural, o ganho do processamento desses textos é relevante, para a identificação e classificação de termos críticos, em diversas áreas, como: Biomedicina, Bioinformática e Mídias Sociais.

Zhang e Elhadad [Zhang e Elhadad 2013], fazem uso do REN para extrair entidades nomeadas de textos biomédicos, os quais incluem dados clínicos e hospitalares que podem conter nomes de doenças e medicamentos, por exemplo. Os autores informam que, devido a adoção de registros eletrônicos de saúde, há uma quantidade significativa de textos biomédicos disponíveis. Por meio da aplicação do REN neste domínio, é possível: identificar doenças e nomes de drogas em sumários de alta hospitalar; detectar ocorrências de genes e nomes de proteínas em resumos de artigos biomédicos. Tais aplicações beneficiam o processamento automático de linguagem natural biomédica.

Rocktäschel [Rocktäschel et al. 2012] aplica o REN para identificar menções de termos químicos em dados textuais. Neste trabalho é desenvolvida uma ferramenta que anota nomes relativos a entidades químicas, drogas e fórmulas moleculares. Rocktäschel destaca a importância da identificação eficaz destes termos, em aplicações que incluem: a reconstrução computadorizada de redes metabólicas e a recuperação de informações sobre substâncias, no desenvolvimento de medicamentos. O autor afirma que utilizando o REN para anotar tais termos, pode-se evitar a realização manual desta tarefa, o que exige um grupo de peritos, que sistematicamente escaneiam publicações relevantes deste domínio para a extração de informações.

Mídias sociais são utilizadas massivamente para expor informações em formato de dados textuais. O trabalho de Liu [Liu et al. 2011] aplica o REN em textos de postagens do Twitter. O autor afirma que, a maioria dos sistemas de REN são focados principalmente em textos formais, como artigos de notícias, logo não alcançam bons resultados quando submetidos a textos de postagens do Twitter. Liu combina os classificadores KNN (*K-Nearest Neighbors*) e CRF (*Conditional Random Field*) em um *framework* de aprendizagem de máquina, utilizando os rótulos *Person*, *Location*, *Organization*, *Product* e *Other*. Neste trabalho, a aplicação do REN é designada para ser eficaz na extração de informação de *tweets*, que são constituídos de poucos dados textuais e apresentam uma linguagem informal.

A identificação e rotulação de entidades nomeadas é aplicada em trabalhos que analisam os documentos textuais na Web [Speck e Ngomo 2014], dado que boa parte dos seus dados estão no formato de linguagem natural não estruturada [Iskold 2006]. A exemplo do trabalho desenvolvido por Vavliakis [Vavliakis et al. 2013], o qual integra o REN com: mapas de tópicos dinâmicos, agrupamento de tópicos e técnicas de detecção de pico. O trabalho de Vavliakis utiliza uma base de dados composta por 7 milhões de blogs da web, afim de identificar eventos divulgados na Web, com os rótulos P (pessoa), O (organizações), L (locais) e OE (outras entidades).

1.4 Organização do Trabalho

No Capítulo 2 é apresentada a fundamentação teórica relacionada ao tema da pesquisa, com tópicos relacionados ao tratamento de dados textuais não estruturados, à

tarefa de Reconhecimento de Entidades Nomeadas e às medidas adotadas para calcular a eficácia da identificação e classificação das entidades nomeadas. No Capítulo 3 são destacados trabalhos relacionados ao tema da pesquisa, destacando diferenças com o trabalho desta dissertação. O Capítulo 4 trata da metodologia de pesquisa empregada, abordando definições realizadas para execução dos experimentos e do estudo comparativo. No Capítulo 5 são apresentados os experimentos realizados para permitir comparar as distintas técnicas de REN, com aprendizagem de máquina. O Capítulo 6 apresenta discussões e conclusões obtidas neste trabalho.

Capítulo 2

Fundamentação Teórica

Este capítulo aborda temas relacionados à área desta pesquisa. Após uma breve explanação sobre EI (Seção 2.1), a Seção 2.2 aborda o pré-processamento, como etapa de preparação de dados textuais. O Reconhecimento de Entidades Nomeadas é discutido na Seção 2.3, apresentando um breve histórico do REN (Seção 2.3.1), seguido do conceito das três abordagens em que o REN pode ser aplicado (seções 2.3.2, 2.3.3, e 2.3.4). Por fim, a Seção 2.3.5 traz um resumo comparativo de técnicas utilizadas para reconhecer entidades nomeadas.

2.1 Extração de Informação

O conceito de EI está relacionado a qualquer processo que, de forma seletiva, estrutura e combina dados dispostos em um ou mais textos [Cowie e Lehnert 1996], os quais são apresentados no formato de linguagem natural não estruturada. O pré-processamento (Seção 2.2) é uma etapa crucial para preparar os dados textuais em um formato que permita a extração de informações. Logo, a EI está associada a diversas técnicas que processam dados textuais, possibilitando a seleção de termos relevantes.

Piskorski e Yangarder [Piskorski e Yangarber 2013] apontam alguns tipos de tarefas de EI, que extraem termos críticos para a geração de informação, entre as quais estão: REN, *Relation Extraction* (RE) e *Event Extraction* (EE). Tais tarefas contribuem para a estruturação de um dado textual em linguagem natural.

O REN, enquanto tarefa que atende a objetivos da EI, identifica e classifica termos críticos em um texto, possibilitando a extração de dados relevantes [Mansouri et al. 2008]. Tais dados identificam em textos, ocorrências de pessoas, locais, organizações, entre outras entidades. Por meio do REN, pode-se também extrair informações das relações entre as entidades.

A relação entre entidades nomeadas, compondo a tarefa de RE, permite identificar casos onde termos de um determinado rótulo A, estão associados a termos de um rótulo B, por exemplo. Na tarefa de RE, são detectados e classificados, relacionamentos entre entidades identificadas no texto. Como a relação de localização extraída das entidades Brasília e Brasil, por meio da frase: *'Brasília é a capital do Brasil'*.

Já a tarefa de EE, identifica eventos em textos, obtendo informações estruturadas sobre tais eventos. Idealmente identificam informações como: quem fez o que a quem, quando, onde [Piskorski e Yangarber 2013]. Normalmente, EE envolve a extração de diversas entidades as relações que possuem, sendo considerada uma das tarefas de EI mais complexas.

Tarefas de extração de informação podem atender a diversos domínios. O conjunto das estruturas extraídas (entidades nomeadas, relacionamento entre as entidades e eventos) está condicionado à tarefa aplicada e ao tipo de informação que deseja ser extraída [Nadeau 2005].

Assim como as estruturas extraídas podem variar, o tipo de base textual, não estruturada, pode variar [Sarawagi 2008]. Os dados podem ser originados de artigos de notícias, páginas na Web, receitas médicas, postagens em redes sociais, entre outros. Cada tipo de base de dados possui um conjunto de termos que correspondem às estruturas extraídas que desejam ser trabalhadas. Por exemplo, receitas médicas podem conter informações como nomes de medicamentos, dosagem, doenças, ou estado do paciente. Outro exemplo é o uso de artigos de notícias para extrair nomes de empresas, pessoas e locais.

O *Reuters Corpus* [Lewis et al. 2004] é um exemplo de base textual não estruturada, constituída de artigos de notícias, no formato de linguagem natural. Esta base é utilizada para compor os corpora, de treino e teste, disponibilizados na conferência CoNLL-03 (*Conference on Natural Language Learning*), os quais são anotados para entidades nomeadas. A etapa de anotação das entidades nomeadas requer um pré-processamento dos artigos de notícias, que separa cada termo por linha, separa as sentenças, adiciona as classes gramaticais dos termos e seus respectivos rótulos de entidade nomeada.

Sarawagi [Sarawagi 2008] descreve algumas aplicações que utilizam a EI, dentre as quais estão: rastreamento de notícias, uma aplicação clássica que rastreia de forma automática tipos específicos de eventos de fontes de notícias; serviços de atendimento ao consumidor, por meio de informações que as empresas coletam de seu clientes; sistemas de gerenciamento de informações pessoais, que organizam dados pessoais, como documentos, e-mails, entre outros em um formato interligado de forma estruturada.

Os processos, incluindo o de REN, que extraem informações, permitem diversas formas de consultas em abundantes fontes de dados não estruturados [Sarawagi 2008].

A EI auxilia ainda a possibilidade de consultas e obtenção de informação quando coexistem dados estruturados e não estruturados.

2.2 Pré-Processamento

Devido à natureza não estruturada dos textos escritos em linguagem natural, a extração de informação desse tipo de dado precisa passar por etapas que permitem processar seus conteúdos [Osenova e Kolkovska 2002]. O pré-processamento textual caracteriza-se justamente como essa etapa, seu objetivo é organizar um texto em linguagem natural, de modo que possa ser processado automaticamente.

Diversas técnicas de pré-processamento são conhecidas e utilizadas, quando se trabalha com dados textuais. As técnicas de pré-processamento como *Tokenization*, *Segmentation*, *Capitalization* e *POS tagging* estão entre as mais usadas para auxiliar o processo de extração de informação de textos [Osenova e Kolkovska 2002]. Os itens abaixo descrevem algumas das principais técnicas de pré-processamento:

- ***Tokenization*:** permite separar o texto em unidades, palavras, possibilitando identificar cada um desses termos e extrair suas informações.

  Ex: |Marcos| |recebeu| |a| |carta| |de| |sua| |mãe|.

- ***Segmentation*:** separa o texto em sentenças, separadas por ponto, o que permite selecionar cada segmento do texto.

  Ex: <s>Marcos recebeu a carta de sua mãe</s>. <s>Nenhuma resposta foi enviada</s>.

- ***Capitalization*:** possibilita identificar termos com ocorrência de letra inicial maiúscula ou palavras completamente em letras maiúsculas. Auxilia, por exemplo a identificar nomes próprios.

  Ex: <Cap>Marcos</Cap> recebeu a carta de sua mãe.

- ***POS Tagging*:** o part-of-speech tagging, etiqueta os termos quanto a sua classe gramatical, determinado se o termo é artigo, substantivo, verbo e entre outros.

  Ex: Marcos_PROP[1] recebeu_V-FIN[2] a_ART[3] carta_N[4] de_PRP[5] sua_PRON-DET[6] mãe_N.

[1]PROP - Substantivo próprio
[2]V-FIN - Verbo finito
[3]ART - Artigo
[4]N - Substantivo
[5]PRP - Preposição
[6]PRON-DET - Pronome determinativo

- ***Stemming*:** processo que coloca o termo em sua forma radical, permitindo agrupar todos os termos derivados do mesmo radical, auxilia a singularizar os termos.

  Ex: Marcos **receb**[eu] a **cart**[a] de sua mãe.

- ***Lemmatization*:** processo que torna o termo em sua forma deflexionada ou lematizada, agrupando diferentes flexões de um termo.

  Ex: Marcos receber o carta de ele mãe.

- ***Stop Words*:** identifica termos que serão desconsiderados do texto, comumente por não acrescentarem informação relevante. Exemplos desses termos são artigos e preposições.

  Ex: Marcos recebeu carta mãe.

O pré-processamento é a etapa que permite que os dados textuais sejam processados. Cada técnica de pré-processamento é aplicada de acordo com a necessidade do trabalho desenvolvido, logo não se faz obrigatória a utilização de todos os tipos de pré-processamentos de dados textuais.

Relacionado ao REN, o pré-processamento dos textos é essencial para obter informações dos termos que possibilitem identificá-los como entidade nomeada e classificá-los [Osenova e Kolkovska 2002]. A etapa do pré-processamento auxilia também na geração de *features* para sistemas baseados em aprendizagem de máquina [Kaur e Gupta 2010], e ainda apoia a criação de regras que identifiquem e classifiquem as entidades nomeadas [Nadeau e Sekine 2007].

2.3 Reconhecimento de Entidades Nomeadas

O REN [Nadeau e Sekine 2007] é a tarefa de identificar e classificar termos relevantes para a compreensão de um dado textual. Os termos que podem ser considerados entidades variam de acordo com o domínio de interesse e são comumente atribuídos a nomes que referenciam pessoas, organizações, locais, entre outros.

O estudo na área de REN é aplicado a diversas áreas e os sistemas de REN são divididos em três tipos de categorias, de acordo com a forma como o reconhecimento das entidades nomeadas é realizado: REN baseado em aprendizagem de máquina (Seção 2.3.2), que faz uso de técnicas de aprendizagem automática, por meio de treino e teste; REN baseado em regras (Seção 2.3.3), que utiliza dicionários, heurísticas e condições para reconhecer as entidades nomeadas; REN baseado em abordagem híbrida (Seção 2.3.4), as quais reúnem regras e aprendizagem de máquina [Liu et al. 2011]. A Seção 2.3.1, aborda um breve histórico do REN, enquanto a Seção 2.3.5 traz um resumo de técnicas de REN.

2.3.1 Breve histórico

O início do estudo do REN foi incentivado principalmente por conferências como, a *Message Understanding Conference* (MUC) [Grishman e Sundheim 1996] que foi a primeira conferência a lidar com a avaliação do Reconhecimento de Entidades Nomeadas na década de 1990 [Nothman et al. 2013]. Suas primeiras edições eram caracterizadas como uma avaliação conjunta na área de Extração de Informação. Tanto na sexta edição (MUC-6) quanto na sétima (MUC-7) realizados respectivamente em 1995 e 1998, houve um foco na avaliação de sistemas de Reconhecimento de Entidades Nomeadas para o inglês.

Ao longo de suas sete edições foram implementadas a identificação das entidades nomeadas e determinadas suas categorias como: nomes próprios (ENAMEX), termos temporais (TIMEX) e expressões numéricas (NUMEX) [Nadeau e Sekine 2007]. Cada uma dessas categorias possuíam subcategorias. ENAMEX com subcategorias para organização (ORGANIZATION), lugar (LOCATION) e pessoa (PERSON); TIMEX com subcategorias para medidas indicadoras de tempo (TIME) e data (DATE); por fim NUMEX podendo ser percentuais (PERCENT) e valores monetários (MONEY).

Já a conferência *Automatic Content Extraction* (ACE) [Doddington et al. 2004] surgiu em 1999, após a sétima edição do MUC (MUC-7), com o objetivo de desenvolver a capacidade de extrair informação de fontes multimídia, incluindo textos. Seus esforços foram dedicados ao Reconhecimento de Entidades Nomeadas, gerando a determinação de desafios nesta área de pesquisa, tais como: *Entity Detection and Tracking* (EDT) − para identificar as menções a entidades, seja um nome, uma descrição ou pronome; *Relation Detection and Characterization* (RDC) − detecção e caracterização de relações entre pares de entidades.

Outro evento que é bastante referenciado ao REN é a *Conference on Natural Language Learning* (CoNLL) que é uma conferência voltada ao desenvolvimento de técnicas de processamento de linguagem natural que utilizam aprendizagem de máquina. CoNLL-2002 e CoNLL-2003, foram duas edições voltadas a tarefa de REN. Sistemas de REN foram desenvolvidos para ambas edições, sendo disponibilizados corpora com entidades rotuladas, divididos em corpus de treinamento e teste, que consideram quatro tipos distintos de entidades: PER, ORG, LOC, MISC, os quais são atribuídos respectivamente a nomes de pessoas, organizações, locais e *miscellaneous* (incluindo nomes de obras literárias e de arte)[Ratinov e Roth 2009]. O corpus do CoNLL-2003 possui versão em inglês e alemão [Nadeau e Sekine 2007].

Relacionado ao REN para textos em português O HAREM é um evento de avaliação conjunta para sistemas de reconhecimento de entidades mencionadas em coleções de documentos em português [Nadeau e Sekine 2007]. Dois eventos principais do HAREM são o primeiro HAREM, ocorrido em 2004, e o segundo HAREM, ocorrido em 2008. Ambos possuem coleções de textos distintos com as entidades marcadas, além de coleções douradas, entidades reconhecidas e classificadas por pessoas especializa-

das. As entidades consideradas segundo HAREM são: Abstração; Acontecimento; Coisa; Local; Obra; Organização; Pessoa; Tempo; Valor e Outro.

2.3.2 Reconhecimento de Entidades Nomeadas baseado em Aprendizagem de Máquina

A identificação e a classificação das entidades é realizada de forma automática por sistemas de Reconhecimento de Entidades Nomeadas baseados em aprendizagem de máquina. Esse tipo de abordagem utiliza duas fases: treinamento e teste. O treinamento consiste em dispor de textos, com exemplos de ocorrências das entidades nomeadas já rotuladas. Enquanto a fase de teste, disponibiliza textos a serem rotulados, possibilitando a obtenção dos resultados da eficácia do sistema.

A aprendizagem de máquina é uma abordagem, na qual a classificação da entidade é realizada mediante uma técnica de aprendizagem, podendo ser definida, de acordo com a forma de treinamento, como: supervisionada, semi-supervisionada ou não supervisionada [Nadeau e Sekine 2007].

A aprendizagem de máquina supervisionada faz uso de um corpus anotado (entidades já identificadas e classificadas) constituído de diversos textos. O corpus é utilizado para treinar e testar classificadores de REN, com base nos exemplos de ocorrências das entidades nomeadas. Exemplos desses classificadores incluem a utilização do *Hidden Markov Model* (HMM) [Eddy 1996], *Maximum Entropy Markov Model* (MEMM) [McCallum et al. 1999], *Conditional Random Field* (CRF) [Lafferty et al. 2001] e *Support Vector Machine* (SVM) [Hearst et al. 1998].

Já a aprendizagem de máquina semi-supervisionada e não supervisionada são aplicadas em situações, nas quais o corpus não possui quantidade significativa de entidades anotadas [Speck e Ngomo 2014]. Em tais casos, são adotadas técnicas que identificam e classificam as entidades nomeadas, com base em características do contexto em que são empregadas no texto.

As seções seguintes abordam modelos presentes em diversos trabalhos, que utilizam a aprendizagem de máquina para reconhecer entidades nomeadas, descrevendo suas principais características. Incluindo o HMM e MEMM, que calculam a probabilidade de um termo x pertencer a um rótulo y. Dessa forma ambos os modelos adotam que dada uma sequência de observações $X = \{x\}_{t=1}^{T}$ presumindo-se que existe uma sequência de estados subjacentes $Y = \{y\}_{t=1}^{T}$. Seja S um conjunto finito de estados, e O um conjunto finito de possíveis observações, ou seja, $x_t \in O$ e $y_t \in S$ para todo t.

No processo de Reconhecimento de Entidades Nomeadas, cada x_t observado é a identificação da palavra na posição t, e cada estado y_t é a classe do termo observado em questão, que pode ser Pessoa, Local, Organização, entre outras. Cada identificador x_t para um termo é constituído das *features* que apresentam diversas

características do termo. O cálculo da probabilidade busca relacionar a ocorrência desses identificadores com os rótulos.

Hidden Markov Model

O Modelo Oculto de Markov (HMM − *Hidden Markov Model*) [Eddy 1996] utiliza probabilidade para modelar dados sequenciais. Seu uso na área de Processamento de Linguagem Natural já é difundido, sendo usado para tarefas como *part-of-speech* (POS) *tagging*, segmentação de texto e REN.

Este modelo considera dois pressupostos independentes. Primeiro assume que cada estado depende apenas do seu estado anterior, ou seja, cada estado y_t é independente de todos os seus demais antecessores $y_1, y_2, ..., y_{t-2}$, dado o estado antecessor y_{t-1}. O segundo pressuposto é que a variável de observação x_t depende apenas do estado atual y_t. A partir dessas informações pode-se especificar um HMM usando três distribuições probabilísticas: primeiro, a distribuição de $p(y_1)$ sobre os estados iniciais; segundo, a distribuição de transição $p(y_t|y_{t-1})$; e terceiro, a distribuição de observação $p(x_t|y_t)$. Que é a probabilidade conjunta de uma sequência de estado y e uma sequência de observação x, como mostrado na Equação 2.1.

$$p(y, x) = p(y)p(x|y) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t) \tag{2.1}$$

O HMM é um modelo generativo, que descreve como um rótulo y_t pode gerar probabilisticamente um identificador x_t (composto por suas características). O que pode ser observado na Equação 2.1 pelo cálculo da probabilidade $p(y, x)$.

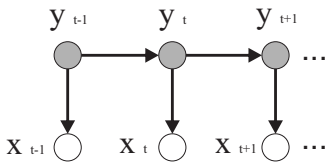

Figura 2.1: Representação gráfica do HMM.

A Figura 2.1 apresenta de forma gráfica como o HMM organiza a sequência de dados do conjunto de observação e o conjunto de estados. No HMM cada observação x_t depende apenas do estado y_t, ao passo que este estado y_t depende do estado y_{t-1}.

Maximum Entropy Markov Model

O Modelo de Máxima Entropia de Markov (MEMM − *Maximum Entropy Markov Model*) [McCallum et al. 1999] é uma estrutura probabilística, na qual as probabilidades de transição são dadas por uma regressão. Relacionado ao REN, fornece a probabilidade de uma palavra pertencer a uma classe.

$$p(y_1, ..., y_t|x_1, ..., x_t) = \prod_{t=1}^{T} p(y_t|y_{t-1}, x_t) \tag{2.2}$$

O MEMM calcula a probabilidade $p(y|x)$ que, por sua vez, maximiza a probabilidade condicional $p(y_1, ..., y_t|x_1, ..., x_t)$. Neste modelo esta probabilidade condicional é fatorada levando em consideração as probabilidades de transições de Markov, onde esta probabilidade de transição, para um rótulo específico, depende apenas da observação nesta posição e no rótulo da posição anterior, como mostrado na Equação 2.2 [McCallum et al. 1999].

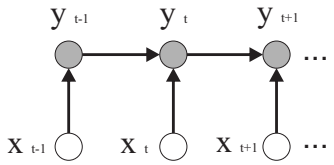

Figura 2.2: Representação gráfica do MEMM.

O modelo gráfico referente ao MEMM é demonstrado na Figura 2.2, na qual cada x_t é relacionado ao conjunto de observação e cada y_t ao conjunto de estados. O modelo MEMM difere do HMM por apresentar a função de probabilidade $p(y_t|, x_t)$, ou seja, cada observação x_t também pode depender do estado anterior y_{t-1}.

Condicional Random Field

Condicional Random Field (CRF) [Lafferty et al. 2001] são modelos de grafos não direcionados. Um caso especial que corresponde a autômatos finitos probabilísticos, condicionalmente treinados. O CRF pode facilmente incorporar um grande número arbitrário de características não independentes [Saha e Ekbal 2013].

Segundo Pereira et al. (2001), considerando X uma variável aleatória sobre sequência de dados a serem rotulados, e Y uma variável aleatória sobre correspondentes sequências rotuladas. Todos os $y_i \in Y$ variam sobre um alfabeto finito de rótulos γ. As variáveis X e Y são distribuídas de forma conjunta, porém num quadro discriminativo, gerando um modelo condicional $p(Y|X)$ de observações e sequências de rótulos.

Com base nessas informações pode-se definir $G = (V, E)$ um grafo tal que $Y = (Y_v)_{v \in V}$, assim Y é indexado pelos vértices de G. Então (X, Y) é considerado CRF no caso, onde condicionado em X, as variáveis aleatórias Yv obedecem a propriedade de Markov com respeito ao grafo: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, onde $w \sim v$ significa que w (que representa um vértice diferente de v) e v são vizinhos em G.

Semi-CRF

Para utilizar o modelo CRF é necessário fazer uso de um grande corpus anotado, no qual os termos já estão rotulados, para a realização do treinamento e aprendizagem, caracterizada como supervisionada. Entretanto, a rotulação destes termos é muito custosa, principalmente em casos em que são feitas manualmente por pessoas. Por esta razão técnicas que obtém boa precisão utilizando apenas pequenas quantidades de dados rotulados, são boas alternativas [Sutton 2012].

O semi-CRF [Cohen e Sarawagi 2004] é uma estratégia para tais casos. Este modelo considera mais que os estados y_{i-1}, o rótulo atual y_i, e o termo de entrada x, ou seja ao avaliar as características de um termo de entrada x, pode avaliar, por exemplo, uma sentença inteira.

Para o uso do semi-CRF é denotada uma segmentação $s = (s_1, ..., s_p)$ de x. Dado o segmento $s_j = (t_j, u_j, y_j)$, no qual t_j representa a posição de início, u_j a posição de fim e $y_j \in Y$ é o rótulo do segmento [Sarawagi e Cohen 2004], indica-se que é marcado um y_j para todos os x_i's entre $i = t_j$ e $i = u_j$. Assim uma avaliação da função característica no j-ésimo segmento é uma função de $(y_j, y_{j-1}, x, t_j, u_j)$, ou seja a atribuição de um rótulo é aplicada para um conjunto de termos que é determinado por meio das definições acima, por exemplo, rotulando não apenas cada termo de um texto, mas cada sentença.

Este modelo pode ser direcionado a casos em que se deseja atribuir simultaneamente o mesmo rótulo para uma quantidade contínua de termos, em que o modelo de treinamento precisa decidir o tamanho deste conjunto de termos [Sutton et al. 2004]. O rótulo de um segmento de termos pode depender apenas do segmento de características e do rótulo do segmento anterior.

Support Vector Machine

Support Vector Machine (SVM) [Hearst et al. 1998] é uma abordagem usada em diversos campos, analisando dados e reconhecendo padrões. O REN é beneficiado pelo uso do SVM para categorizar textos, e tem alcançado alta precisão, apesar de utilizar um grande número de termos tomados como *features* [Saha e Ekbal 2013].

Suponha a existência de um conjunto de dados de treinamento para um problema de classificação: $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ $x_i \in R^D, y \in \{+1, -1\}$. Cada x_i é o vetor de características da i-ésima amostra dos dados de treinamento, e y_i é o rótulo ao qual x_i pertence [Wu et al. 2008].

Na sua forma básica um SVM é projetado para aprender um hiperplano linear, separando as classes, usando conjuntos de exemplos positivos e negativos com uma distância mínima entre os dois conjuntos [Jonnalagadda et al. 2013], termo conhecido como *maximal margin*.

Dos métodos, que compõem classificadores baseados em aprendizagem de máquina, apresentados nesta seção, o estudo comparativo é realizado entre os classificadores CRF, MEMM e HMM. Além disso, o treinamento destes classificadores é supervisionado, uma vez que conta com corpora já anotados para as entidades nomeadas.

2.3.3 Reconhecimento de Entidades Nomeadas baseado em Regras

O Reconhecimento de Entidades Nomeadas baseado em regras analisa diversas características dos termos e da forma como são organizados no texto [Chiticariu et al. 2010]. A análise dessas características possibilita estabelecer as regras que permitem identificar os termos com potencial de serem entidades nomeadas e classificá-los. Técnicas de Reconhecimento de Entidades Nomeadas baseadas em regras costumam utilizar dicionários e gazetteers [Cohen e Sarawagi 2004], contendo uma vasta quantidades de nomes de entidades de um determinado domínio, além de heurísticas, algoritmos e condições.

Os termos "dicionários" e *"gazetteers"'* são frequentemente usados como listas de termos que possibilitam classificar uma entidade nomeada [Nadeau e Sekine 2007]. A inclusão destas listas é uma forma de expressar a relação "é um" (a exemplo, Paris é uma cidade). Ou seja, pela identificação de um termo (Paris) presente em uma lista de cidades, sua probabilidade de ser uma cidade, num dado texto, é alta.

O uso de *gazetteers* pode estar direcionado a listas contendo nomes de cidades e estados de um país, nomes de bairros e ruas, sendo utilizados para classificar os termos que pertencerem a categoria Local. Há também o uso desse tipo de lista de nomes para reconhecer pessoas, onde são coletados diversos nomes, principalmente entre os mais comuns, auxiliando a classificação dos termos de rótulo Pessoa.

Diversas características dos termos são consideradas para extrair informações que possibilitem classificar as entidades nomeadas [Cowie e Lehnert 1996]. A classe gramatical que o termo pertence exemplifica uma das informações que o termo possui (POS *tagging*), a identificação dessa etiqueta atribuída ao termo também é utilizada para identificar termos como entidade nomeada.

A estrutura interna dos termos também é considerada para auxiliar o desenvolvimento das regras de REN [Chiticariu et al. 2010]. Por exemplo, nomes de organizações, na sua forma completa, costumam apresentar designadores da corporação, como Corp., Inc., Ltda. e S.A.. De modo que essa característica ajuda a estabelecer a regra que os termos com esses designadores são classificados como entidades nomeadas da categoria organização. Distintas regras podem ser determinadas para reconhecer as entidades nomeadas, dependendo da percepção de quem desenvolve essa abordagem e do tipo de dado que está sendo trabalhado.

Neste trabalho não é utilizada a abordagem baseada em regras. O estudo comparativo desta dissertação destina-se a classificadores baseados em aprendizagem

de máquina, utilizando distintos corpora para treinar e testar os classificadores em cenários diferenciados.

2.3.4 Reconhecimento de Entidades Nomeadas baseado em abordagens Híbridas

Sistemas de Reconhecimento de Entidades Nomeadas baseados em abordagens híbridas [Srihari et al. 2000] combinam técnicas de aprendizagem de máquina e baseadas em regras. O objetivo desses sistemas é melhorar os resultados, buscando aproveitar-se das características e vantagens das duas abordagens.

Um exemplo de técnica para REN híbrida é vista no trabalho de Rocktäschel et al. (2012) que apresenta um sistema para reconhecer entidades nomeadas que utiliza *Condicional Random Field* combinado com o uso de um dicionário. O CRF é treinado para reconhecer e classificar as entidades e o dicionário melhora os resultados auxiliando na classificação dos termos que pertencem a sua categoria.

Um sistema de REN que use um modelo de aprendizagem de máquina juntamente com *gazetteer* que contenha, por exemplo, nomes de cidades e estados, além de um *gazetteer* com uma lista de nomes comuns de pessoas tem mais chances de obter resultados superiores por utilizar mais de um modelo para reconhecer e classificar as entidades.

Por conta desse potencial existem diversos trabalhos que aplicam abordagens híbridas. Um exemplo dessa aplicação é vista no trabalho desenvolvido por Liu e Ekbal (2013), que fazem uso de *gazetteers*, CRF e agrupamentos para o REN. Outra abordagem híbrida é usada no artigo de Thachenko e Simmanovsky (2012), que combinam o classificador CRF junto a *gazetteers*. Tais exemplos apresentam resultados superiores, quando comparado com abordagens que utilizam apenas uma técnica.

Abordagens híbridas não são utilizadas para comparar os classificadores. O foco é na comparação dos classificadores baseados em aprendizagem de máquina supervisionada. O objetivo desta definição é contribuir com a comparação destes classificadores, quando submetidos a textos em português. A utilização de uma abordagem híbrida pode ser utilizada em um trabalho futuro.

2.3.5 Resumo Comparativo das principais técnicas

Levando em consideração a existência de sistemas baseados em regras, aprendizagem de máquina e híbridos, a técnica adotada para identificar e classificar as entidades tende a variar. Cada técnica é definida para melhor aplicar o REN no domínio utilizado. Esta seção demonstra distintas aplicações de REN, compostas por técnicas variadas.

Determinar quais rótulos de entidades são consideradas é uma definição importante para a realização do REN. Os trabalhos nessa área de pesquisa não somente envolvem a identificação e classificação de entidades do tipo Pessoa, Organização e Local, mas também para extrair entidades com rótulos de Genes, abreviações e fórmulas moleculares, entre outros, como no trabalho de Zhang et al. (2013).

A Tabela 2.1 apresenta artigos publicados de REN, contendo informações da abordagem usada, quais ferramentas de REN e/ou de Processamento de Linguagem Natural foram utilizadas, quais categorias de entidades são consideradas, o(s) corpus usado para auxiliar na identificação e classificação das entidades, e as medidas que cada sistema usa para obter os resultados.

Alguns dos trabalhos apresentados na Tabela 2.1 fazem uso de mais de um classificador para identificar e classificar as entidades nomeadas. Parte dos trabalhos publicados utilizam uma combinação que varia entre modelos de aprendizagem de máquina (necessitando de um conjunto de dados já rotulados para treinamento), *gazetteers* ou dicionários, e regras no formato de heurísticas para auxiliar o REN.

A diversidade de domínio de estudo nos trabalhos que fazem uso do REN, está relacionada com a variedade de suas aplicações. Os trabalhos presentes na Tabela 2.1 adotam rótulos de entidades nomeadas relevantes para cada aplicação, como genes e proteínas, além de rótulos mais genéricos como: Pessoa, Local e Organização.

O corpus adotado para cada trabalho está relacionado ao conjunto de rótulos de entidades nomeadas selecionados. Os corpora CoNLL-03, SCAI, GENIA exemplificam bases textuais que possuem ocorrências de termos, cujos rótulos são relevantes para domínios distintos.

Os trabalhos mostrados na Tabela 2.1, revelam ainda o uso das medidas: Precisão, Cobertura e F1. A Seção 4.3.1 descreve com mais detalhes essas medidas, que servem para verificar a eficácia de cada sistema de REN, além de auxiliar definições relacionadas a fase de desenvolvimento dos sistemas, para que alcancem resultados desejáveis, como por exemplo, os dados de entrada para os classificadores.

A tarefa de REN é importante para anotar dados textuais, identificando e classificando termos críticos, e auxiliando a EI. Dadas as diferentes abordagens para realizar o REN, é visto um diversificado número de técnicas que fazem uso de distintos classificadores, dicionários e *gazetteers*.

Entre os classificadores baseados em aprendizagem de máquina, é observado que boa parte dos trabalhos utilizam o CRF. Este classificador é utilizado para compor a comparação, principalmente com os corpora em português, dado que a maioria dos trabalhos de comparação que incluem os classificadores CRF, MEMM e HMM são voltados unicamente para o idioma inglês.

Tabela 2.1: Distintas técnicas aplicadas em trabalhos sobre REN.

Artigos	**Técni. de REN**	**Entidades**	**Corpus**	**Medidas**
[Fersini et al. 2014]	CRF, Semi-CRF	Pessoa, Local, Organização, Miscellaneous; Anúncios para apartamentos e Paper entities	CoNLL-03, US50, Cora, Advertisements	Precisão, Cobertura, F1
[Liu e Zhou 2013]	CRF, Gazetteers, Clustering	Pessoa, Local, Organização, Produto, Outro	Anotado usando a API do Twitter	Precisão, Cobertura, F1
[Saha e Ekbal 2013]	Naive Bayes, DT, MBL, HMM, MEMM, CRF, SVM	Pessoa, Local, Organização, Miscellaneous	Anotado usando Bengali news, NERSSEAL	Precisão, Cobertura, F1
[Ling e Weld 2012]	CRF	Pessoa, Organização, Local, Produto, Artigo, Evento, Prédio	Textos da Wikipedia	Precisão, Cobertura, F1
[Tkachenko e Simanovsky 2012]	CRF, Gazetteers	Pessoa, Local, Organização, Miscellaneous; Proteína, DNA, RNA, linha e tipo de célula	CoNLL-03, OntoNotes v4, NLPBA 2004	Precisão, Cobertura, F1
[Li et al. 2012]	Algoritmo dinâmico, Segment ranking	Pessoa, Local, Organização, Miscellaneous	SIN, SGE	Precisão, Cobertura, F1
[Rocktäschel et al. 2012]	CRF, Dicionário	Nomes triviais, Drogas, Abreviações, Fórmulas Moleculares, IUPAC Entidades	SCAI	Precisão, Cobertura, F1
[Atkinson e Bull 2012]	SVM, HMM	Gene/proteína	GENIA, BioCreative	Precisão, Cobertura, F1
[Liu et al. 2011]	KNN, semi-CRF, Gazetteers	Pessoa, Local, Organização, Produto	Anotado usando Twigg SDK	Precisão, Cobertura, F1

Capítulo 3

Trabalhos Relacionados

A tarefa de REN é utilizada em diversos trabalhos. Os rótulos de entidades que, comumente, se resumem em Pessoa, Local e Organização, podem mudar de acordo com a área de pesquisa ou o domínio de interesse. Trabalhos que analisam ferramentas e modelos de REN também são realizados, objetivando sua avaliação. Os trabalhos relacionados apresentados na Seção 3.1 avaliam distintos sistemas de NER, dentre os quais estão inclusos os que utilizam aprendizagem de máquina. Enquanto a Seção 3.2 demonstra trabalhos que envolvem a avaliação das *features* para o REN, também incluindo sistemas baseados em aprendizagem de máquina.

3.1 Comparações baseadas em ferramentas de Reconhecimento de Entidades Nomeadas

O trabalho de Mansouri [Mansouri et al. 2008] faz uma avaliação de modelos de Reconhecimento de Entidades Nomeadas e compara os resultados de sistemas baseados em regras, aprendizagem de máquina e híbridos. Para essa avaliação, os autores utilizam as definições e o corpus da conferencia MUC. A fim de obter informações das características de cada abordagem. Além disso, apresentam um método de REN baseado em aprendizagem de máquina chamado de Fuzzy Support Vector Machine (FSVM).

Para comparar as diferentes abordagens de REN são usados os sistemas baseados em regras: IsoQuest Inc [Krupka e Hausman 1998], NYU System [Grishman 1995] e U. of Manitoba [Lin 1995]; os sistemas baseados em aprendizagem de máquina: Nymble [Bikel et al. 1997], MENE [Borthwick et al. 1998], IdentiFinder[1], Support Vector Machine, Association Rule Mining, Maximum Entropy; e os sistemas híbridos: LTG [Grover et al. 2010], NYU Hybrid [Borthwick et al. 1998]. A avaliação

[1]http://www.bbn.com/technology/identifinder

de cada sistema demonstra que as abordagens baseadas em regras obtêm bons resultados em seus domínios específicos, ou seja, nos casos em que a criação de regras beneficiam ocorrências específicas de entidades nomeadas.

Quando aplicada em textos genéricos, abordagens de aprendizagem de máquina são as melhores soluções por serem independentes de limitações das regras criadas. Porém a alta performance dessa abordagem depende dos dados de treino. O modelos híbridos apresentam bons valores de precisão e cobertura. Entretanto a probabilidade de acerto deste tipo de abordagem pode ser reduzida na tentativa de aumentar a precisão de reconhecimento de entidades nomeadas, por meio do uso de quantidades excessivas de regras reparadoras, uma vez que essas regras são definidas para casos muito específicos, e não genéricos, podendo afetar a rotulação das entidades, realizada na aprendizagem de máquina.

Mansouri faz uso de somente uma base textual (MUC) para comparar os sistemas de REN. O trabalho desenvolvido nesta dissertação diferencia-se por apresentar quatro corpora distintos, em dois idiomas (inglês e português), a fim de contribuir com mais domínios de interesse.

No trabalho de Marrero [Marrero et al. 2009], um conjunto de ferramentas de REN são analisadas. A avaliação dessas ferramentas constitui-se de uma especificação resumida de suas técnicas, a análise da qualidade das ferramentas, de acordo com seus resultados obtidos por meio das medidas de precisão, *recall* (cobertura) e F1, de forma geral e por tipo de entidade. As ferramentas avaliadas neste trabalho são: Supersense-CONLL, Supersense-WNSS e Supersense-WSJ [Ciaramita e Altun 2006], Afner [Van Zaanen et al. 2007], Annie [Cunningham et al. 2002], Freeling [Atserias et al. 2006], TextPro [Pianta et al. 2008], YooName [Nadeau et al. 2006], ClearForest [Iskold 2006], Lingpipe [Carpenter 2007]. Supersense-WNSS e Clearforest obtêm melhores resultados nos testes realizados, enquanto a ferramenta Afner obteve os menores resultados. Uma das conclusões que os autores demonstram é, que não apenas a quantidade de categorias de entidades que os sistemas reconhecem é importante, mas também a acurácia dos resultados.

Rizzo e Troncy [Rizzo e Troncy 2012] desenvolveram um trabalho que analisa e compara dez populares ferramentas de REN disponíveis na Web, além de disponibilizar um framework que permite reunir essas ferramentas e avaliar os seus resultados. As ferramentas disponíveis nesse framework, denominado NERD (*Named Entity Recognition and Disambiguation*), são: AlchemyAPI[2], DBpedia Spotlight[3], Evri[4], Extrac-

[2]http://www.alchemyapi.com
[3]http://dbpedia.org/spotlight
[4]http://www.evri.com/developer

tiv[5], Lupedia[6], Wikimeta[7], OpenCalais[8], Yahoo! Content Extraction[9], Zemanta[10] e Saplo[11]. O framework desenvolvido nesse trabalho permite unificar as saídas de cada uma dessas ferramentas já existentes nas seguintes categorias de entidades nomeadas: Thing, Amount, Animal, Event, Function, Location, Organization, Person, Product e Time, disponibilizando-as no mesmo ambiente. Um dos testes realizados nesse framework demonstra que a ferramenta DBpedia Spotlight reconhece poucas categorias de entidades, Zemanta possui performance de classificação superior a DBpedia e AlchemyAPI demonstra forte habilidade para reconhecer entidades do tipo Person, e considerável resultado para Organization. Como parte dos trabalhos futuros os autores destacaram investigar como a combinação das ferramentas afeta os resultados.

De forma semelhante a Marrero et al. (2009), Rizzo e Troncy reúnem ferramentas de REN já definidas, indicando que os critérios para a identificação e classificação as entidades nomeadas (como conjunto de *features* usadas) não é abordado. A pesquisa desta dissertação reúne comparações de diferentes classificadores de REN, além de apresentar distintos conjuntos de *features* e suas contribuições para o REN.

O trabalho de Abacha e Zweigenbaum [Abacha e Zweigenbaum 2011] realiza um estudo comparativo de classificadores de REN médicas. Os autores apresentam e comparam três classificadores baseados em regras e aprendizagem de máquina, assumindo duas abordagens: na primeira é usado um identificador (*chunker*) para a classificação das entidades médicas; na segunda abordagem são usados os classificadores de aprendizagem de máquina para, simultaneamente, identificar e classificar as entidades. Os resultados obtidos demonstram que as abordagens híbridas obtêm melhores performances. Para realização do estudo comparativo os autores utilizam as ferramentas: Treetagger-chunker, OpenNLP [Atserias et al. 2006], MetaMap [Aronson 2001], libSVM, e CRF.

O trabalho de Kang [Kang et al. 2012] utiliza uma abordagem que faz uso de um conjunto de técnicas de REN para extração de informação em registros clínicos. O objetivo destacado pelos autores é melhorar o desempenho de uma variedade de sistemas de REN, combinando os seus resultados individuais. São utilizados dois sistemas baseados em dicionários: MateMap [Aronson 2001] e Peregrine [Schuemie et al. 2007]; e cinco sistemas baseados em modelos estatísticos: ABNER [Settles 2005] − sistema para REN em textos de biologia molecular, baseado em CRF; Lingpipe [Carpenter 2007] − suíte de bibliotecas em Java para processamento de linguagem natural, baseado em HMM; OpenNLP Chunker[12] − quite de ferra-

[5]http://extractiv.com
[6]http://lupedia.ontotext.com/
[7]http://www.wikimeta.com
[8]http://www.opencalais.com
[9]http://developer.yahoo.com/search/ content/V2/contentAnalysis.html
[10]http://www.zemanta.com
[11]http://www.saplo.com/
[12]https://opennlp.apache.org/

mentas para PLN, baseado em MEMM; JNET[13] – sistema genérico de REN, utiliza o modelo CRF e StanfordNer[14] – sistema de REN desenvolvido pelo Grupo de Processamento de Linguagem Natural de Stanford, baseado em linear chain CRF. Os autores concluem que o uso combinado (utilização de mais de um classificador em conjunto) das ferramentas de REN demonstra, substancialmente, melhores resultados que qualquer uma ferramenta individualmente.

A pesquisa desta dissertação limita-se a utilizar classificadores baseados em aprendizagem de máquina, e comparar seus resultados individuais, dado que são usados corpora em português, e que a grande maioria dos trabalhos que comparam sistemas de REN são voltados para o idioma inglês. Além disso, o uso de sistemas já desenvolvidos não permite uma exploração mais detalhada do processo de REN, uma vez que diversos fatores, como a escolha das *features*, modelo de aprendizagem, tipos de dados e até mesmo as categorias de entidades nomeadas, podem influenciar os resultados de cada sistema.

3.2 Comparações baseadas no uso de Features

O trabalho desenvolvido por Tkachenko e Simanovsky [Tkachenko e Simanovsky 2012] explora o uso de *features* para avaliar como contribuem para o REN. Os autores estudam um conjunto de *features* usadas no REN supervisionado, explorando variações de suas combinações e comparando seu impacto na performance do reconhecimento das entidades. Neste trabalho, são estudadas *features* descritas como de conhecimento local, que podem ser extraídas da estrutura interna do termo e do seu contexto, e *features* de conhecimento externo, como *POS tags*, *gazetteers*, entre outros. A fim de melhorar o REN também são adotadas *features* para ocorrências específicas de tipos de termos, como termos compostos separados por hífens e termos numéricos. O estudo do uso dessas *features* é feito com o desenvolvimento de um sistema baseado em CRF. São utilizadas as bases do CoNLL 2003 e a Versão 4 do *OntoNotes* CNN. Neste trabalho é observado que o uso dos termos de contexto, como *features*, atingem bons valores na identificação e classificação das entidades num intervalo de três termos (o termo atual, o antecessor e o sucessor). Os autores observam ainda, que o comportamento das *features* baseadas nas características internas dos termos dependem da especificidade das entidades. Além de chegarem a conclusão de que o uso de *gazetteers* ainda é útil para o REN.

Com base nessas informações do trabalho de Tkachenko e Simanovsky é possível concluir que o conjunto de *features* usadas numa abordagem de REN supervisionada pode afetar o seu desempenho. Esse é um fator de estudo também considerado na pesquisa dessa dissertação. Porém, além de explorar os impactos das *features* no

[13]http://www.julielab.de/
[14]http://nlp.stanford.edu/software/CRF-NER.shtml

REN, avaliar suas relações com distintas bases de dados é outro fator que é considerado, uma vez que as diferenças entre as características das bases pode adicionar uma informação sobre como cada conjunto de *features* influencia o REN.

Outro trabalho que explora o uso de *features* é o desenvolvido por Settles [Settles 2004]. O qual apresenta um *framework* para reconhecer simultaneamente as ocorrências de entidades das categorias Proteína, DNA, RNA, Linha e Tipo de Célula com CRF. Os conjuntos das *features* utilizados são divididos entre ortográficas e semânticas. As *features* ortográficas são descritas como as que representam características que generalizam as ocorrências dos termos, essas features são retiradas da estrutura interna dos termos, como identificar palavras com letra maiúscula, termos numéricos, números romanos, afixos (prefixos e sufixos), entre outros. Já as *features* semânticas são usadas para associar alguns termos com rótulos de entidades nomeadas, levando em consideração características em comum que possuem. É testado o REN com apenas as *features* ortográficas, obtendo resultado da medida F_1 de 69.8. O segundo teste incluindo as *feautres* semânticas atinge a medida F_1 de 69.5. Settles afirma que o uso de *features* semânticas, mesmo com a medida F_1 menor, traz um aumento nos valores para as medidas precisão e cobertura, para os rótulos de entidades RNA e Linhas de Célula.

O trabalho de Settles observa como dois conjuntos distintos de *features* afeta o REN, com o uso do CRF. O trabalho de Settles é voltado à área de Biomedicina, incluindo *features* que contribuem para identificar e classificar termos deste domínio. O trabalho desenvolvido nesta dissertação apresenta um conjunto de *features* que atendam a mais de um domínio de assunto ou idioma, por serem extraídas das suas características internas (morfológicas).

Técnicas para seleção de *features* são estudadas no trabalho de Saha [Saha et al. 2009]. O objetivo demonstrado por Saha é selecionar o melhor conjunto de *features* que melhore o REN para Biomedicina, incluindo entidades nomeadas de categorias como proteínas, genes, DNA, e RNA. O experimento desse trabalho envolveu o uso do MEMM como classificador, sendo treinado com a base de dados JNLPBA 2004[15].

O sistema no estudo dos autores, é inicialmente desenvolvido com um conjunto de *features* que passa por técnicas de redução, a fim de diminuir a quantidade das *features* utilizadas. O conjunto inicial de *features* definido para o classificador de Máxima Entropia possui: o termo − incluindo o atual, dois antecessores e dois sucessores; etiquetas das entidades nomeadas antecessoras, capitalização e informação de dígitos; caracteres especiais − ocorrência de pontos, hífens, entre outros sinais; normalização do termo − cada termo no formato lematizado; prefixos e sufixos de cada termo; informação do POS *tag*; e *trigger words* (substantivos e verbos que ocorrem frequentemente ates de uma entidade nomeada).

As seleções das *features*, usadas por Saha são: seleção informativa dos termos, a

[15]http://www.bbn.com/technology/identifinder

qual considera os termos da adjacência; termos internos de uma entidade nomeada; termos externos às entidades nomeadas, representando os termos que com frequência ocorrem logo após uma entidade nomeada; por fim a redução dos valores de algumas *features* (caso o termo seja considerado como não relevante para auxiliar o REN, de acordo com as seleções listadas acima, a *feature* para esse termo possuirá o valor *null*). O classificador deste trabalho é testado com variações das *features*, objetivando identificar suas contribuições. Fica concluído por esse estudo, que o conjunto total das *features* permite o melhor reconhecimento das entidades nomeadas.

O trabalho de Saha [Saha et al. 2009] apresenta o estudo do uso das *features*, sendo avaliada suas contribuições para o REN. É demonstrado que variações no conjunto das *features* influenciam o reconhecimento das entidades nomeadas. Este trabalho, entretanto, não avalia o impacto da variação do uso das *features* com outros classificadores, como CRF por exemplo, o qual é utilizado em trabalhos de REN, por atingir valores de maior eficácia na identificação e classificação de entidades nomeadas.

A pesquisa proposta por essa dissertação inclui avaliar diferentes conjuntos de *features*, utilizando distintas bases de dados. A fim de obter informações relacionadas ao REN em distintas condições, não se limitando a domínios mais específicos. Por conta disso, não são inclusas *features* para reconhecer rótulos como genes e proteínas, de modo que as *features* selecionadas auxiliem mais domínios de assunto e idioma, auxiliando, assim, futuros trabalhos de forma mais simples e generalizada.

Capítulo 4

Processo de Avaliação

Este capítulo contém uma descrição do processo de avaliação elaborado para realizar o estudo comparativo dos classificadores. Este processo é composto pelas seguintes etapas (Figura 4.1): 1) seleção da base de dados, 2) pré-processamento e 3) seleção dos classificadores. Por fim é produzido um relatório comparando os experimentos.

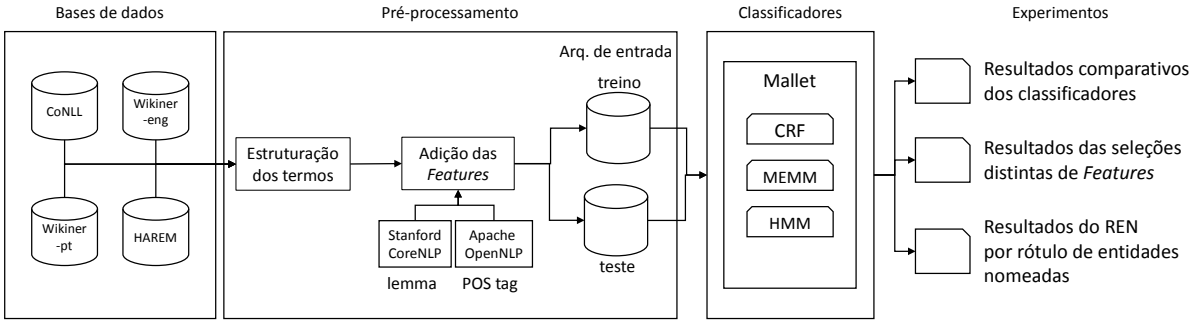

Figura 4.1: Etapas do processo de avaliação de classificadores.

A Seção 4.1 apresenta as bases de dados que podem ser selecionadas, e a definição dos rótulos de entidades nomeadas utilizados, os quais estão condicionados às bases que já são rotuladas. A Seção 4.2 descreve a etapa de pré-processamento, que consiste em preparar os dados para serem utilizados nos experimentos. Já a Seção 4.2.1 destaca a seleção das *features*, adquiridas dos termos, para serem usadas como entrada dos classificadores baseados em aprendizagem de máquina, determinando assim os parâmetros relacionados ao desenvolvimento da pesquisa.

A Seção 4.3 aborda definições para avaliação comparativa dos classificadores. Apresenta as medidas definidas para comparar a eficácia dos classificadores (Seção 4.3.1). Já na Seção 4.3.2, são brevemente abordados métodos utilizados para validação dos classificadores.

Com base na avaliação de trabalhos relacionados a REN (Tabela 2.1), há um uso frequente de técnicas que implementam classificadores baseados em aprendizagem de máquina. Os classificadores HMM, MEMM, e CRF são comparados estudando a contribuição das *features*, de forma individual e em conjunto, observando a eficácia dos classificadores em distintos corpora e na relação de identificação e classificação dos rótulos de entidades nomeadas.

4.1 Bases de dados

Os modelos de aprendizagem de máquina implementados neste estudo comparativo, por serem supervisionados, requerem a utilização de um corpus já rotulado para as entidades nomeadas. Com a disponibilidade desse corpus é possível treinar os modelos de classificação das entidades, para então realizar os testes que possibilitam avaliar os resultados de cada classificador.

O treinamento supervisionado requer grande número de textos do corpus, para que os classificadores sejam capazes de representar o maior número possível de exemplos de ocorrências das entidades em textos [Nadeau e Sekine 2007]. Uma base com menos de 100 textos, por exemplo, não constitui uma quantidade suficiente para treinar, de forma supervisionada, os classificadores. Bases com um grande número de textos facilitam o aprendizado dos classificadores, dispondo de vários exemplos nos quais as entidades são empregadas no texto.

As bases de dados utilizadas neste trabalho foram adquiridas de forma gratuita. Alguns corpora, como MUC-6 e MUC-7 (Seção 2.3.1) requerem um pagamento para aquisição dos textos que compõem suas bases. No exemplo destes corpora, a obtenção dos dados é possível de duas formas: 1) cadastrando-se como membro da LDC (*Linguistic Data Consortium*) [Liberman e Cieri 1998], o que oferece descontos para pagar pelos dados; 2) Obtendo uma licença, como não membro da LDC, tendo que concordar com um acordo de aquisição dos dados, para efetuar o pagamento.

Optou-se ainda, por trabalhar com corpora nos idiomas inglês e português. Os corpora em inglês já são utilizados em diversos trabalhos de REN, como [Fersini et al. 2014] e [Tkachenko e Simanovsky 2012]. No caso das bases em português são escolhidas as mais relatadas em trabalhos de Reconhecimento de Entidades Nomeadas que utilizam o idioma, como [Nothman et al. 2013] e [Santos et al. 2006].

4.1.1 CoNLL-03

Uns dos corpora usados para treinamento e testes é o disponibilizado pela conferência CoNLL-03, que é gerado a partir de um conjunto de artigos de notícias, no idioma inglês, provenientes do *Reuters Corpus*. Os arquivos de dados textuais do corpus

CoNLL-03, possuem quatro colunas separados por um espaço em braço. Cada palavra encontra-se separada por uma linha e há uma linha em branco ao final de cada sentença. O primeiro item de cada linha é a palavra seguida de *part-of-speech* (POS) *tag*, o terceiro item é definido como *syntatic chunk* que adiciona uma informação sintática para a palavra, por último a categoria da entidade: PER, ORG, LOC e MISC, indicando respectivamente Pessoa, Organização, Local e *Miscellaneous*. O rótulo 'O' ocorre caso o termo não seja uma entidade.

A Figura 4.2 apresenta um trecho do arquivo de treinamento, do corpus CoNLL-03, demonstrando seu formato padrão. A primeira linha desta figura é composta por "Fischler NNP I-NP I-PER", onde Fischler é o termo do texto, NNP é o POS *tag* para este termo, I-NP é o *syntatic chunk*, já o I-PER é o rótulo dado ao termo, indicando que é uma entidade nomeada referente a uma pessoa.

Fischler	NNP	I-NP	I-PER
proposed	VBN	I-NP	O
EU	NNP	I-NP	I-MISC
measures	VBZ	I-VP	O
after	IN	I-PP	O
reports	NNS	I-NP	O
from	IN	I-PP	O
Britain	NNP	I-NP	I-LOC
and	CC	I-NP	O
France	NNP	I-NP	I-LOC
.	.	O	O

Figura 4.2: Formato do corpus CoNLL-03.

A base CoNLL-03 é composta por três arquivos dispostos nesse formato. Um arquivo de treino, juntamente com mais dois arquivos de testes ("eng.testa" e "eng.testb"), referenciados nesse trabalho como conll-test1 e conll-test2, respectivamente. Os dados referentes ao número de artigos, sentenças e termos, além do número de termos de cada rótulo de entidade estão apresentados na Tabela 4.1, informando a dimensão de cada arquivo de treino e teste, e a relação entre o quantidade total de termos com as ocorrências de entidades nomeadas (LOC, MISC, ORG e PER).

Tabela 4.1: Dados quantitativos da base CoNLL-03.

Arquivos de Dados	Artigos	Sentenças	Termos	LOC	MISC	ORG	PER
conll-train	946	14987	203621	8286	4556	10001	11128
conll-test1	216	3466	51362	2094	1264	2092	3149
conll-test2	231	3684	46435	1919	909	2491	2773

O corpus conll-test1 é disponibilizado pelo CoNLL-03 para auxiliar o desenvolvimento do sistema de REN, não sendo utilizado para testar sua eficácia. Portanto, os experimentos que envolvem os corpora derivados desta base apresentam valores referentes ao REN com o corpus conll-test2.

4.1.2 WikiNER

As bases provenientes do WikiNER [Nothman et al. 2013] são constituídas por diversos artigos da Wikipedia[1]. No total são disponíveis corpora para o REN em nove idiomas: Inglês, Alemão, Espanhol, Francês, Italiano, Holandês, Polonês, Português e Russo.

As bases de dados são caracterizadas, quanto a aquisição dos artigos que as constitui, como *Popular labelled* corpus e *Random labelled* corpus. Apenas um corpus, com o idioma inglês, é definido como *Popular*, contendo, entre outros, 1000 dos artigos mais acessados na Wikipedia em 2008 [Nothman et al. 2013]. As bases de dados criadas por seleção aleatória de artigos da Wikipedia foram desenvolvidas para os nove idiomas.

Os rótulos de entidade nomeada presentes nestes corpora são PER, ORG, LOC e MISC. A Tabela 4.2 apresenta as estatísticas para todos os corpora disponíveis pelo WikiNER, na qual pode-se avaliar o tamanho, por meio no número de artigos, além de uma distribuição aproximada dos rótulos de entidades, páginas de desambiguação (DAB), e termos que não são entidades nomeadas (NON).

Tabela 4.2: Dados estatísticos das bases de dados disponíveis pelo WikiNER.

Corpus		Num. de artigos	Distribuição dos rótulos (%)					
			LOC	ORG	PER	MISC	NON	DAB
POPULAR	Inglês	2322	28	11	11	16	30	4
RANDOM	Inglês	2531	20	10	26	18	16	10
RANDOM	Alemão	872	19	11	33	13	12	12
RANDOM	Espanhol	203	28	10	19	19	20	4
RANDOM	Francês	210	22	5	25	20	20	8
RANDOM	Italiano	203	30	4	23	19	18	6
RANDOM	Holandês	286	34	9	17	15	17	8
RANDOM	Polonês	210	36	4	30	13	11	6
RANDOM	Português	202	38	6	17	15	19	5
RANDOM	Russo	223	30	8	26	14	13	9

Dados os nove corpora disponíveis, este trabalho comparativo faz uso do corpus constituído de artigos selecionados aleatoriamente no idioma inglês e português. O corpus em inglês com artigos populares não é utilizado para que haja o mesmo critério entre os demais corpora do WikiNER.

As Tabelas 4.3 e 4.4 mostram os dados dos corpora, para treino e teste, em inglês e português respectivamente, derivados da base WikiNER. Os corpora são disponíveis sem a divisão para treino e teste. Esta divisão é realizada levando em consideração a quantidade de termos de cada corpus, evitando cortar sentenças. Com base nisso,

[1]https://www.wikipedia.org/

os corpora de treino possuem cerca de 70% da base original, enquanto os corpora de treino possuem aproximadamente 30%.

Tabela 4.3: Dados quantitativos dos corpora, em inglês, derivados das bases WikiNER.

Corpus	Termos	Sentenças	LOC	MISC	ORG	PER
wikiner-eng-train	2450061	98750	87717	90785	61464	102172
wikiner-eng-test	1049591	42756	35491	39375	27982	42427

Tabela 4.4: Dados quantitativos dos corpora, em português, derivados das bases WikiNER.

Corpus	Termos	Sentenças	LOC	ORG	PER
wikiner-pt-train	2153779	84979	157794	21520	59207
wikiner-pt-test	923197	36432	42036	14543	35218

O corpus em inglês derivado do WikiNER possui um número superior de artigos, em relação ao corpus CoNLL-03, possibilitando estudar o comportamento do REN em corpora de tamanhos distintos. De forma semelhante, o WikiNER em português contém quantidade de dados textuais superiores ao corpus CoNLL-03. A utilização desta base em português é uma contribuição para o estudo do REN neste idioma.

No corpus de treino 'wikiner-ent-train' há mais ocorrência de entidades nomeadas para os rótulos PER, com mais de 102 mil termos, e MISC com mais de 90 mil termos. Enquanto o corpus 'wikiner-pt-train' possui mais 157 mil termos com o rótulo LOC, contra mais de 59 mil termos com rótulo PER. As diferenças entre estas bases permitem observar como a eficácia no reconhecimento de tais rótulos é adquirira pelos classificadores.

Os corpora em inglês, derivados do WikiNER e CoNLL-03 consequentemente, possuem os mesmos rótulos de entidades nomeadas. O mesmo critério é adotado entre os corpora em português utilizados, permitindo não somente comparar a eficácia dos classificadores de modo geral (na tarefa de REN), mas também por rótulo de entidade nomeada.

4.1.3 HAREM

A Avaliação Conjunta para Sistemas de Reconhecimento de Entidades Mencionadas − HAREM, constitui-se como um evento que avalia sistemas de reconhecimento de entidade nomeadas da língua portuguesa [Santos et al. 2006]. O primeiro HAREM (2004) e o segundo HAREM (2008) são dois eventos destacados por disponibilizar

coleções de documentos textuais no formato de bases de dados para serem utilizados pelos sistemas de REN.

A coleção do primeiro HAREM contém textos da Web, jornalísticos, textos provenientes de revistas, literários e políticos, contabilizando na sua totalidade 466355 termos. O português destes textos é proveniente de países como Portugal, Brasil e Angola. Há também uma versão desse corpus denominada de Coleção Dourada, contendo 89241 termos e 3851 entidades.

O segundo HAREM possui um corpus com uma coleção de 1040 documentos, dos quais 129 pertencem a sua Coleção Dourada. Os padrões para este corpus seguem os mesmos critérios do primeiro. As categorias de entidade nomeadas utilizadas em ambas as coleções são Pessoa, Organização, Local, Tempo, Acontecimento, Obra, Abstração, Coisa, Valor e Outro. A Tabela 4.5 apresenta a relação de documentos, e rótulos de entidades nomeadas para cada coleção dourada. O corpus utilizado nos experimentos deste trabalho, é composto por textos das duas coleções do HAREM, buscando adquirir uma base maior. As coleções douradas do HAREM, são anotadas por pessoas que revisaram os textos para atribuir os rótulos de entidades nomeadas, aos seus devidos termos.

Tabela 4.5: Relação dos rótulos de entidades das duas coleções douradas do HAREM.

Rótulos de	Corpus	
Entidades	CD HAREM 1	CD HAREM 2
PESSOA	822	2035
ORGANIZACAO	575	958
LOCAL	844	1239
TEMPO	355	1189
ACONTECIMENTO	54	302
OBRA	189	437
ABSTRACCAO	205	278
COISA	163	304
VALOR	328	352
OUTRO	14	8

O corpus do HAREM utilizado neste estudo comparativo faz uso de, aproximadamente, 70%, em relação ao número de termos da base, para realização do treino e 30% para ser usado no teste dos classificadores de REN. Em relação aos dois corpus citados anteriormente, a base de dados do HAREM é a menor. O tamanho do corpus é também um dos fatores utilizados na comparação dos classificadores.

A Tabela 4.6 apresenta os dados relativos a quantidade de termos, sentenças e ocorrências de rótulos de entidades nomeadas para os corpora de treino e teste, montados a partir das coleções do HAREM. O número de termos é o menor nesta base, assim como o número de ocorrências de entidades nomeadas.

Tabela 4.6: Dados quantitativos dos corpora derivados das coleções do HAREM.

Corpus	Termos	Sentenças	LOC	ORG	PER
harem-train	158989	7353	3419	3506	5262
harem-test	18028	445	549	270	998

Os corpora derivados das coleções do HAREM, como demonstrado na Tabela 4.6, apresentam menor quantidade de termos e, consequentemente, ocorrências de entidades nomeadas. A utilização destes corpora auxilia a comparação dos classificadores de entidades nomeadas no idioma português.

No que diz respeito aos corpora em português, derivados das bases WikiNER e HAREM, a diferença dos números de termos, sentenças e rótulos de entidades nomeadas é expressivo. Essa diferença possibilita comparar o comportamento dos classificadores de REN, quando submetidos a corpora em português com tamanhos e números de ocorrência de entidades nomeadas distintos.

4.2 Pré-Processamento

Como já mencionado na Seção 2.2, a etapa de pré-processamento é crucial para preparar os dados que são usados por cada classificador. O pré-processamento realizado neste estudo comparativo segue esse propósito e é aplicado nos corpora selecionados para a realização dos experimentos.

O pré-processamento é dividido em cinco principais etapas: 1) separação dos termos, 2) separação das sentenças, 3) aquisição do POS *tagging*, 4) aquisição do lema dos termos e 5) extração de características internas dos termos. Estas etapas preparam os arquivos de treino e teste para o formato de entrada dos classificadores.

A primeira etapa de pré-processamento aplicada é a separação de cada termo (*tokenization*). Os classificadores reconhecem o primeiro termo de cada linha, do arquivo de entrada, como a palavra que precisa ser classificada. O próprio termo também é utilizado como uma *feature*, auxiliando as etapas de treinamento e teste.

A segunda etapa consiste na separação das sentenças. Este processo auxilia os classificadores identificarem cada trecho de texto. A separação das sentenças também auxilia determinar os limites de início e fim para algumas *features* de contexto (Tabela 4.7), que usam características de termos antecessores e sucessores e que não podem ultrapassar o tamanho de uma sentença. Além de permitir detectar o primeiro termo de cada sentença, o qual recebe uma *feature* específica (Seção 4.2.1).

A terceira etapa é a aquisição do POS *tagging*. Esta etapa só é necessária para os corpora em português, visto que os corpora em inglês utilizados já possuem essa característica para cada termo. A adição do POS *tagging* no idioma português é

realizado com o uso do quite de ferramentas de aprendizagem de máquina para processamento de linguagem natural, *Apache OpenNLP*[2], que recebe cada termo do corpus e informa seu POS *tagging*.

Na quarta etapa, é realizado um procedimento para adquirir o lema dos termos. A lematização de um termo o coloca em uma forma simplificada, ignorando tempos verbais, em casos de verbos, especificações de gênero, número e grau. Esta etapa de pré-processamento é realizada com o auxílio do quite de ferramentas em Java *Stanford CoreNLP* [Manning et al. 2014], que recebe o termo e informa sua forma lematizada.

A última etapa consiste em extrair as características de cada termo, explorando sua estrutura interna. São identificados termos com ocorrências de números, hifens e letras maiúsculas, além do comprimento de cada termo (que informa o número de letras que esse termo possui) e da verificação da posição do termo no texto, para indicar se este termo é o primeiro de uma sentença. Esta etapa consiste em extrair *features* para os termos, como melhor detalhado na Seção 4.2.1. O pré-processamento permite que os dados fiquem prontos para serem submetidos aos classificadores, uniformizando assim a entrada para os classificadores.

4.2.1 Seleção de Features

Dado que os modelos baseados em aprendizagem de máquina, utilizados neste estudo comparativo, calculam a probabilidade de um termo de entrada x pertencer a um rotulo y, é necessário definir como esses dados de entrada serão organizados e quais características desses termos serão analisadas. A análise das principais *features* utilizadas em outras pesquisas da área auxiliam nesta definição.

A definição das *features* utilizadas nos modelos de classificação de aprendizagem de máquina supervisionada leva em consideração quais, dessas *features*, são usualmente presentes nos trabalhos de Reconhecimento de Entidade Nomeada, que fazem uso desses classificadores. Trabalhos como o de Fersini [Fersini et al. 2014], Liu [Liu e Zhou 2013], Saha [Saha e Ekbal 2013], Ekbal [Ekbal e Saha 2012], Ling [Ling e Weld 2012], Rocktäschel [Rocktäschel et al. 2012] e Atkinson [Atkinson e Bull 2012] contribuíram para a determinação das *features*.

A maioria das *features* são extraídas das características internas dos termos, o que é vantajoso para trabalhos de diversos domínios e idiomas. As *features* são adquiridas na etapa de pré-processamento, e são utilizadas como entrada para os classificadores de entidades nomeadas. A seguir são listadas as *features* utilizadas para extrair as características de cada termo.

1. **w:** termo do texto sendo analisado no momento;

[2]https://opennlp.apache.org/

2. **p:** POS *tagging*, ou seja, a informação sintática de cada termo, etiquetando-os como Substantivo (NN), Substantivo Próprio (NNP), Verbo (VB), Preposição (IN), entre outros;
3. **cap:** determina se o termo está escrito com todas ou a primeira letra maiúscula;
4. **num:** indica se o termo é um número;
5. **hif:** informa que o termo é composto com a ocorrência do hífen;
6. **fst:** identifica se o termo é o primeiro da sentença;
7. **Prefixo:** representando o prefixo do termo, em duas *features*: pfx1, que indica a primeira letra do termo, e pfx2, sendo as duas primeiras letras;
8. **Sufixo:** informa a última letra do termo − sfx1, e as duas últimas − sfx2;
9. **length:** informa o comprimento do termo, ou seja, o número de caracteres;
10. **lemma:** indica a forma lematizada do termo;
11. **sig:** Identifica para sinais de pontuação, como DOT − ocorrência de ponto. E também COMMA − ocorrências de vírgula.

Para a definição do conjunto de *features*, é levado em consideração, selecionar características que podem ser extraídas de qualquer tipo de base textual. Portanto não ocorre nesse estudo comparativo, casos de uso de *features* específicas para um determinado domínio de estudo, ou idioma. Trabalhos que utilizam o REN para anotar genes e proteínas [Settles 2005], por exemplo, possuem *features* que são próprias para estes tipos de termos, e não se aplicariam em demais domínios. Desta maneira, a comparação voltada às *features* pode ser útil para trabalhos mais variados que façam uso de REN.

As alterações no conjunto de características extraídas das palavras modificam a saída de um classificador de entidades nomeadas [Tkachenko e Simanovsky 2012]. Tais variações auxiliam a determinar o melhor conjunto de *features* que devem ser utilizados, visando ajustar a exatidão da identificação e classificação dos termos.

As *features* selecionadas para este trabalho informam características para cada uma das palavras, como dados de seus formatos internos. Além disso, são selecionadas *features* de contexto, que servem para indicar características dos termos antecessores e sucessores.

No que diz respeito a obtenção das características dos termos baseado na sua adjacência, as *features* que indicam o termo (w), o POS *tagging* (p) e os termos com letra maiúscula (cap) são comparadas em intervalos de 3, 5 e 7 sequência de termos. A Tabela 4.7 demonstra como as *features* de contexto são aplicadas em relação ao termo atual, para cada intervalo.

Tabela 4.7: Intervalo de adjacência da sequência de termos.

Intervalo de adjacência	Features
3	w-1, w, w+1
5	w-2, w-1, w, w+1, w+2
7	w-3, w-2, w-1, w, w+1, w+2, w+3

Em cada intervalo adotado, são incluídas, para cada palavra, as *features* 'w', 'p' e 'cap' de termos antecessores e sucessores, consequentemente aumentando o número de informações de entrada para o classificador de entidades nomeadas. Tais *features* permitem identificar relações entre as palavras que possam relacionar com os rótulos de entidade nomeada.

A Figura 4.3 mostra o formato das *features* que podem ser extraídas de um termo 'w', colocadas lado a lado, na mesma linha. Neste exemplo é adotado o intervalo de 5 termos adjacentes para compor as *features* de contexto. Todos os termos possuem suas *features* distribuídas neste formado, por linha.

w p cap fst num hif pfx2 pfx1 sfx2 sfx1 w-2 w-1 w+1 w+2 p-2 p-1 p+1 p+2 length lemma sig

Figura 4.3: Organização das *features* para cada termo.

A Figura 4.4 apresenta um trecho de texto retirado do corpus de treinamento do CoNLL-03. A aquisição das *features* para os termos dessa sentença é apresentada na Figura 4.5, a qual inclui *features* dos dois termos antecessores e sucessores de cada termo.

A primeira linha da Figura 4.5 contém as *features* para a palavra Fischler, que é a primeira da sentença, seguindo a ordem apresentada na Figura 4.3. As *features* relativas ao 'num', 'hif' e 'sig' não ocorrem nesta linha, por não estarem relacionadas ao termo Fischler. Enquanto as *features* 'p-2', 'p-1' não ocorrem porque, como se trata do primeiro termo, não há POS *tagging* dos dois termos antecessores.

Fischler proposed EU measures after reports from Britain and France.

Figura 4.4: Sentença do corpus CoNLL-03.

Algumas das *features* aparecem apenas quando há a sua ocorrência relacionada ao termo. A quantidade de *features* depende de cada termo, caso o termo inicie com a letra minúscula a *feature* 'cap' não é mostrada, assim como a *feature* 'fst' apenas aparece se o termo for o primeiro da sentença.

Para comparar o impacto das *features* em cada classificador é adotado um modelo de comparação semelhante ao presente no trabalho de Saha [Saha et al. 2009], no

Fischler NNP cap fst pfx2=Fi pfx1=F sfx2=er sfx1=r <START1>@-2 <START0>@-1 proposed@1 EU@2 VBN@1 NNP@2 8lgth [Fischler]

proposed VBN pfx2=pr pfx1=p sfx2=ed sfx1=d <START0>@-2 Fischler@-1 EU@1 measures@2 NNP@-1 NNP@1 VBZ@2 8lgth [propose]

EU NNP cap pfx2=EU pfx1=E sfx2=de sfx1=e Fischler@-2 proposed@-1 measures@1 after@2 NNP@-2 VBN@-1 VBZ@1 IN@2 2lgth [eu]

measures VBZ pfx2=me pfx1=m sfx2=es sfx1=s proposed@-2 EU@-1 after@1 reports@2 VBN@-2 NNP@-1 IN@1 NNS@2 8lgth [measure]

after IN pfx2=af pfx1=a sfx2=er sfx1=r EU@-2 measures@-1 reports@1 from@2 NNP@-2 VBZ@-1 NNS@1 IN@2 5lgth [after]

reports NNS pfx2=re pfx1=r sfx2=ts sfx1=s measures@-2 after@-1 from@1 Britain@2 VBZ@-2 IN@-1 IN@1 NNP@2 7lgth [report]

from IN pfx2=fr pfx1=f sfx2=om sfx1=m after@-2 reports@-1 Britain@1 and@2 IN@-2 NNS@-1 NNP@1 CC@2 4lgth [from]

Britain NNP cap pfx2=Br pfx1=B sfx2=in sfx1=n reports@-2 from@-1 and@1 France@2 NNS@-2 IN@-1 CC@1 NNP@2 7lgth [Britain]

and CC pfx2=an pfx1=a sfx2=nd sfx1=d from@-2 Britain@-1 France@1 .@2 IN@-2 NNP@-1 NNP@1 O@2 3lgth [and]

France NNP cap pfx2=Fr pfx1=F sfx2=ce sfx1=e Britain@-2 and@-1 .@1 <END0>@2 NNP@-2 CC@-1 O@1 6lgth [France]

. . and@-2 France@-1 <END0>@1 <END1>@2 CC@-2 NNP@-1 DOT

Figura 4.5: Features adquiridas para os termos da sentença mostrada na Figura 4.4.

qual são comparados os resultados do REN para cada intervalo de adjacência, e em seguida é apresentada contribuição da adição gradual das *features*. O objetivo dessa comparação é identificar a contribuição, individual e em conjunto, que cada *feature* apresenta na tarefa de reconhecer entidades nomeadas, obtendo a combinação que melhor auxilia o REN para cada base de dados e rótulos de entidades nomeadas utilizados.

4.3 Avaliação

Esta seção apresenta conceitos adotados para calcular a eficácia dos classificadores, além de definições para validar os resultados de comparação. A seção 4.3.1 apresenta as medidas de precisão, cobertura e F1, e como são formuladas. Enquanto a seção 4.3.2 aborda, resumidamente, métodos de validação cruzada, incluindo denominado *k-fold*, aplicado no Capítulo 5.

4.3.1 Medidas

As medidas de avaliação informam o grau de acerto no reconhecimento das entidades nomeadas. A maioria dos trabalhos nessa área utilizam três medidas para esse fim, as quais são: precisão, cobertura (*recall*) e a medida F_β [Davis e Goadrich 2006].

Estas medidas são formuladas através das variáveis: TP (*True Positive*) - indicando que o termo é corretamente identificado, como uma entidade nomeada, e recebe a classificação correta do rótulo de entidade nomeada; FP (*False Positive*) - caso em que um termo que não é entidade nomeada, ou não pertence a um determinado rótulo, é classificado como tal; FN (*False Negative*) - indica que o termo de um determinado rótulo de entidade nomeada, não é classificado como tal.

A precisão é a medida que calcula a porcentagem de acerto na classificação do termos. Ajuda a indicar a qualidade do sistema no processo de classificação das entidades. A Equação 2.3 apresenta o cálculo da precisão, formulada através da razão entre as entidades TP e a soma de TP e FP.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{4.1}$$

Cobertura é a medida que calcula o número total de entidades identificadas sobre o número total de entidades presentes no texto. Ou seja, essa medida indica a porcentagem de entidades selecionadas corretamente. A Equação 2.4 mostra o cálculo dessa medida que é formulada pela razão de TP e a soma de TP e FN.

$$\textit{Cobertura} = \frac{TP}{TP + FN} \tag{4.2}$$

Já a medida F_β, caracteriza-se como um valor que, harmonicamente, combina as duas medidas anteriores. Originalmente esta formula pode ser descrita da seguinte maneira: $F = \frac{(\beta^2+1)P*R}{(\beta^2*P)+R}$, em que P indica a medida de precisão, e R indica a medida de cobertura. Com a atribuição de $\beta = 1$, esta medida é formulada como mostrado na Equação 2.5.

$$F_1 = \frac{2 * P * R}{P + R} \tag{4.3}$$

Considerando como exemplo, um texto contendo 100 ocorrências de entidades nomeadas cujo rótulo é Pessoa. O sistema classifica como Pessoa 80 termos, dos quais 50 estão classificados corretamente (TP). Neste caso a cobertura, para este rótulo é de 50%, indicando que dentre os 100 termos presentes no texto 50 são identificados e classificados corretamente. Já a precisão do REN fica em torno de 62%, em relação aos 80 classificados como Pessoa.

4.3.2 Validação

A avaliação dos classificadores é utilizada para comparar seus resultados, no REN, quando submetidos a distintos corpora. Para demonstrar os resultados das medidas de precisão, cobertura e F1, as bases de dados são divididas em corpora de treino e teste. A definição destas porções de dados é baseada no método de avaliação conhecido na literatura como *holdout* ([Kohavi et al. 1995]).

O método *holdout* particiona os dados em dois conjuntos mutualmente exclusivos, os quais são direcionados para realização de treinamento e teste, respectivamente. Na utilização deste método, é comum designar 2/3 dos dados para treino e 1/3 para teste. A separação destes conjuntos de dados são apresentadas nas tabelas 4.1, 4.3, 4.4 e 4.6.

Além desta validação, através do *holdout*, utilizado para comparar a eficácia dos classificadores, definiu-se empregar o método de validação cruzada ([Bailey e Elkan 1993]) *k-fold* ([Kohavi et al. 1995]). Neste método de validação cruzada, a base de dados D é dividida em k subconjuntos de dados (D_1, D_2, ..., D_k), de aproximadamente mesmo tamanho e mutualmente exclusivos. O classificador é treinado e testado k vezes, alternando o subconjunto k, que é utilizado para teste, enquanto os demais subconjuntos $(k-1)$ treinam o classificador.

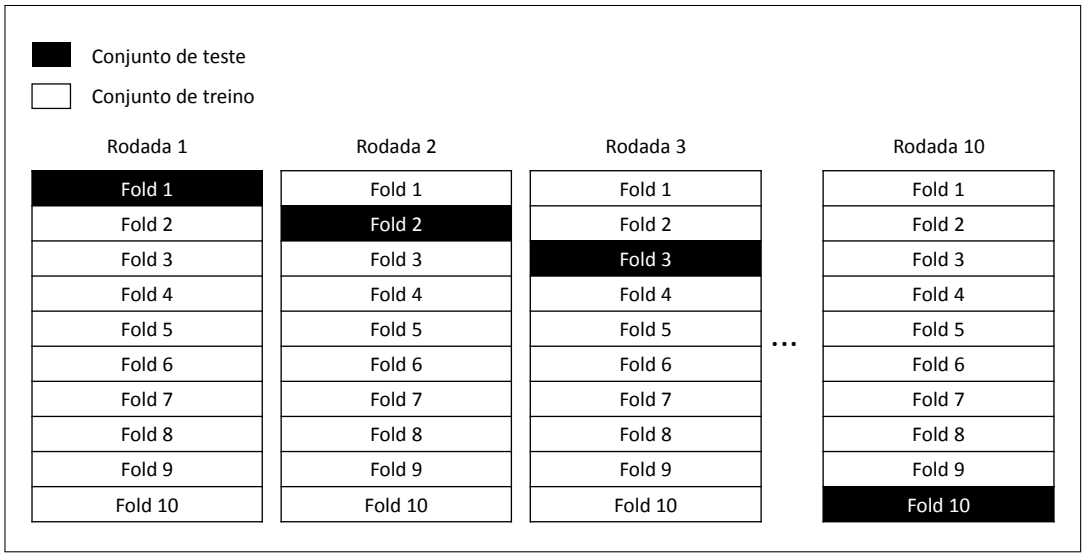

Figura 4.6: Ilustração do método k-fold.

Na Figura 4.6, é apresentado um esquema para exemplificar o método *k-fold*. Neste exemplo, o conjunto de dados é divido em dez subconjuntos, de igual tamanho. Cada rodada utiliza um dos subconjuntos para teste, enquanto os nove subconjuntos restantes compõem os dados de treino. São realizadas dez rodadas, para que cada subconjunto seja utilizado para o teste uma vez. Por fim, para obter-se a eficácia geral do classificador, utiliza-se a média dos resultados adquiridos nas dez rodadas.

Segundo Witten e Frank [Witten e Frank 2005] extensivos testes mostram que definir 10 subconjuntos, no método *k-fold* é a melhor forma para obter uma estimativa precisa na classificação. Por meio da validação cruzada, há uma distribuição melhor dos dados, para treino e teste do classificador.

Neste trabalho, o método *k-fold* é empregado para validar os resultados de comparação dos classificadores com o corpus em português, derivado da base WikiNER (Seção 5.1). A validação cruzada utilizada, consta na divisão do corpus em dez, subconjuntos, seguindo o esquema exemplificado na Figura 4.6, permitindo que todos os subconjuntos deste corpus sejam testados durante as etapas, informando a eficácia dos classificadores, por meio das medidas de precisão, cobertura e F1.

O Cápitulo 5 traz os experimentos realizados, que possibilitam comparar os classificadores de REN, baseados em aprendizagem de máquina. Comparando ainda, a eficácia destes classificadores, quando submetidos a textos em português.

Capítulo 5

Experimentos Comparativos

Este capítulo apresenta os experimentos realizados, que permitem comparar classificadores baseadas em aprendizagem de máquina. O uso de diferentes *features* e corpus, são também utilizados no estudo comparativo.

O primeiro experimento (Seção 5.1) permite comparar qual dos classificadores, baseados em aprendizagem de máquina, possui melhores resultados para a identificação e classificação das entidades nomeadas. Para isso os classificadores são comparados nos corpora definidos na Seção 4.1. Este experimento serve também, para definir qual classificador é utilizado no estudo que avalia a seleção de *features*.

Nesta mesma seção, há uma validação cruzada, realizada por meio do método *k-fold*, a qual permite comparar a eficácia dos classificadores, no corpus em português derivado do WikiNER. Uma vez que a grande maioria dos trabalhos comparativos de REN, envolve apenas classificadores treinados para o idioma inglês.

O estudo do uso de diferentes conjuntos de *features* no reconhecimento de entidades nomeadas constitui a Seção 5.2, demonstrando como cada característica dos termos contribui, individualmente e em conjunto, para o reconhecimento das entidades nomeadas.

A Seção 5.3 estuda o comportamento dos classificadores para diferentes rótulos. Os resultados da identificação e classificação são apresentados para cada rótulo de entidade nomeada, avaliando as relações de exatidão que as envolvem.

Os modelos de REN que compõem os classificadores comparados neste trabalho são desenvolvidos com o auxílio do quite de ferramentas Mallet [McCallum 2002]. Este quite de ferramentas é implementado em Java para processamento estatístico de linguagem natural, classificação de documentos, Clustering, Topic Modeling, Extração de Informação, entre outras aplicações de aprendizado de máquina para processamento textual.

Cada classificador é comparado, quanto ao seu desempenho na identificação e classificação das entidades nomeadas, por meio das medidas de precisão, cobertura e a

medida F1 (Seção 4.3.1). A contribuição de cada *feature* é analisada nos corpora adotados por este estudo comparativo, os quais possuem características distintas, como o número de documentos, a quantidade de rótulos de entidades nomeadas, o número de suas ocorrências e a extensão do vocabulário.

5.1 Estudo comparativo sobre a identificação e classificação de Entidades Nomeadas

Cada um dos classificadores calcula a probabilidade de um termo de entrada fazer parte de um rótulo de entidade nomeada. Nesta seção os classificadores são comparados quando submetidos aos corpora derivados das bases selecionadas: CoNLL-03, WikiNER-eng, WikiNER-pt e HAREM. Neste experimento, os corpora WikiNER-pt e HAREM são constituídos de textos em português, a fim de contribuir com a comparação de classificadores baseados em aprendizagem de máquina neste idioma.

Deseja-se também observar com este experimento se, em bases com características como idioma, tamanho e número de rótulos de entidades nomeadas distintos, ocorre algum tipo de variação de desempenho entre os classificadores.

São definidos os mesmos padrões de entrada dos dados, para a comparação dos classificadores. De modo que, para todos os corpora, a entrada dos dados é constituída do conjunto completo de *features* extraídas de cada termos. As *features* de contexto, que adicionam características dos termos antecessores e sucessores, não são incluídas, uma vez que a Seção 5.2 compara a contribuição dessas *features*, inclusive demonstrando o aumento da acurácia do REN.

A Tabela 5.1 apresenta o resultado dos experimentos (por meio das medidas de precisão, cobertura e F1) realizados com o CRF, MEMM, e HMM nos corpora selecionados. Nesta tabela é possível observar a diferença entre as medidas, gerais de REN, de cada classificador.

Tabela 5.1: Comparação dos Classificadores CRF, MEMM e HMM para cada corpus.

Corpus	Medidas	CRF	MEMM	HMM
CoNLL-03	Prec.	**0,7647**	0,7102	0,6668
	Cobert.	**0,7875**	0,7169	0,5667
	F1	**0,7759**	0,7135	0,6127
WikiNER-eng	Prec.	**0,7993**	0,7038	0,7880
	Cobert.	**0,7805**	0,6607	0,6581
	F1	**0,7897**	0,6816	0,7172
WikiNER-pt	Prec.	**0,7961**	0,7290	0,7891
	Cobert.	**0,7909**	0,6729	0,6466
	F1	**0,7934**	0,6998	0,7107
HAREM	Prec.	0,6011	0,5547	**0,6820**
	Cobert.	**0,5160**	0,5003	0,1588
	F1	**0,5346**	0,5260	0,2576

Os valores mostrados por este experimento revelam que, de modo geral, o classificador que utiliza CRF supera os demais classificadores nos corpora testados, confirmando conclusões obtidas em trabalhos como o de [Lafferty et al. 2001]. Quando submetido ao corpus HAREM, o HMM adquire valor de precisão superior ao CFR e MEMM, porém a cobertura é inferior, atingindo cerca de 16% das entidades nomeadas presentes no corpus. Essa diferença indica que mesmo com um valor superior de precisão, o CRF classificou mais termos identificados como entidade nomeada, por possuir cobertura de, aproximadamente, 52%.

Nos corpora em português, o comportamento do classificador CRF também demonstra melhor acurácia no REN, comparado com o MEMM e HMM. Os resultados do CRF no corpus WikiNER-pt revelam que a precisão atinge aproximadamente 80%, contra próximos 73% e 79% dos classificadores MEMM e HMM, respectivamente. Já a cobertura do CRF, neste mesmo corpus, demonstra uma diferença maior de resultados, em comparação com o MEMM e HMM, apontando que o CRF identificou mais termos como entidade nomeada.

O corpus HAREM apresenta os menores valores de REN para todos os classificadores. Os classificadores CRF e MEMM identificaram aproximadamente metade dos termos que são entidades nomeadas, dos quais 60% são classificados corretamente pelo CRF, enquanto o MEMM classifica corretamente 55%. A cobertura do HMM é a mais inferior, identificando aproximadamente 16% das entidades nomeadas presentes no corpus.

No corpus WikiNER-pt o CRF atinge valores competitivos com os corpora em inglês. O CRF, neste corpus, atinge praticamente a mesma precisão, e uma cobertura cerca de 1% maior, em relação ao corpus WikiNER-eng. A superioridade de valores do CRF é um pouco maior, quando comparado com o resultado obtido do corpus CoNLL-03.

Em um corpus com menor quantidade de termos, como o HAREM, o classificador baseado no modelo oculto de Markov (HMM) não demonstra a mesma eficácia para realizar o REN, considerando os resultados obtidos pela Tabela 5.1. Este classificador não atinge valores de cobertura competitivos, significando que o HMM revela uma relação de dependência para identificar as entidades nomeadas, de acordo com a quantidade destes termos no corpus de treinamento.

A utilização do método *k-fold* é aplicado para executar uma validação cruzada, sobre os classificadores. Neste processo, o corpus derivado do WikiNER, com textos em português, é dividido em dez arquivos (Tabela 5.2) que são utilizados pelos classificadores CRF, MEMM e HMM. São realizados, para cada classificador, dez treinos (contando com nove arquivos) e testes, sempre alternando o arquivo de teste.

Tabela 5.2: Dados quantitativos para os dez subconjuntos de dados derivados do corpus WikiNER-pt.

Subconjuntos	Termos	Sentenças	LOC	ORG	PER
WikiNER-pt1	308309	11968	24447	3397	6466
WikiNER-pt2	308535	11752	4208	18715	10055
WikiNER-pt3	308348	11922	20258	3054	11453
WikiNER-pt4	308320	11967	20342	2508	9888
WikiNER-pt5	307625	12648	25905	2385	6079
WikiNER-pt6	307682	12604	31101	2179	4275
WikiNER-pt7	308161	12114	17026	3789	10990
WikiNER-pt8	308225	12077	14187	4046	11587
WikiNER-pt9	308238	12036	13045	5521	11649
WikiNER-pt10	307985	12314	14804	4976	11982

Os dados mostrados na Tabela 5.2 informam o número de termos, sentenças e entidades nomeadas dos rótulos: LOC, ORG e PER. A divisão dos dez subconjuntos buscou selecionar a mesma quantia de dados textuais, entretanto evitando interromper as sentenças entre os arquivos. Por conta deste fator, o número de termos não é precisamente igual. Existe uma variação no tamanho dos subconjuntos, porém não reflete uma diferença expressiva, respeitando a condição de dez subconjuntos de dados, com tamanhos aproximadamente iguais e mutualmente exclusivos.

A diferença na quantidade de termos, para cada rótulo de entidade nomeada, demonstra como é dada a distribuição destas entidades no corpus WikiNER-pt. O rótulo LOC, na maioria dos subconjuntos, ocorre mais vezes, seguido de PER e ORG. A exceção desta ordem é vista no subconjunto 'WikiNER-pt2', onde ocorrem mais termos de rótulo ORG.

A Tabela 5.3 apresenta os resultados, para os classificadores CRF, MEMM e HMM, quando submetidos ao método de validação cruzada *k-fold*. Cada atributo desta tabela contém os valores de precisão, cobertura e F1, referentes ao teste realizado

com um subconjunto da base WikiNER-pt. Para cada resultado de teste, o treinamento dos classificadores é executado com o os nove subconjuntos restantes. Cada classificador é treinado e testado dez vezes.

Tabela 5.3: Validação cruzada, da comparação dos classificadores CRF, MEMM, e HMM, utilizando subconjuntos derivados da base WikiNER-pt.

Subconjuntos	CRF			MEMM			HMM		
	Prec.	Cob.	F1	Prec.	Cob.	F1	Prec.	Cob.	F1
WikiNER-pt1	**0,8831**	**0,8577**	**0,8703**	0,8325	0,7643	0,7970	**0,8781**	0,7412	0,8039
WikiNER-pt2	0,87383	0,8417	0,8575	**0,8638**	**0,8192**	**0,8410**	0,8676	**0,7504**	**0,8048**
WikiNER-pt3	0,8574	0,8311	0,8441	0,8408	0,7952	0,8174	0,8336	0,7358	0,7817
WikiNER-pt4	0,8792	0,8213	0,8493	0,7929	0,7251	0,7575	0,8423	0,6807	0,7530
WikiNER-pt5	0,8541	0,8447	0,8494	0,6947	0,7199	0,7071	0,8364	0,7264	0,7776
WikiNER-pt6	0,8164	0,8138	0,8151	0,7707	0,7805	0,7756	0,7855	0,6856	0,7322
WikiNER-pt7	0,8245	0,8063	0,8154	0,7953	0,7970	0,7961	0,8005	0,7150	0,7554
WikiNER-pt8	0,8123	0,8015	0,8069	0,7558	0,7482	0,7520	0,7931	0,7094	0,7489
WikiNER-pt9	0,7815	0,7783	0,7800	0,7714	0,7644	0,7679	0,7902	0,6945	0,7393
WikiNER-pt10	0,8226	0,8099	0,8162	0,6744	0,5822	0,6249	0,8040	0,6996	0,7482

Os valores da Tabela 5.3 validam os eficácia superior do CRF, em comparação aos classificadores MEMM e HMM. Para o CRF, quando o subconjunto 'WikiNER-pt1' é utilizado para o teste, e os demais subconjuntos treinam o classificador, as medidas de precisão, cobertura e F1 atingem os maiores valores, respectivamente próximos 88%, 86% e 87%.

Já para o classificador MEMM, o maior eficácia do REN, é alcançada utilizando, para teste, o subconjunto 'WikiNER-pt2'. Neste caso, a precisão, a cobertura e a medida F1, são aproximadamente 86%, 82% e 84%, respectivamente. Neste mesmo subconjunto, o classificador HMM alcança maior cobertura e F1, em cerca de 75% e 80%, respectivamente. A precisão com maior valor, no entanto, é adquirida no corpus de teste 'WikiNER-pt1'.

Enquanto o CRF é em todos os dez subconjuntos, com medidas acima de 80% em nove execuções, o comportamento dos classificadores MEMM e HMM variam no alcance de melhor eficácia. O classificador MEMM é superior ao HMM, segundo a media F1, em sete casos. Pois, quanto utilizados os subconjuntos 'WikiNER-pt1', 'WikiNER-pt5' e 'WikiNER-pt10', o HMM obtém maior eficácia.

Tabela 5.4: Média dos resultados mostrados na Tabela 5.3.

	CRF			MEMM			HMM		
Média	Prec.	Cob.	F1	Prec.	Cob.	F1	Prec.	Cob.	F1
10 Subconjuntos	0,8404	0,8206	0,8304	0,7792	0,7496	0,7636	0,8231	0,7138	0,7645

A Tabela 5.4 apresenta a média dos resultados comparativos, através da validação cruzada, presentes na Tabela 5.3. A média destes valores informa que, no corpus

WikiNER-pt, o CRF apresenta precisão, cobertura e F1, de 84%, 82% e 83%, respectivamente. Enquanto o MEMM possui aproximadamente, 78%, 75% e 76% de precisão, cobertura e F1, respectivamente. Por fim, o HMM aponta precisão média próxima a 82% (superior ao MEMM), cobertura próxima de 71% e F1 de cerca de 76%. Segundo esses valores, percebe-se que a cobertura do MEMM é superior à demonstrada pelo HMM, entretanto o mesmo não pode ser dito da precisão.

Uma vez que, inclusive em textos em português, o CRF apresenta valores superiores quando comparado aos classificadores MEMM e HMM, o estudo que compara a contribuição das *features*,individualmente, em grupo, na Seção 5.2, utiliza este classificador.

5.2 Estudo comparativo para diferentes seleções de features

As comparações sobre distintas seleções de *features* avaliam suas contribuições para o REN. Cada combinação distinta possui um impacto na identificação e classificação das entidades nomeadas. Os experimentos realizados nesta seção visam observar qual a melhor condição para utilizar características que são adquiridas dos termos. A Seção 5.2.1 apresenta os experimentos que avaliam o impacto das *features* individualmente, além de demonstrar o comportamento do REN para cada par de *features*. Enquanto a Seção 5.2.2 contém experimentos que estudam o comportamento das *features* de contexto.

5.2.1 Estudo para avaliar o impacto individual das features

Os experimentos desta seção avaliam a contribuição individual das *features* extraídas de cada termo, sem considerar sua relação com termos antecessores e sucessores. Neste estudo, as *features* são adicionadas individualmente ao conjunto de *features* avaliadas. Com essa adição, pretende-se observar quais *features* contribuem mais para aumentar os valores das medidas de reconhecimento de entidades nomeadas.

Cada uma das *features* extraídas das palavras, objetiva auxiliar a identificação e classificação das entidades. A observação da contribuição individual das *features* é demonstrada por duas etapas. Na primeira etapa são executados experimentos com apenas a própria palavra e a *feature* em questão, demonstrado por meio da Tabela 5.5. Já na segunda etapa, Tabela 5.7, são realizados testes que adicionam, uma a uma, as *features*, observando quais são mais relevantes para melhorar o desempenho do REN.

Cada *feature* indica uma característica do termo. O conjunto dessas características é usado para diferenciá-los, permitindo identificá-los e classificá-los como entidades

nomeadas. A própria palavra é uma *feature* que pode ser usada como entrada do classificador baseado em aprendizagem de máquina. Juntamente com a palavra a ser rotulada, cada *feature* é colocada individualmente, e verificada a acurácia do CRF para realizar o REN.

O número de iterações de treino é distinto para cada *feature* usada, uma vez que influenciam o cálculo da probabilidade para identificar e classificar as entidades nomeadas de formas diferenciadas. Por conta deste fato, os experimentos da primeira etapa, possuem o número de iterações fixado, para cada corpus.

Na base CoNLL-03, o número de iterações é fixado em 78, dado que o classificador CRF encerra seu treinamento com essa quantidade de iterações, quando recebe apenas o termo como entrada para realizar o REN. Enquanto para a base WikiNER-eng, o número de iterações é fixado em 40, valor de iterações de treinamento quando usadas as *features* 'w' e 'cap'. Para cada corpus, as demais *features* realizaram o treinamento com mais iterações. Entretanto, busca-se com o experimento mostrado na Tabela 5.5, comparar a contribuição para o REN, que as *features* selecionadas possuem com a mesma quantidade de iterações de treinamento.

Tabela 5.5: Comparação das features individualmente. Utilizando CRF e as bases CoNLL-03, com iteração igual a 78, e a base WikiNER-eng, com iteração igual a 40.

Features	CoNLL-03			WikiNER-eng		
	Prec.	Cobert.	F1	Prec.	Cobert.	F1
w, affixes	**0,7328**	**0,7492**	**0,7400**	**0,5796**	**0,5743**	**0,5753**
w, cap	0,6766	0,6738	0,6637	0,5040	0,4907	0,4760
w, p	0,6717	0,6418	0,6530	0,5159	0,4720	0,4701
w, sig	0,6936	0,4465	0,5331	0,4489	0,1339	0,2037
w, num	0,7089	0,4362	0,5327	0,4541	0,0948	0,1502
w, fst	0,7054	0,4130	0,5197	0,6017	0,0999	0,1711
w, hif	0,6997	0,4198	0,5119	0,4422	0,1026	0,1639
w	0,6804	0,3993	0,4877	0,4265	0,0900	0,1460
w, length	0,5643	0,4379	0,4805	0,3795	0,0760	0,1210
w, lemma	0,5318	0,4056	0,4602	0,4317	0,1239	0,1843

Segundo o experimento mostrado na Tabela 5.5, a *feature* 'affixes' (formada pelo prefixo e sufixo de cada palavra) atinge maior valor para as medidas de precisão, cobertura e F1. As *features* 'p' e 'cap' são características dos termos que também merecem destaque, uma vez que atingem valores superiores às demais *features*, em ambos os corpora. As medidas de cobertura apresentadas pelo CRF no corpus CoNLL-03, e com as *features* 'affixes', 'p' e 'cap', destacam que acima de 64% das entidades nomeadas são identificadas, enquanto as demais *features*, individualmente, identificam menos de 45% das entidades nomeadas.

Um comportamento semelhante é observado com os resultados do CRF no corpus WikiNER-eng, com um diferencial ainda maior. Enquanto as *features* 'affixes', 'p' e

'cap' alcançam cobertura de, respectivamente, cerca de 57%, 49% e 47%, as demais *features* não são capazes de identificar, individualmente, mais que 13% das entidades nomeadas. Vale lembrar que, para o corpus WikiNER-eng, são feitas apenas 40 iterações de treinamento, entretanto o experimento demonstra o potencial de contribuição individual das *features* 'affixes', 'p' e 'cap', em comparação às demais.

A Tabela 5.6 apresenta a relação de *features*, que são adicionadas gradualmente, e seus respectivos códigos de referência. Estes códigos são utilizados para representar cada conjunto de *features*, selecionado como entrada do classificador CRF, para observar a contribuição que a adição das *features* proporciona para o REN, como demonstrado Tabela 5.7.

Tabela 5.6: Códigos para referenciar features

Features	Cód. Features
w	w
w, p	wp
w, p, cap	wpc
w, p, cap, num	wpcn
w, p, cap, num, hif	wpcnh
w, p, cap, num, hif, fst	wpcnhf
w, p, cap, num, hif, fst, sig	wpcnhfs
w, p, cap, num, hif, fst, sig, affixes	wpcnhfsa
w, p, cap, num, hif, fst, sig, affixes, length	wpcnhfsal
w, p, cap, num, hif, fst, sig, affixes, length, lemma	wpcnhfsall

Da mesma forma determinada para a comparação individual das *features*, mostrada na Tabela 5.5, o número de iterações para o treinamento do classificador CRF é fixado para cada corpus. O objetivo dessa definição é identificar qual das *features* contribui com um maior diferencial nas medidas de precisão, cobertura e F1, ao ser adicionada ao conjunto anterior. No corpus CoNLL-03, ao ser adicionada a *feature* 'cap' o treino é encerrado após 68 iterações, logo para este corpus os resultados são mostrados obedecendo este mesmo critério. No corpus WikiNER-eng, ao adicionar a *feature* 'cap', o classificador encerra o treinamento com 85 iterações, por este motivo, os valores para este corpus são referentes a esta quantidade de iterações de treino.

Tabela 5.7: Comparação, por meio da adição gradual ao conjunto das *features*. Utilizando CRF com os corpora CoNLL-03, com iteração igual a 68 e WikiNER-eng, com iteração igual a 85.

Features	CoNLL-03			WikiNER-eng		
	Prec.	Cobert.	F1	Prec.	Cobert.	F1
w	0,6970	0,3659	0,4799	0,6349	0,2956	0,4034
wp	0,6685	0,6225	0,6329	0,6072	0,5838	0,5865
wpc	0,5521	0,6613	0,5918	0,6341	0,6129	0,6086
wpcn	0,6023	0,6757	0,6322	0,6156	0,61105	0,5986
wpcnh	0,6350	0,6708	0,6452	0,6147	0,6153	0,6009
wpcnhf	0,6438	0,6745	0,6583	0,6241	0,6044	0,5978
wpcnhfs	0,6026	0,6820	0,6350	0,6495	0,6287	0,6263
wpcnhfsa	0,7056	0,7467	0,7252	0,6693	0,6638	0,6626
wpcnhfsal	**0,7215**	0,7688	**0,7435**	0,6751	0,6666	0,6683
wpcnhfsall	0,7211	**0,7690**	0,7434	**0,7003**	**0,6946**	**0,6950**

Já os resultados dos experimentos apresentados na Tabela 5.7 mostram, por meio da adição de cada uma das *features*, que o reconhecimento de entidades nomeadas é mais eficaz, à medida que mais *features* são adicionadas. Esse comportamento já era esperado, uma vez que mais informações sobre as palavras auxiliam a identificá-las e classificá-las como entidades nomeadas.

A Tabela 5.7 revela que, quando usado o corpus CoNLL-03, a inclusão da *feature* 'affixes' é a que mais contribui para aumentar as medidas de REN em relação ao grupo anterior. A adição desta característica permitiu que a medida F1 aumentasse de, aproximadamente 63% para cerca de 72%. O mesmo comportamento é observado quando usado o corpus WikiNER-eng, no qual a variação, ao acrescentar a característica dos prefixos e sufixos, da medida F1 foi de aproximadamente 62% para 66%. Esses resultados validam a contribuição dessa *feature*, reforçando a importância de sua presença no conjunto de *features* para trabalhos de REN.

Quando adicionada a *feature* 'p', há um acréscimo considerável na identificação das entidades nomeadas. O valor da precisão, no corpus CoNLL-03 reduz de aproximadamente 70% para 67%. Assim como há uma redução da precisão, no corpus WikiNER-eng, de aproximadamente 63% para 60%. O ganho na acurácia do REN, entretanto, é demonstrada por meio dos valores da medida F1, que refletem a maior cobertura na identificação das entidades nomeadas, em ambos os corpora.

A inclusão da *feature* 'cap', no corpus CoNLL-03, reduz a precisão do REN, de aproximadamente 67% para 55%. Entretanto, há um ganho de cobertura, de 62% para 66%, demonstrando que esta *feature* mais auxilia a identificação das entidades nomeadas. Entretanto, no corpus WikiNER-eng, além de contribuir com o aumento da cobertura de, aproximadamente, 58% para 61%, a precisão também é beneficiada, aumentando de próximos 61% para 63%. Averiguar o porquê deste comportamento, na base em português, é uma tarefa que ficará para um trabalho futuro.

A fim de observar a correlação das *features*, a Tabela 5.8 apresenta os resultados do REN, por meio da medida F1, para cada combinação de de par de *features*. Porém, diferentemente do estabelecido nos experimentos das Tabelas 5.5 e 5.7, os resultados apresentados na Tabela 5.8 não são referentes a testes com o mesmo número de iteração de treino, dado que cada combinação de pares de *features* requer uma quantidade respectiva de iterações, durante a fase de treinamento, para identificar e classificar as entidades nomeadas. Os resultados da Tabela 5.8 mostram quais destes pares de *features* possuem maior e menor acurácia. Os resultados são referentes a execução do classificador CRF treinado e testado com a base CoNLL-03.

Tabela 5.8: Combinação em par das features, utilizando o classificador CRF e a base CoNLL-03.

Features	p	cap	num	hif	fst	sig	affixes	length	lemma
p	0,6993	0,5918	0,6124	0,6889	0,7178	0,6884	**0,7597**	0,6892	0,7180
cap	0,5918	0,6637	0,6251	0,6379	0,7174	0,6391	**0,7378**	0,6619	0,6973
num	0,6124	0,6251	0,6068	0,3758	0,6500	0,6462	**0,7535**	0,5549	0,6735
hif	0,6889	0,6379	0,3758	0,6163	0,6139	0,5653	**0,7391**	0,5528	0,6651
fst	0,7178	0,7174	0,6500	0,6139	0,6233	0,3621	**0,7640**	0,6086	0,6576
sig	0,6884	0,6391	0,6462	0,5653	0,3621	0,6063	**0,7588**	0,5118	0,6892
affixes	0,7597	0,7378	0,7535	0,7391	**0,7640**	0,7588	0,7388	0,7500	0,7595
length	0,6892	0,6619	0,5549	0,5528	0,6086	0,5118	**0,7500**	0,6194	0,6466
lemma	0,7180	0,6973	0,6735	0,6651	0,6576	0,6892	**0,7595**	0,6466	0,4602

Os resultados da Tabela 5.8, confirmam os resultados apresentados na Tabela 5.5, em relação ao uso da *feature* 'affixes'. Os resultados da medida F1, são superiores quando combinam cada uma das *features* com 'affixes', com todos os valores superiores à 70%. Exceto a própria *feature* 'affixes' que apresenta valor superior quando combinada com a *feature* 'fst'.

A *feature* 'p', além da combinação com 'affixes' que é cerca de 76%, atinge valor aproximado à 72% quando combinada com as *features* 'lemma' e 'fst'. Já os menores valores da medida F1 são vistos com as combinações de 'p' com 'cap', atingindo aproximadamente 59%, e 'p' com 'num', atingindo cerca de 61%.

Quando 'cap' faz par com as demais *features*, depois de 'affixes', o maior valor da medida F1 é da combinação com 'fst', assim como foi ocorre com 'p'. Além da combinação de 'cap' e 'p', a combinação de 'cap' e 'num', também demonstra valor inferior para o REN, segundo os valores da Tabela 5.8.

A *feature* 'num' possui a pior combinação, segundo o valor da medida F1, quando combinada com 'hif', alcançando aproximadamente 37%. Já o segundo maior valor, dentre pares de *features* com 'num', é visto na combinação com 'lemma', alcançando cerca de 67%. Quando Combinada com 'fst' a *feature* 'num' adquire o terceiro maior valor: 65%.

O uso da *feature* 'hif' é melhor aproveitado, além de 'affixes', quando combinado com 'p', trazendo a medida F1 a aproximadamente 69%, e quando faz par com 'lemma', com F1 $\cong$ 66%. Os pares que atingem menores valores são vistos nas combinações do 'hif' com a *feature* 'num' (já destacada no parágrafo anterior) e com 'length', F1 $\cong$ 55%.

A *feature* 'fst' quando faz par com 'p' e 'cap', atingem cerca de 72% para o REN. Não é tão eficaz quando faz par com a *feature* 'sig', com medida F1 próxima a 36%. Assim como 'length' não é um par dos mais eficazes quando combinado com 'fst', a medida F1 é cerca de 61%.

Por sua vez, a *feature* 'sig' demonstra melhores (além do par com 'affixes') combinações com as *features* 'p' e 'lemma', atingindo aproximadamente 69% na medida F1. Já quando combinada com 'length' atinge o segundo menor valor de F1, aproximadamente 51%.

No que diz respeito a *feature* 'length', a combinação com 'affixes' possui maior valor para a medida F1, 75%, e o segundo maior valor para o par com a *feature* 'p' (F1 $\cong$ 69%). A *feature* 'lemma' não combinada com outra *feature* possui medida F1 de aproximadamente 46%, entretanto o par com 'length' tem o menor valor de F1, pouco mais de 64%.

Em resumo, a *feature* 'affixes', individualmente e em conjunto, apresenta melhores resultados de classificação. A Tabela 5.8 revela que os piores resultados são obtidos com as combinações que incluem as *features* 'num' e 'length'.

A comparação de conjuntos distintos de *features* revela que, individualmente cada característica de um termo pode contribuir para auxiliar o aprendizado, durante o treinamento do classificador, permitindo a identificação e classificação das entidades nomeadas. A Seção 5.2.2 apresenta experimentos que estudam a relação de *features* referentes ao contexto em que o termo está empregado no texto.

5.2.2 Estudo sobre a variação da adjacência das features

Neste experimento é avaliado em qual intervalo de adjacência dos termos, os valores para as medidas que indicam a porcentagem de acerto no REN são maiores. Os textos em linguagem natural são compostos por palavras (termos) dispostas em sequência, de modo que toda palavra possui uma adjacência.

As *features* relacionadas a cada palavra do texto (w), ao POS *tag* (p) e a ocorrência de termos com letra inicial maiúscula, ou completamente maiúsculas (cap), são submetidas a testes de variação da adjacência neste experimento. As *features* 'p' e 'cap' são selecionadas, para compor as *features* de contexto, pois são as características extraídas dos termos que, individualmente, mais contribuíram para a identificação de entidades nomeadas (Tabelas 5.5 e 5.7), considerando que as *features* referentes ao afixo dos termos são compostas por dois prefixos e sufixos. Três variações de

adjacências são testadas. A Tabela 5.9 demonstra quais *features* são adicionadas para cada intervalo de adjacência.

Tabela 5.9: Variação do intervalo de adjacência para das palavras, do POS Tag e ´cap'.

Referência	Features do intervalo de adjacência
w3	w-1, w, w+1
w5	w-2, w-1, w, w+1, w+2
w7	w-3, w-2, w-1, w, w+1, w+2, w+3
p3	p-1, p, p+1
p5	p-2, p-1, p, p+1, p+2
p7	p-3, p-2, p-1, p, p+1, p+2, p+3
c3	cap-1, cap, cap+1
c5	cap-2, cap-1, cap, cap+1, cap+2
c7	cap-3, cap-2, cap-1, cap, cap+1, cap+2, cap+3

A adição das *features* de adjacência da palavra e do POS *tag* se comportam de forma distinta, em relação a *feature* 'cap'. Enquanto as duas primeiras ocorrem para todos os termos, palavras com letras maiúsculas ocorrem, apenas, em alguns termos. Desta forma, para todos os termos do texto, exceto o primeiro e o último, serão adicionadas *features* das palavras anteriores e sucessores. Já no que diz respeito a *feature* 'cap', apenas os termos que sucedem e/ou precedem uma palavra com letra maiúscula são acrescidas dessas *features*.

A inclusão de *features* de adjacência do 'cap' avaliam, se o acréscimo dessas características contribuem para melhorar o reconhecimento de entidades nomeadas. A Figura 5.1, mostra um trecho da base de dados CoNLL-03, com quatro termos dos quais dois são capitalizados. Já na Figura 5.2 é exposto como as *features* que indicam adjacência a um termo marcado como 'cap' é inclusa, além das características de adjacência da palavra e POS *tag* antecessor e sucessor.

Fischler proposed EU measures.

Figura 5.1: Sentença do corpus CoNLL-03.

```
Fischler NNP cap <START0>@-1 proposed@1 VBN@1
proposed VBN Fischler@-1 NNP@-1 cap@-1 EU@1 NNP@1 cap@1
EU NNP cappproposed@-1VBN@-1 measures@1 VBZ@1
measures VBZ EU@-1 NNP@-1 cap@-1 .@1 O@1
. . measures@-1 VBZ@-1 <END>@1
```

Figura 5.2: Features w, p, cap, w-1, p-1, cap-1, w+1, p+1 e cap+1 extraídas da sentença exibida na Figura 5.1.

Como exibido na Figura 5.2, quando um termo é sucessor de uma palavra que recebe a *feature* 'cap', este termo possui a *feature* cap@-1. Ao mesmo tempo que, os termos que são antecessores de uma palavra com 'cap' recebem a *feature* cap@1. Fazer uso dessas características é válido, uma vez que boa parte das entidades nomeadas, são nomes próprios, e iniciam com letras maiúsculas.

A Tabela 5.10 e a Tabela 5.11 apresentam os resultados dos testes realizados com variações da inclusão de *features* de contexto, com o corpus CoNLL-03. Já as Tabelas 5.12 e 5.13 mostram os resultados do mesmo procedimento, de comparação das *features* de contexto, com o corpus WikiNER-eng. São atribuídas características dos termos que antecedem e sucedem cada uma das palavras, e avaliado, em duas etapas, o melhor intervalo de adjacência para incluir no conjunto das *features*. Na primeira etapa são feitos testes, apenas com as *features* 'w', 'p' e 'cap' (Tabelas 5.10 e 5.12). Em seguida, para complementar e validar os resultados, são feitos testes que utilizam todas as *features* para cada um dos intervalos de adjacência (Tabelas 5.11 e 5.13).

Tabela 5.10: Variação das features de adjacência dos termos. Utilizando CRF e a base CoNLL-03.

Features	Precisão	Cobert.	F1
w3, p3, c3	0,7862	0,7966	0,7905
w5, p5, c5	**0,7944**	**0,8027**	**0,7982**
w7, p7, c7	0,7936	0,7977	0,7953

Tabela 5.11: Variação das features de adjacência dos termos, incluindo todo o conjunto de features. Utilizando CRF e a base CoNLL-03.

Features	Precisão	Cobert.	F1
w3, p3, c3, num, hif, fst, sig, affixes, length, lemma	0,8128	0,8279	0,8202
w5, p5, c5, num, hif, fst, sig, affixes, length, lemma	**0,8219**	**0,8325**	**0,8270**
w7, p7, c7, num, hif, fst, sig, affixes, length, lemma	0,8146	0,8288	0,8215

Os testes realizados na base CoNLL-03, demonstram que o intervalo de cinco sequência de termos, permite adquirir melhores características dos termos no texto, e consequentemente ter resultados superiores nas medidas de precisão, cobertura e F1. Não há uma diferença tão expressiva nos valores para cada intervalo de adjacência adotado.

Nas Tabelas 5.10 e 5.11, a precisão e cobertura do classificador CRF aumenta, cerca de 1% com o uso das *features* 'w5', 'p5' e 'c5', em comparação às *features* 'w3', 'p3' e 'c3'. A diferença em porcentagem dos valores do intervalo de cinco termos, em relação ao de sete termos adjacentes, também não é tão expressiva. Mesmo com os valores das medidas de REN aproximados, uma pequena diferença já aponta para um conjunto relevante de termos que podem ser identificados e classificados corretamente.

Tabela 5.12: Variação das features de adjacência dos termos. Utilizando CRF e a base WikiNER-eng.

Features	Precisão	Cobert.	F1
w3, p3, c3	0,7852	0,7814	0,7799
w5, p5, c5	0,8295	**0,8097**	**0,8180**
w7, p7, c7	**0,8299**	0,8066	0,8158

Tabela 5.13: Variação das features de adjacência dos termos, incluindo todo o conjunto de features. Utilizando CRF e a base WikiNER-eng.

Features	Precisão	Cobert.	F1
w3, p3, c3, num, hif, fst, sig, affixes, length, lemma	0,8291	0,8140	0,8206
w5, p5, c5, num, hif, fst, sig, affixes, length, lemma	0,8362	0,8169	0,8265
w7, p7, c7, num, hif, fst, sig, affixes, length, lemma	**0,8378**	**0,8196**	**0,8277**

O intervalo de três termos em sequência, nos resultados demonstrados na Tabela 5.12, demonstra valores com uma inferioridade maior ao visto na Tabela 5.10. Ao utilizar os conjuntos de distintas *features* de contexto, com o corpus WikiNER-pt, a precisão do classificador atinge, aproximadamente 78% com o intervalo de três termos adjacentes, contra cerca de 83% nos intervalos de cinco e sete adjacência de termos. De forma semelhante, a cobertura é inferior com as *features* de contexto 'w3', 'p3' e 'c3', identificando cerca de 78% das entidades nomeadas, enquanto ao utilizar as *features* de contexto 'w5', 'p5' e 'c5', o classificador CRF identifica, aproximadamente 81% das entidades nomeadas.

A inclusão de *features* de contexto demonstram, nos experimentos realizados neste capítulo, capacidade de melhorar a acurácia do REN. A comparação dos resultados apresentados pelo CRF na Tabela 5.1, para os corpora CoNLL-03 (F1≅ 78%) e WikiNER-eng (F1≅ 79%), com os resultados das Tabelas 5.11 e 5.13 confirmam a contribuição das *features* de contexto.

Na Seção 5.3, a comparação dos classificadores CRF, MEMM e HMM inclui os resultados de precisão, cobertura e F1 por rótulo de entidade nomeada, quando as *features* de contexto são adicionadas ao conjunto, revelando os rótulos que são identificados e classificados com maior acurácia.

5.3 Comparação com base nos rótulos de entidades nomeadas

As comparações dispostas nesta seção observam os rótulos de entidades nomeadas (PER, LOC, ORG e MISC) que são melhor identificados e rotulados pelos classificadores. O objetivo desta comparação é verificar qual o comportamento dos valores de acurácia, do reconhecimento dos rótulos de entidades nomeadas, com diferentes classificadores e bases de dados.

A Seção 5.2.2, apresenta resultados que demonstram maior acurácia para os classificadores de REN, quando são inclusas *features* de contexto. Por conta desta contribuição, os classificadores adotam, como entrada, o conjunto completo das *features* selecionadas (Seção 4.2.1), juntamente com as *features* de contexto 'w5', 'p5' e 'c5', que adicionam características dos dois termos antecessores e sucessores ao atual (Tabela 5.9).

A Tabela 5.14 mostra os valores de precisão, cobertura e F1 para cada um dos rótulos de entidades nomeadas, provenientes dos testes executados com os corpora em inglês: CoNLL-03 e WikiNER-eng. Estes valores são referentes ao classificador CRF. Enquanto a Tabela 5.16 possui os valores do CRF com os corpora em português: WikiNER-pt e HAREM.

Além disso, as Tabelas 5.15 e 5.17 demonstram, quantitativamente, como é dada a classificação dos termos em entidades nomeadas (PER, LOC, ORG e MISC) e dos termos que são rotulados como não sendo entidades nomeadas (O). Estes valores permitem observar a quantidade de termos que são atribuídos aos seus respectivos rótulos, e os que são classificados de forma incorreta. Os dados são equivalentes ao classificador CRF, quando treinado e testado com os corpora derivados das bases CoNLL-03, WikiNER-eng, WikiNER-pt e HAREM.

Tabela 5.14: Medidas por rótulo de entidade nomeada, utilizando o classificador CRF e as bases CoNLL-03 e WikiNER-eng.

Rótulo de	CoNLL-03			WikiNER-eng		
Entidade Nomeada	Prec.	Cobert.	F1	Prec.	Cobert.	F1
PER	**0,8712**	**0,9217**	**0,8957**	**0,8717**	**0,9207**	**0,8955**
LOC	0,8566	0,8744	0,8654	0,8369	0,8657	0,8511
ORG	0,7895	0,7816	0,7856	0,8319	0,7261	0,7754
MISC	0,7703	0,7525	0,7613	0,8045	0,7551	0,779

Analisando os valores presentes nas Tabelas 5.14 e 5.16, fica claro que o rótulo de entidades nomeada referente a ocorrência de nomes de pessoas (PER) possui maior exatidão. Todos os valores para este rótulo superam 87% nos corpora derivados das bases CoNLL-03, WikiNER-eng e WikiNER-pt, enquanto na base HAREM os valores superam 70%. Para ambos os corpora em inglês, o classificador identifica e classifica corretamente (cobertura) cerca de 92% dos termos cujo rótulo é PER. A precisão, para o corpus CoNLL-03, é cerca de 87%, praticamente o mesmo valor de precisão adquirido no corpus WikiNER-eng, indicando que do total classificado como PER, cerca de 13% pertencia a outro rótulo. A semelhança do comportamento do classificador CRF em ambos os corpora em inglês pode ser observada através da medida F1, próximo dos 89%.

O comportamento da identificação e classificação de entidades nomeadas, para o rótulo LOC é semelhante entre os corpora em inglês submetidos ao CRF. Este rótulo de entidade nomeada é o segundo melhor identificado e classificado. As medidas de precisão, cobertura e F1 para o rótulo LOC, em ambos os corpora variam aproximadamente entre 84% e 87%. Estes valores demonstram um aproveitamento relativamente bom para reconhecer os termos referentes a nomes de locais.

Os rótulos ORG e MISC, que fazem referência, respectivamente, a nomes de organizações e demais nomes de entidades, possuem valores inferiores de precisão e cobertura, de acordo com os resultados da Tabela 5.14. O CRF, no corpus CoNLL-03, cobre cerca de 78% das entidades de rótulo ORG, e do total classificado com este rótulo, classifica corretamente, aproximadamente 79%. Enquanto no corpus WikiNER-eng, o CRF cobre cerca de 73% das entidades de rótulo ORG, apontando uma precisão de, aproximadamente 83%. A cobertura das entidades de rótulo MISC varia em torno de 75%, enquanto a precisão, no corpus CoNLL-03, é de 77%, contra próximos 80% no corpus WikiNER-eng.

Tabela 5.15: Número de termos classificados por rótulos, utilizando o CRF e as bases CoNLL-03 e WikiNER-eng.

	CoNLL-03					WikiNER-eng				
Rótulos	PER	LOC	ORG	MISC	O	PER	LOC	ORG	MISC	O
PER	**2560**	46	104	13	50	**39025**	1182	418	1271	489
LOC	61	**1681**	91	40	46	1294	**30697**	955	1544	968
ORG	199	141	**1949**	61	141	1665	2149	**20290**	2557	1321
MISC	39	37	70	**684**	79	2288	2084	2050	**29665**	3288
O	82	62	244	89	**38077**	460	532	637	1772	**900915**

Os valores da Tabela 5.15 demonstram que, no corpus CoNLL-03, 104 dos termos cujo rótulo é PER, são classificados erroneamente com o rótulo ORG. Enquanto no corpus WikiNER-eng, o rótulo que, de forma errada, mais recebe termos referentes a nomes de pessoa é MISC, com 1271 ocorrências. Estes valores constituem falsos negativos (FN), o que reduz a cobertura para o rótulo PER. Já em relação aos falsos positivos (FP), a maior quantidade de termos classificados como PER, de

forma errada são 199 termos, os quais de fato pertencem ao rótulo ORG. Enquanto no corpus WikiNER-eng, o maior número de FP para o rótulo PER é referente a termos de rótulo MISC, no valor de 2288 termos. Tais FP comprometem a precisão do CRF para o rótulo PER.

No corpus WikiNER-eng, para os rótulos LOC, ORG e O, o maior valor, que compõe seus respectivos FN, é visto na atribuição dada ao rótulo MISC. 1544 dos termos que pertencem ao rótulo LOC, são classificados erroneamente como MISC, assim como 2557 dos termos de rótulo ORG, acabam sendo classificados com o rótulo MISC. Até mesmo entre os termos que não são entidades nomeadas 1772 são classificados como MISC.

Já no corpus CoNLL-03, essa relação dos rótulos classificados incorretamente, compondo os FN de cada rótulo, é mais distribuída. Dos termos que pertencem ao rótulo LOC, 91 são classificados como ORG. Enquanto, 199 dos termos cujo rótulo é ORG, são classificados como PER. Entre os termos de rótulo MISC, 79 são classificados como não entidade nomeada (O). Finalizando, 244 termos que não são entidades nomeadas, são classificados como ORG.

Tabela 5.16: Medidas por rótulo de entidade nomeada, utilizando o classificador CRF e as bases WikiNER-pt e HAREM.

Rótulo de	WikiNER-pt			HAREM		
Entidade Nomeada	Prec.	Cobert.	F1	Prec.	Cobert.	F1
PER	**0,8942**	0,8814	**0,8878**	0,7157	**0,7064**	**0,7110**
LOC	0,8224	**0,9337**	0,8745	**0,7211**	0,5803	0,6431
ORG	0,8361	0,7192	0,7733	0,5371	0,3481	0,4225

No que diz respeito aos testes executados com os corpora em português WikiNER-pt e HAREM (Tabela 5.16), o comportamento do CRF para o REN, na relação dos rótulos de entidades nomeadas que são melhores identificados e classificados, é semelhante ao observado com os corpora em inglês. Repetindo a ordem dos rótulos que possuem maiores valores. O Rótulo PER apresenta maior valor de F1, seguido dos rótulos LOC e ORG.

Comparando os resultados do CRF, nos corpora em português, percebe-se que os valores, por rótulo de entidade nomeada, são superiores no corpus WikiNER-pt. O rótulo PER possui medida F1 de aproximadamente 71% no corpus HAREM, contra cerca de 89% no corpus WikiNER-pt. A diferença desta medida é maior para o rótulo LOC, de próximos 64% contra 87%. Enquanto para o rótulo ORG distancia-se ainda mais, no HAREM aproxima-se de 42% contra 77% obtidos no corpus WikiNER-pt.

De acordo com os valores da Tabela 5.17, o rótulo LOC é mais afetado pelos falsos negativos, no corpus WikiNER-pt. Dado que 2613 termos de rótulo PER, 2087 termos de rótulo ORG, e até mesmo 3777 termos que não são entidades nomeadas, são classificados como LOC. Enquanto no corpus HAREM, isso ocorre para o rótulo O. De modo que 244, 115 e 122 dos termos cujo rótulos são, respectivamente, PER,

Tabela 5.17: Número de termos classificados por rótulos, utilizando o CRF e as bases WikiNER-pt e HAREM.

	WikiNER-pt				HAREM			
Rótulos	PER	LOC	ORG	O	PER	LOC	ORG	O
PER	**31042**	2613	271	1292	**705**	38	11	244
LOC	684	**39250**	369	1733	71	**318**	44	115
ORG	728	2087	**10460**	1268	34	20	94	**122**
O	2261	3777	1410	**823952**	175	57	26	**15331**

LOC e ORG não são reconhecidos como entidades nomeadas. Inclusive a cobertura para o rótulo ORG é inferior a 50%, porque dos 270 termos deste rótulos, contidos no corpus de teste derivado do HAREM, apenas 94 são classificados corretamente.

Afim de realizar uma validação em relação aos resultados por rótulos obtidos, as Tabelas 5.18 e 5.19 demonstram os valores, por rótulo de entidade nomeada, da medida F1 para os quatro corpora de teste, com os classificadores MEMM e HMM, respectivamente.

Tabela 5.18: Medidas por rótulo de entidade nomeada, utilizando o classificador MEMM e as bases CoNLL-03, WikiNER-eng, WikiNER-pt e HAREM.

Rótulo de	CoNLL-03	WikiNER-eng	WikiNER-pt	HAREM
Entidade Nomeada	F1	F1	F1	F1
PER	**0,8673**	**0,8530**	**0,8668**	**0,7175**
LOC	0,8322	0,7885	0,8506	0,6292
ORG	0,7311	0,6916	0,7306	0,4664
MISC	0,7320	0,7112		

Tabela 5.19: Medidas por rótulo de entidade nomeada, utilizando o classificador HMM e as bases CoNLL-03, WikiNER-eng, WikiNER-pt e HAREM.

Rótulo de	CoNLL-03	WikiNER-eng	WikiNER-pt	HAREM
Entidade Nomeada	F1	F1	F1	F1
PER	0,5394	**0,8006**	0,7715	**0,2974**
LOC	**0,7136**	0,7685	**0,8171**	0,2731
ORG	0,5792	0,6611	0,5101	0,2012
MISC	0,5748	0,6315		

Como já mencionado, os experimentos demonstrados pelas Tabelas 5.18 e 5.19 possuem o intuito de validar os resultados mostrados nas tabelas 5.14 e 5.16. A primeira validação observada é que mesmo para os classificadores MEEM e HMM o rótulo de entidades nomeadas, referente a nomes de pessoas é o melhor identificado e classificado, como pode ser verificado pelos valores F1 das tabelas 5.18 e 5.19. Este rótulo

só não possui o maior valor de F1 nos experimentos que utilizam o classificador HMM juntamente com os corpora CoNLL-03 e WikiNER-pt. Nestes experimentos o rótulo LOC é superior à PER, demonstrando um comportamento diferenciado em relação ao classificador HMM no reconhecimento de termos que indicam locais.

Entretanto, a acurácia do REN para um determinado rótulo de entidade nomeada pode estar relacionada com o número de ocorrência destas categorias de entidades. A Tabela 5.20 agrupa o número de termos para cada rótulo, nos corpora selecionados para treinamento. Os valores da Tabela 5.20 mostram que o rótulo PER só não ocorre com maior frequência no corpus 'wikiner-pt-train'.

Tabela 5.20: Número de Rótulos de entidade nomeada por corpora de treino.

Rótulos	conll-train	wikiner-eng-train	wikiner-pt-train	harem-train
PER	**11128**	**102172**	59207	**5262**
LOC	8286	87717	**157794**	3419
ORG	10001	61464	21520	3506
MISC	4556	90785		

Comparando os resultados dos classificadores de REN, por rótulo de entidades, com o número de ocorrência das entidades, percebe-se que o CRF atinge medida F1 superior para o rótulo PER, nos quatro corpora, mesmo quando este rótulo não ocorre em maior número, nos corpora de treino. O classificador MEMM também atinge valor superior, para a medida F1, para o rótulo PER, nos quatro corpora.

Já no que diz respeito ao classificador HMM, a acurácia do REN, por rótulo de entidade nomeada, na maioria dos casos segue a relação de maior ocorrência, em cada corpus. No corpus 'wikiner-pt-train' há mais termos de rótulo LOC, e o HMM atinge maior valor de F1 no reconhecimento desta categoria de entidade nomeada. O mesmo ocorre com o reconhecimento das entidades de rótulo PER, que possuem maior valor da medida F1, quando treinados com os corpora 'wikiner-eng-train' e 'harem-train'. A exceção a esse comportamento ocorre com o uso do corpus 'conll-train', no qual há um maior número de entidades com rótulo PER, entretanto o maior valor para a medida F1 é adquirido para o rótulo ORG.

Os experimentos realizados neste capítulo confirmam a eficácia do CRF pra o REN, comparado ao MEMM e HMM, inclusive nos corpora constituídos de textos em português. As *features* referentes ao POS *tagging*, *capitalization* e afixos demonstraram maior contribuição para auxiliar a identificação e classificação das entidades nomeadas. Quando incluídas as *features* de contexto, a acurácia dos classificadores apresenta melhoria, a qual é ligeiramente maior com os intervalos de cinco e sete termos em sequência. As características dos corpus utilizados para treinamento dos classificadores afeta o REN, por rótulo de entidade nomeada. De modo que a comparação destes classificadores deve considerar todos esses fatores para demonstrar como identificam e classificam as entidades nomeadas.

Capítulo 6

Considerações Finais

A tarefa de Reconhecimento de Entidades Nomeadas é presente em trabalhos de diferentes domínios, uma vez que permite a anotação de dados textuais não estruturados, auxiliando tarefas de processamento de linguagem natural. Com a grande quantidade de técnicas já utilizadas para identificar e classificar termos críticos em textos, este trabalho apresenta um estudo comparativo de classificadores baseados em aprendizagem de máquina, para diferentes conjuntos de *features* e corpora.

Tais experimentos revelam que, entre os três classificadores testados, os melhores resultados, por meio das medidas de precisão, cobertura e F1, são alcançados, pelo CRF. A superioridade dos valores dos testes com o CRF são validados pelo méto de validação cruzada *k-fold*, além de demonstrar o nível de acerto, com base nos rótulos de entidades nomeadas.

O estudo comparativo para diferentes conjuntos de *features*, serve para indicar, individualmente e em conjunto, a contribuição de cada característica extraída dos termos e de seu contexto no texto. Os testes individuais das *features* mostram que, além do próprio termo, características extraídas de sua estrutura interna, como prefixo, sufixo e *capitalization*, além do POS *tagging* servem para aumentar eficácia do REN, de forma significativa. Já os testes que incluem *features* de contexto, ou seja as características dos termos antecessores e sucessores, demonstram de modo geral, valores superiores para os intervalos de 5 e 7 sequência de termos.

O estudo comparativo desenvolvido neste trabalho contribui para demonstrar as diferenças, na identificação e classificação de entidades nomeadas, de classificadores baseados em aprendizagem de máquina, que estão entre os principais utilizados para o REN. Contribui também com o estudo comparativo da contribuição das *features* para melhorar os resultados do classificador CRF, com dois corpora distintos. Além de demonstrar a comparação dos classificadores em corpora nos idiomas inglês e português, o que contribui para a difusão de trabalhos envolvendo REN neste último idioma.

Dada a diversidade de trabalhos que apresentam Reconhecimento de Entidades No-

meadas, estes experimentos servem para auxiliar futuros trabalhos, que lidem com a execução desta tarefa, utilizando classificadores baseados em aprendizagem de máquina, principalmente quando utilizados textos em português. Podendo ajudar a definição de alguns fatores, como a escolha do conjunto de *features*, corpus a ser utilizado e classificador.

Pesquisas futuras, para este estudo desenvolvido, incluem a adição de mais corpora, idiomas e rótulos de entidades nomeadas. Uma vez que é crescente o desenvolvimento de trabalhos que realizam o REN em textos publicados em mídias sociais, como por exemplo o Twitter.

Há uma diversidade de domínios que utilizam o Reconhecimento de Entidades Nomeadas, como por exemplo a Bioinformática. Trabalhos futuros que envolvem mais domínios podem contribuir para um estudo comparativo mais abrangente, incluindo entidades nomeadas de rótulos como: gene, proteína, entre touros.

Outra atribuição futura para este estudo comparativo é o desenvolvimento de novas observações do uso de diferentes técnicas de REN. O estudo comparativo do uso combinado de distintos classificadores baseados em aprendizagem de máquina é um exemplo desta atribuição futura, além da utilização de técnicas de REN baseadas em abordagens híbridas.

Referências Bibliográficas

[Abacha e Zweigenbaum 2011] Abacha, A. B. e Zweigenbaum, P. (2011). Medical entity recognition: A comparison of semantic and statistical methods. In *Proceedings of BioNLP 2011 Workshop*, pp. 56–64. Association for Computational Linguistics.

[Aronson 2001] Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, pp. 17. American Medical Informatics Association.

[Atkinson e Bull 2012] Atkinson, J. e Bull, V. (2012). A multi-strategy approach to biological named entity recognition. *Expert Systems with Applications*, 39(17):12968–12974.

[Atserias et al. 2006] Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., e Padró, M. (2006). Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of Language Resources and Evaluation Conference*, volume 6, pp. 48–55.

[Bailey e Elkan 1993] Bailey, T. L. e Elkan, C. (1993). Estimating the accuracy of learned concepts.". In *Proc. International Joint Conference on Artificial Intelligence*. Citeseer.

[Bikel et al. 1997] Bikel, D. M., Miller, S., Schwartz, R., e Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pp. 194–201. Association for Computational Linguistics.

[Borthwick et al. 1998] Borthwick, A., Sterling, J., Agichtein, E., e Grishman, R. (1998). Nyu: Description of the mene named entity system as used in muc-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7*. Citeseer.

[Carpenter 2007] Carpenter, B. (2007). Lingpipe for 99.99% recall of gene mentions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume 23, pp. 307–309.

[Chiticariu et al. 2010] Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., e Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical*

Methods in Natural Language Processing, pp. 1002–1012. Association for Computational Linguistics.

[Ciaramita e Altun 2006] Ciaramita, M. e Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 594–602. Association for Computational Linguistics.

[Cohen e Sarawagi 2004] Cohen, W. W. e Sarawagi, S. (2004). Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89–98. Association for Computing Machinery.

[Cowie e Lehnert 1996] Cowie, J. e Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1):80–91.

[Cunningham et al. 2002] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Ursu, C., Dimitrov, M., Dowman, M., Aswani, N., e Roberts, I. (2002). Developing language processing components with gate (a user guide). *University of Sheffield, Sheffield UK*, 5.

[Davis e Goadrich 2006] Davis, J. e Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM.

[Doddington et al. 2004] Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., e Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. *International Conference on Language Resources and Evaluation*, 4:837â–840.

[Eddy 1996] Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, 6(3):361–365.

[Ekbal e Saha 2012] Ekbal, A. e Saha, S. (2012). Multiobjective optimization for classifier ensemble and feature selection: an application to named entity recognition. *International Journal on Document Analysis and Recognition*, 15(2):143–166.

[Fersini et al. 2014] Fersini, E., Messina, E., Felici, G., e Roth, D. (2014). Soft-constrained inference for named entity recognition. *Information Processing & Management*, 50(5):807–819.

[Grishman 1995] Grishman, R. (1995). The nyu system for muc-6 or where's the syntax? In *Proceedings of the 6th conference on Message understanding*, pp. 167–175. Association for Computational Linguistics.

[Grishman e Sundheim 1996] Grishman, R. e Sundheim, B. (1996). Message understanding conference-6: A brief history. In *COLING*, volume 96, pp. 466–471.

[Grover et al. 2010] Grover, C., Tobin, R., Alex, B., e Byrne, K. (2010). Edinburgh-ltg: Tempeval-2 system description. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 333–336. Association for Computational Linguistics.

[Hearst et al. 1998] Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., e Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

[Iskold 2006] Iskold, A. (2006). Clearforest: a top-down approach to semantic web. `http://readwrite.com/2006/12/21/clearforest`. [Online: acessado em 11-Outubro-2015].

[Jonnalagadda et al. 2013] Jonnalagadda, S., Cohen, T., Wu, S., Liu, H., e Gonzalez, G. (2013). Using empirically constructed lexical resources for named entity recognition. *Biomedical informatics insights*, 6(Suppl 1):17.

[Kang et al. 2012] Kang, N., Afzal, Z., Singh, B., Van Mulligen, E. M., e Kors, J. A. (2012). Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics*, 45(3):423–428.

[Kaur e Gupta 2010] Kaur, J. e Gupta, V. (2010). Effective approaches for extraction of keywords. *Journal of Computer Science*, 7(6):144–148.

[Kohavi et al. 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pp. 1137–1145.

[Krupka e Hausman 1998] Krupka, G. e Hausman, K. (1998). Isoquest inc.: Description of the netowl (tm) extractor system as used for muc-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pp. 1–10.

[Lafferty et al. 2001] Lafferty, J., McCallum, A., e Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289.

[Lewis et al. 2004] Lewis, D. D., Yang, Y., Rose, T. G., e Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397.

[Liberman e Cieri 1998] Liberman, M. e Cieri, C. (1998). The creation, distribution and use of linguistic data: the case of the linguistic data consortium. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pp. 159–164.

[Lin 1995] Lin, D. (1995). University of manitoba: description of the pie system used for muc-6. In *Proceedings of the 6th conference on Message understanding*, pp. 113–126. Association for Computational Linguistics.

[Ling e Weld 2012] Ling, X. e Weld, D. S. (2012). Fine-grained entity recognition. In *Proceedings of the Association for the Advancement of Artificial Intelligence*. Citeseer.

[Liu et al. 2011] Liu, X., Zhang, S., Wei, F., e Zhou, M. (2011). Recognizing named entities in tweets. *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 39:359–367.

[Liu e Zhou 2013] Liu, X. e Zhou, M. (2013). Two-stage ner for tweets with clustering. *Information Processing & Management*, 49(1):264–273.

[Manning et al. 2014] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., e McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pp. 55–60.

[Mansouri et al. 2008] Mansouri, A., Affendey, L. S., e Mamat, A. (2008). Named entity recognition approaches. *International Journal of Computer Science and Network Security*, 8(2):339–344.

[Marrero et al. 2009] Marrero, M., Sánchez-Cuadrado, S., Lara, J. M., e Andreadakis, G. (2009). Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

[McCallum et al. 1999] McCallum, A., Pereira, F., Ave, P., e Park, F. (1999). Maximum entropy markov models for information extraction and segmentation. *Proceedings of the Seventeenth International Conference on Machine Learning*, 3:591–598.

[McCallum 2002] McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`. [Online: acessado em 28-Setembro-2015].

[Nadeau 2005] Nadeau, D. (2005). Balieâbaseline information extraction: Multilingual information extraction from text with machine learning and natural language techniques. Technical report, Technical report, University of Ottawa.

[Nadeau e Sekine 2007] Nadeau, D. e Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Nadeau et al. 2006] Nadeau, D., Turney, P., e Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity.

[Nothman et al. 2013] Nothman, J., Ringland, N., Radford, W., Murphy, T., e Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

[Osenova e Kolkovska 2002] Osenova, P. e Kolkovska, S. (2002). Combining the named-entity recognition task and np chunking strategy for robust pre-processing. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September*, pp. 20–21.

[Pianta et al. 2008] Pianta, E., Girardi, C., e Zanoli, R. (2008). The textpro tool suite. In *Proceedings of Language Resources and Evaluation Conference*. Citeseer.

[Piskorski e Yangarber 2013] Piskorski, J. e Yangarber, R. (2013). Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*, pp. 23–49. Springer.

[Ratinov e Roth 2009] Ratinov, L. e Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147–155. Association for Computational Linguistics.

[Rizzo e Troncy 2012] Rizzo, G. e Troncy, R. (2012). Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 73–76. Association for Computational Linguistics.

[Rocktäschel et al. 2012] Rocktäschel, T., Weidlich, M., e Leser, U. (2012). Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

[Saha e Ekbal 2013] Saha, S. e Ekbal, A. (2013). Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, 85:15–39.

[Saha et al. 2009] Saha, S. K., Sarkar, S., e Mitra, P. (2009). Feature selection techniques for maximum entropy based biomedical named entity recognition. *Journal of biomedical informatics*, 42(5):905–911.

[Santos et al. 2006] Santos, D., Seco, N., Cardoso, N., e Vilela, R. (2006). Harem: An advanced ner evaluation contest for portuguese. In *Proceedings of Language Resources and Evaluation Conference*, pp. 1986–1991.

[Sarawagi 2008] Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3):261–377.

[Sarawagi e Cohen 2004] Sarawagi, S. e Cohen, W. W. (2004). Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, pp. 1185–1192.

[Schuemie et al. 2007] Schuemie, M. J., Jelier, R., e Kors, J. A. (2007). Peregrine: Lightweight gene name normalization by dictionary lookup. In *Proc of the Second BioCreative Challenge Evaluation Workshop*, pp. 131–133.

[Settles 2004] Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 104–107. Association for Computational Linguistics.

[Settles 2005] Settles, B. (2005). Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

[Speck e Ngomo 2014] Speck, R. e Ngomo, A.-C. N. (2014). Ensemble learning for named entity recognition. In *The Semantic Web–ISWC 2014*, pp. 519–534. Springer.

[Srihari et al. 2000] Srihari, R., Niu, C., e Li, W. (2000). A hybrid approach for named entity and sub-type tagging. In *Proceedings of the sixth conference on Applied natural language processing*, pp. 247–254. Association for Computational Linguistics.

[Sutton 2012] Sutton, C. (2012). An introduction to conditional random fields. *Foundations and TrendsÂ® in Machine Learning*, 4(4):267–373.

[Sutton et al. 2004] Sutton, C., Rohanimanesh, K., e McCallum, A. (2004). Dynamic conditional random fields. *Twentyfirst international conference on Machine learning ICML 04*, 4:99.

[Tkachenko e Simanovsky 2012] Tkachenko, M. e Simanovsky, A. (2012). Named entity recognition: Exploring features. In *Proceedings of KONVENS*, volume 2012, pp. 118–127.

[Van Zaanen et al. 2007] Van Zaanen, M., Mollá, D., et al. (2007). A named entity recogniser for question answering. Pacific Association for Computational Linguistics.

[Vavliakis et al. 2013] Vavliakis, K. N., Symeonidis, A. L., e Mitkas, P. A. (2013). Event identification in web social media through named entity recognition and topic modeling. *Data & Knowledge Engineering*, 88:1–24.

[Witten e Frank 2005] Witten, I. H. e Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wu et al. 2008] Wu, Y. C., Lee, Y. S., e Yang, J. C. (2008). Robust and efficient multiclass svm models for phrase pattern recognition. *Pattern Recognition*, 41:2874–2889.

[Zhang e Elhadad 2013] Zhang, S. e Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–1098.