PUCRS

FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

DANIELA OLIVEIRA FERREIRA DO AMARAL

RECONHECIMENTO DE ENTIDADES NOMEADAS NA ÁREA DA GEOLOGIA: BACIAS SEDIMENTARES BRASILEIRAS

Porto Alegre
2017

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

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

RECONHECIMENTO DE ENTIDADES NOMEADAS NA ÁREA DA GEOLOGIA: BACIAS SEDIMENTARES BRASILEIRAS

DANIELA OLIVEIRA FERREIRA DO AMARAL

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

Orientador: Prof. Renata Vieira

Porto Alegre
2017

Ficha Catalográfica

A485r Amaral, Daniela Oliveira Ferreira do

Reconhecimento de Entidades Nomeadas na Área da Geologia : Bacias Sedimentares Brasileiras / Daniela Oliveira Ferreira do Amaral . – 2017.

107 f.

Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Renata Vieira.

1. Reconhecimento de Entidades Nomeadas. 2. Conditional Random Fields. 3. Corpus. 4. Geologia. 5. Bacia Sedimentar Brasileira. I. Vieira, Renata. II. Título.

Student's Daniela Oliveira Ferreira do Amaral

Reconhecimento de Entidades Nomeadas na Área da Geologia: Bacias Sedimentares Brasileiras

This Thesis has been submitted in partial fulfillment of the requirements for the degree of Doctor of Computer Science, of the Graduate Program in Computer Science, School of Technology of the Pontifícia Universidade Católica do Rio Grande do Sul.

Sanctioned on September 14, 2017.

COMMITTEE MEMBERS:

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS)

Profa. Dra. Mara Abel (UFRGS)

Prof. Dr. Paulo Miguel Torres Duarte Quaresmo (Universidade de Évora)

Profa. Dra. Renata Vieira (PPGCC/PUCRS - Advisor)

DEDICATÓRIA

Dedico a minha tese a Antonio, meu amado esposo e ao nosso filho João Pedro, preciosidade em nossas vidas, por representarem a fonte de minha inspiração.

“Todas as vitórias ocultam uma abdicação.”
(Simone de Beauvoir)

AGRADECIMENTOS

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

Ao meu amado esposo Antonio, meu eterno muito obrigada pelo constante apoio, carinho, companheirismo, compreensão e amor sempre dispensados a mim durante toda a minha jornada acadêmica. Agradeço ao maior presente divino que recebi, meu filho João Pedro. És o ser mais precioso do mundo e fonte de nossa inspiração.

À maninha Fabrize, por toda a preocupação e apoio nunca negados a mim.

À minha querida orientadora, professora Renata Vieira, o meu respeitoso e profundo agradecimento pela confiança depositada em mim novamente. Nossa história acadêmica iniciou no mestrado. Renata, obrigada por todos os teus ensinamentos e amizade.

Agradeço à minha admirável amiga de todas as horas e colega, “Sandrinha” Collovini, pelos ensinamentos, palavras de incentivo, motivação e carinho.

Aos meus colegas do laboratório de Processamento da Linguagem Natual (PLN), em especial, Artur, Daniela Schmidt, Henrique, Evandro, Carolina, Vinicius, Bolivar pelo carinho, amizade e troca de conhecimentos.

A colega e amiga Renata De Paris e aos bolsistas de Iniciação Científica, Maiki e Anny, pela responsabilidade na realização de relevantes tarefas em conjunto para o desenvolvimento desse trabalho.

A uma pessoa que foi mais que uma amiga, uma irmã para mim: “Cidinha” obrigada pelos conselhos e paciência para comigo.

Agradeço a minha instituição do coração, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), por ter me acolhido nesses sete anos de experiência na Pós-graduação. Aos professores e funcionários que tive o privilégio de conhecer e trocar conhecimentos, o meu muito obrigada.

Por fim, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pela oportunidade de cursar o Doutorado.

RECONHECIMENTO DE ENTIDADES NOMEADAS NA ÁREA DA GEOLOGIA: BACIAS SEDIMENTARES BRASILEIRAS

RESUMO

O tratamento da informação textual torna-se cada vez mais relevante para muitos domínios. Nesse sentido, uma das primeira tarefas para Extração de Informações a partir de textos é o Reconhecimento de Entidades Nomeadas (REN), que consiste na identificação de referências feitas a determinadas entidades e sua classificação. REN compreende muitos domínios, entre eles os mais usuais são medicina e biologia. Um dos domínios desafiadores no reconhecimento de EN é o de Geologia, sendo essa uma área carente de recursos linguísticos computacionais.

A presente tese propõe um método para o reconhecimento de EN relevantes no domínio da Geologia, subárea Bacia Sedimentar Brasileira, em textos da língua portuguesa. Definiram-se features genéricas e geológicas para a geração do modelo de aprendizado. Entre as abordagens automáticas para classificação de EN, a mais proeminente é o modelo probabilístico Conditional Random Fields (CRF). O CRF tem sido utilizado eficazmente no processamento de textos em linguagem natural. A fim de gerar um modelo de aprendizado foi criado o GeoCorpus, um corpus de referência para REN Geológicas, anotado por especialistas. Avaliações experimentais foram realizadas com o objetivo de comparar o método proposto com outros classificadores. Destacam-se os melhores resultados para o CRF, o qual alcançou 76,78% e 54,33% em Precisão e Medida-F.

Palavras-Chave: Reconhecimento de Entidades Nomeadas, Conditional Random Fields, Corpus, Geologia, Bacia Sedimentar Brasileira.

NAMED ENTITIES RECOGNITION IN THE GEOLOGY AREA

ABSTRACT

The treatment of textual information has been increasingly relevant in many domains. One of the first tasks for extracting information from texts is the Named Entities Recognition (NER), which consists of identifying references to certain entities and finding out their classification. There are many NER domains, among them the most usual are medicine and biology. One of the challenging domains in the recognition of Named Entities (NE) is the Geology domain, which is an area lacking computational linguistic resources. This thesis proposes a method for the recognition of relevant NE in the field of Geology, specifically to the subarea of Brazilian Sedimentary Basin, in Portuguese texts. Generic and geological features were defined for the generation of a machine learning model. Among the automatic approaches to NE classification, the most prominent is the Conditional Random Fields (CRF) probabilistic model. CRF has been effectively used for word processing in natural language. To generate our model, we created GeoCorpus, a reference corpus for Geological NER, annotated by specialists. Experimental evaluations were performed to compare the proposed method with other classifiers. The best results were achieved by CRF, which shows 76,78% of Precision and 54,33% of F-Measure.

Keywords: Named Entity Recognition, Conditional Random Fields, Corpus, Geology, Brazilian Sedimentary Basin.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

REN – Reconhecimento de Entidades Nomeadas

PLN – Processamento de Linguagem Natural

CRF – Conditional Random Fields

EN – Entidades Nomeadas

SVM – Support Vector Machine

EG – Entidade Geológica

REG – Reconhecimento de Entidade Geológica

SUMÁRIO

1. INTRODUÇÃO

O Reconhecimento de Entidades Nomeadas (REN) é a primeira tarefa para a Extração de Informações (EI) a partir de textos. REN consiste na identificação e classificação de Entidades Nomeadas (EN), o qual abrange os mais variados domínios, como jornalismo, medicina, biologia e geologia [SC07]. De modo geral, as EN são expressões linguísticas que se referem a um nome próprio como uma pessoa, um local ou uma organização. O que constitui um tipo de EN é sua específica aplicação, ou seja, além das EN mais comuns citadas, há as mais especializadas como genes, proteínas [CDF14] e doenças. Como definimos trabalhar no domínio da Geologia, a EN de nosso interesse refere-se a Entidades Geológicas (EG), que consistem em termos específicos no texto, desde que esses façam parte de uma subárea geológica. A detecção bem sucedida de EN em um argumento é importante, a fim de que se possa determinar, seguramente, o tipo de relação que existe entre elas. São exemplos de EN: nomes de pessoas, organizações, locais, nomes de doenças, vírus, bactérias, rochas, bacias sedimentares e formações estratigráficas. A tarefa de REN envolve o processo de detecção de conceitos específicos com base em dados textuais, ou seja, a extração semântica de um termo. Um termo representa um conceito particular em que um autor pretende discutir [KLW15]. Seu objetivo é identificá-lo e reconhecê-lo semanticamente. Assim torna-se possível capturar informações sobre o texto e saber quais assuntos ele trata.

Efetivamente, a informação baseada em textos torna-se cada vez mais relevante sob muitos domínios [Jia12]. Ela configura aspectos determinantes na literatura científica, como em bases de dados, nos fóruns de discussão da medicina, biologia, jornalismo, geologia, entre outros. Estudos revelam que os textos não estruturados possuem um conhecimento valioso e que podem ser melhores explorados automaticamente. Entre as abordagens automáticas para identificação das EN estão as baseadas em regras, gazetteers e métodos estatísticos. Abordagem baseada em regras encontra termos que descrevem as estruturas, as quais nomeam um determinado conceito. Essa abordagem identifica com precisão os padrões conhecidos, no entanto a construção de regras manuais é dispendiosa e demorada. Além disso, as regras criadas para um conceito específico, normalmente, não podem ser aplicadas a outros conceitos. Já a abordagem com base em gazetteers implica em recursos terminológicos existentes, o qual fornece ao sistema o conhecimento lexical, identificando assim a ocorrência dos termos nos textos [MUSC+13]. Por sua vez, a abordagem estatística ou de aprendizado de máquina fixa-se sobre a distribuição da palavra ao longo do texto, associando-a a features locais e não locais.

Técnicas de aprendizado de máquina e de Processamento de Linguagem Natural (PLN) têm ganho popularidade na descoberta de dados, podendo ser úteis a profissionais dos mais variados campos, em função da grande quantidade de documentos científicos

disponíveis na literatura. O aprendizado de máquina é um método de análise de dados que automatiza a contrução de modelos analíticos; ele permite que sejam encontradas respostas automaticamente, por meio de algoritmos que interativamente aprendam com os dados, sem explícitamente preocupar-se em como obter os resultados desejados. Em outras palavras, esse método gera um modelo com a finalidade de dar respostas futuras a outros dados que não são etiquetados. O aspecto interativo do aprendizado de máquina é importante porque, à medida que os modelos são expostos a novos dados, eles são capazes de se adaptar de forma independente [Cir01]. Eles aprendem por meio de cálculos anteriores a produzirem decisões e resultados confiáveis. Devido às novas tecnologias de computação, o aprendizado de máquina, atual, não é como o aprendizado de máquina do passado. Enquanto muitos desses algoritmos têm a capacidade de aplicar automaticamente cálculos matemáticos complexos sob um grande volume de dados, a velocidade com que esses dados são processados ainda é um desenvolvimento recente.

Entre os exemplos expressivos de aplicativos baseados nessa técnica estão: 1) a recomendação on-line da Amazon e do Netflix, que são aplicações para a vida cotidiana; 2) o depoimento de clientes no Twitter, ou seja, a aprendizagem combinada com a criação de features linguísticas; 3) a detecção de fraudes em cartões bancários, um dos usos mais óbvios e importantes em nosso mundo atualmente; 4) o reconhecimento de padrões, pois através deles é possível a detecção de muitos tipos de imagens. Por exemplo, o Serviço Postal dos EUA utiliza o aprendizado de máquina para o reconhecimento da escrita, como o de uma assinatura; 5) automóveis autônomos que aprendem a dirigir em itinerários de rápido acesso. Enfim, o interesse crescente pelo aprendizado de máquina deve-se ao fato de que usar um processamento computacional de tal natureza, torna-se mais barato e mais poderoso, à medida que dispomos, progressivamente, de um grande volume de informações a serem analisadas. Assim, serão oferecidos resultados preciosos de forma mais rápida.

Nesse contexto, verificou-se que REN em domínios específicos, como Medicina, Biologia e Geologia, por meio de aprendizado de máquina, além de ser uma necessidade pela carência de ferramentas automatizadas, é um desafio e uma tarefa extremamente importante na área de PLN, principalmente em documentos do português. Dentre os domínios de pesquisa estudados para a tarefa de classificação de EN, destaca-se o de Geologia, visto que a literatura apresenta uma deficiência no estudo desse tipo de entidades, bem como nos seus relacionamentos [JPM+11][SMG10][Sob12]. Logo, a adequada identificação e classificação de EN sob domínios específicos, o qual inclui o de Geologia, representa um grande desafio aos pesquisadores de PLN. Em especial, isso deve-se ao fato de que as palavras têm diferentes aplicações nos textos, ou seja, há muitas maneiras de mencionar a mesma EN. Por exemplo, "West Bengal"e "WB" se referem a um local idêntico. Isto significa que o mesmo lugar pode ser identificado por vários nomes [SMG10]. Somada à carência de informações semânticas geológicas e à necessidade de soluções automáticas

para capturá-las, está a construção de uma base de dados de referência formada por um conjunto de textos, chamado de corpus.

Observou-se que no domínio de Geologia, existem muitas subáreas, como Estratigrafia, Sedimentologia, Petrografia, entre outras. Logo, para fins de delimitação do escopo desta tese e para a obtenção de resultados mais específicos na sua avaliação, foi definida uma subárea específica, pois a quantidade de EN no domínio de geologia como um todo é demasiado ampla. Dessa forma, o presente trabalho se insere no contexto de REN para o domínio da Geologia, subárea Bacia Sedimentar Brasileira, por meio de aprendizado de máquina.

A escolha desse domínio deve-se ao fato de que o REN Geológicas é pouco encontrado na literatura, o que consequentemente leva a uma carência de recursos automáticos para a área de Geociências [SMG10] [Sob12] [JPM+11]. Já REN para Medicina, Biomedicina e Biologia é apresentado com uma gama bem maior de trabalhos científicos na comunidade Médica e da Computação [Zac12] [AHvdH+15] [CPCT14] [AV14] [ME15] [OTK02]. Para a comunidade geológica, essa subárea é relevante e necessária de ser trabalhada por três motivos. Academicamente, as Bacias Sedimentares Brasileiras são registros de variações climáticas e tectônicas ao longo do tempo, as quais contam a história geológica dessas grandes áreas de sedimentação. Segundo, tais bacias são espaços de grande importância econômica pela questão do petróleo, água, hidrocarbonetos, recursos minerais e fósseis.

Terceiro, o maior obstáculo enfrentado pelos pesquisadores em Geociências e que os consome um tempo enorme é a busca e o agrupamento de informações geológicas sobre determinado assunto. Por exemplo: como extrair a partir da tabela Cronoestratigráfica, considerando o intervalo de tempo geológico do Devoniano (419.2 Ma) ao Oligoceno (23.03 Ma), mais especificamente, os Estágios Aptiano-Albiano (125 100.5) informações relacionadas à formação das rochas geradoras de hidrocarbonetos das bacias da margem brasileira? Essas informações poderão ser obtidas por meio de questionamentos como:

1) Qual o intervalo de tempo geológico de interesse?

2) Quais foram às condições de formação durante esses estágios?

3) Qual bacia está inserida no contexto dos dois questionamentos anteriores?

4) Outras bacias mundiais também comportam rochas geradoras desses estágios?

As informações acima existem e estão dispersas em inúmeros artigos de periódicos científicos, capítulos de livros, livros, resumos de congressos e simpósios, notas explicativas, relatórios entre outros documentos acadêmicos. Agrupar todas essas informações é o ideal. Para isso, é fundamental identificar, primeiramente, as ENs que responderão cada uma das perguntas formalizadas acima, como o tempo Geológico e a classificação de rochas a fim de, posteriormente, realizar a junção de tais informações.

1.1 Objetivos

Dentro desse contexto, o presente trabalho aplica a tarefa de REN para identificar e classificar EN em textos do Português. A escolha de realizar REN para o referido idioma, deve-se ao fato de que a língua portuguesa é carente de recursos automatizados, o que dificulta ainda mais os avanços da pesquisa para essa tarefa. Assim, será apresentado o objetivo geral e detalhados os objetivos específicos.

1.1.1 Objetivo Geral

Esta tese tem por objetivo propor e avaliar métodos para o reconhecimento de EN relevantes no domínio de Geologia (subárea Bacia Sedimentar Brasileira) em textos da língua portuguesa.

1.1.2 Objetivos Específicos

A partir do objetivo geral, os seguintes objetivos específicos foram definidos:

• Realizar um estudo das subáreas da Geologia, a fim de definir uma para a tarefa de REN;

• Estudar o problema da especialização das classes relevantes, utilizadas a partir da subárea escolhida;

• Construir um corpus de referência com a ajuda de especialistas na subárea Bacia Sedimentar Brasileira;

• Elaborar um conjunto de features para aprendizado do reconhecimento das entidades geológicas;

• Gerar modelos baseados no corpus, a partir de aprendizado de máquina, comparando o CRF com outras técnicas (árvores de decisão e Naive Bayes);

• Analisar os resultados.

1.2 Organização da Tese

A tese está organizada da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica do trabalho proposto, como o REN, suas técnicas.Trabalhos relacionados,

abordagens computacionais e ferramentas que realizam a extração e classificação de EN em domínios específicos, em especial no de Geologia, são apresentadas no Capítulo 3. A determinação das classes, construção do corpus, discussão da anotação, ou seja, o estudo do domínio é descrito no Capítulo 4. Na sequência, é detalhada a modelagem do problema, com a descrição das features e a explanação do método de validação das EN extraídas (Capítulo 5). O processo de avaliação como um todo compreende o Capítulo 6. Por fim, as considerações finais são relatadas no Capítulo 7.

2. FUNDAMENTAÇÃO TEÓRICA

A Extração de Informação (EI) consiste em extrair dados específicos a partir de um conjunto homogêneo de documentos em linguagem natural, de acordo com um domínio restrito [NNB09]. Sistemas comuns de EI caracterizam-se por três etapas: análise do texto, seleção de regras e aplicação de regras. A primeira etapa realiza desde a segmentação do texto em sentenças até uma análise linguística mais completa como a identificação de palavras-chave como nomes próprios, verbos de interação e a relação de sucessão e antecessão das palavras que serão extraídas dos textos. A segunda define as regras de extração de informação, que são associadas a "triggers", geralmente, palavras-chave. A presença de 'triggers' ativa a verificação de partes condicionais de regras de correspondência. Por exemplo, uma determinada regra pode estar associada a ocorrência de prefixos e/ou sufixos nas palavras relevantes num conjunto de textos. A terceira e última etapa exprime que ao acionar uma regra, todas as suas condições contextuais são verificadas, e um modelo é prenchido de acordo com os resultados das regras apropriadas para um determinado domínio. Dessa forma, o resultado pode ser a geração de um modelo ou de um texto anotado.

Assim, um passo fundamental para a maioria das tarefas de EI é a identificação, classificação e relacionamento entre Entidades Nomeadas (EN), ou seja nomes próprios, na sua grande maioria. Esse processo é denominado Reconhecimento de Entidades Nomeadas (REN), o qual busca a identificação de expressões linguísticas contidas em textos [BdSVR08] que possam ser associadas a determinadas classes ou categorias. Exemplos de tais expressões linguísticas, usuais em EI, são nomes de pessoas, locais, e organizações. A EI tem se mostrado útil também em áreas de conhecimento altamente especializadas, como biologia e medicina. Uma área igualmente relevante mas para o qual não se encontram muitos trabalhos em EI, é a geologia, área focal deste trabalho. Mais especificamente, neste trabalho tratamos de EI no sub-domínio de Bacias Sedimentares Brasileiras.

As primeiras soluções para REN estabeleceram padrões de regras edificados manualmente, as quais dependem de especialistas para a sua criação e possuem alto custo [HAB+97].

A literatura atual apresenta trabalhos que utilizam técnicas de aprendizado de máquina para o reconhecimento das ENs. As regras para REN são aprendidas automaticamente a partir de dados já etiquetados [Cir01].

Neste capítulo, serão apresentados os fundamentos teóricos que se relacionam com o tema proposto: REN, técnicas de aprendizado de máquina e o domínio de estudo, Bacia Sedimentar Brasileira.

2.1 Reconhecimento de Entidades Nomeadas

O REN implica em encontrar cada menção de EN no texto e etiquetá-lo com sua classe [JM09]. Outros autores [FMS+10] definem REN como a identificação e classificação de expressões linguísticas, na sua maioria nomes próprios, os quais referenciam um termo específico no texto. A classficação mais frequente apresenta as ENs nos tipos: Pessoa, como "Pedro", Local como "Brasil"e Organização como "Faculdade de Informática", incluindo expressões temporais e numéricas como datas, horas, percentuais e monetárias, respectivamente. Há, ainda, as classificações mais especiíficas, que estão de acordo com um domínio restrito como o Biomédico. Por exemplo: a classe Anatomia tem "Veia Poplítea"como EN e a classe Trombo Pet possui a EN "Embolia Pulmonar Bilateral". Já no domínio Químico, a classe Ácido Lático tem como ENs "Ácido Lático", "Ácido Lático"e "Ácido Mercapto Lático", por exemplo.

REN não é uma tarefa simples, pois implica em muitos desafios. Normalmente, o REN separa-se em dois processos: o primeiro de identificação das ENs; e o segundo de classificação das ENs. Um dos obstáculos é a delimitação das ENs na etapa de identificação. Por exemplo, “Bacia do rio Belém” pode ser identificada como uma única EN ("Bacia do rio Belém") ou como duas ENs: “Bacia do rio Belém” e "rio Belém". A etapa de classificação é ainda mais complexa que a etapa de identificação, devido à ambiguidade das palavras, ou seja, à mesma EN pode ser atribuída mais de uma classe, dependendo do contexto. No exemplo, "A Bacia de Campos limita-se ao norte pela Bacia do Espírito Santo", a EN "Bacia do Espírito Santo"é classificada como Bacia Sedimentar, e em "A Bacia de Campos se estende do Estado do Espírito Santo até Arraial do Cabo", a EN "Espírito Santo"é classificada como Local.

A literatura apresenta várias abordagens sobre a tarefa de REN. A abordagem menos complexa caracteriza-se pela consulta de gazetteers, ou seja, ela utiliza listas de nomes próprios, como nome de pessoas, locais, nomes de doenças, compostos químicos, genens e proteínas, conforme a área de aplicação. Outra abordagem é constituída por um conjunto de regras, geralmente, formada por regras intuitivas, baseadas em conhecimento e elaboradas manualmente. Técnicas caracterizadas por empregar aprendizado de máquina como as probabilísticas são outra possibilidade para a tarefa de REN. Tais técnicas empregam um corpus de treino e as ENs que compõem esse corpus são determinadas preliminarmente. Dentre o conjunto de técnicas de aprendizado de máquina dirigidas para REN citam-se os modelos estatísticos como o Maximum Entropy Markov Model (MEMM) [BON03], [CC03], [FDM+05], os modelos de Máxima Entropia [CN03], o Hidden Markov Model (HMM) [BMSW97] e os modelos matemáticos probabilísticos intitulados Conditional Random Fields (CRF) [CD12], [LMP01], [Set04], [SM09]. Finalmente, as abordagens híbridas, as quais combinam técnicas diferentes, como por exemplo, sistemas que utilizam

regras, os quais empregam gazetteers ou léxicos compostos por palavras-chave ou triggers [AL13], [MAM08], [NS07]. A partir das diferentes abordagens aplicadas para REN, muitos sistemas puderam ser avaliados por meio de competições, as quais serão descritas na próxima seção.

2.1.1 Avaliação de Sistemas de Reconhecimento de Entidades Nomeadas

A partir do ano de 1997 os sistemas de REN tiveram a oportunidade de participar de competições, por meio de Conferências de Avaliação Conjunta. Seu objetivo é enriquecer pesquisas que envolvem o estado da arte, de acordo com uma tarefa específica na área de EI [NNB09]. Tais sistemas recebem uma avaliação ao realizarem uma mesma tarefa e seus resultados geram, fundamentalmente, recursos de avaliação, os quais servirão de testes em pesquisas subsequentes [Cri08]. A seguir, serão apresentadas as principais conferências sobre a avaliação de identificação, classificação e relacionamentos entre ENs, bem como as métricas mais usuais de avaliação.

A Message Understanding Conference foi a conferência que iniciou a avaliação conjunta dos sistemas que tratam o REN. Seu objetvo foi o de promover e avaliar pesquisas em EI em textos da Língua Inglesa. Dentre as conferências que o MUC proporcionou, a sua sexta edição introduziu o REN como um quesito de avaliação independente ao de EI [SC07]. As ENs como nomes próprios, acrônimos e outras expressões linguísticas receberam uma das sete classificações: Pessoa, Local, Organização, Data, Hora, Valor Monetário e Porcentagem. Já a sua sétima edição foi marcada por inserir a tarefa de Relação de Modelo, a qual identifica o relacionamento entre duas ou mais ENs [Cri08].

Já a Automatic Content Extraction consistiu numa avaliação conduzida pelo 'National Institute of Standards and Thecnology (NIST). Sua tarefa é detectar as menções de entidades e o encadeamento delas, identificando sua correferência. No vocabulário ACE, as entidades são objetos, as menções são referências a elas, e as relações são expressas entre as entidades [DMP+04]. As classes podem ser: Pessoa, Organização, Local, Instalação e Entidades Geopolíticas, por exemplo, países ou cidades. Já as menções são nomes, expressões nominais ou pronomes. A extração de relações é difícil, uma vez que a extração bem sucedida implica na correta detecção de ambas as menções de um argumento, para determinar seguramente o tipo de relação que existe entre elas.

Uma outra conferência que merece destaque é a Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Organizado pela Linguateca, o HAREM avalia sistemas de REN específicos para corpus da língua portuguesa [San09]. Duas edições o compõem: o Primeiro HAREM, ocorrido no ano de 2004, e o Segundo HAREM, em 2008. A Coleção Dourada (CD) é um subconjunto da coleção do HAREM, sendo utilizada para tarefa de avaliação dos sistemas que tratam REN. O corpus do HAREM é considerado a

principal referência na área de PLN para o Português, e caracteriza-se por ter um conjunto de textos anotados e validados por especialistas (CD), o que facilita a avaliação dos sistemas.
O HAREM empregada a seguinte metodologia: 1) especifica as tarefas que serão avaliadas; 2) define as diretivas de etiquetagem e 3) estabelece a criação das coleções de textos. A avaliação conjunta que o HAREM realiza é feita por meio da comparação do desempenho dos sistemas de vários grupos. Esses grupos realizam a referida avaliação utilizando um conjunto de recursos em comum e uma métrica estabelecida por meio de um consenso. O evento do Segundo HAREM possui uma coleção composta por 1040 documentos, sendo que, dentro deste grupo, encontram-se 129 documentos constituintes da CD.
A segunda edição do HAREM, além de realizar uma avaliação mais justa dos sistemas, incluiu: a tarefa de reconhecer e normalizar expressões classificadas como Tempo e o reconhecimento de relações semânticas entre as ENs.

Os sistemas participantes utilizaram dez classes, as quais foram definidas pelos especialistas do HAREM. São elas: Abstração, Acontecimento, Coisa, Local, Obra, Organização, Pessoa, Tempo, Valor e Outro [Cri08]. Por fim, a finalidade da CD é avaliar os sistemas participantes, como Priberam, R3M, REMBRANDT, SEI-Geo e CaGE de modo a comparar as anotações produzidas por eles com a CD de referência, por meio de medidas de avaliação.

Todas as conferências apresentadas utilizam as seguintes métricas de avaliação: Precisão (P), Abrangência (A) e Medida-F (MF) [Cri08]. A Precisão representa a razão entre o número correto de ENs classificadas pelo sistema sob o número de ENs classificadas por ele. Já a Abrangência corresponde ao valor das entidades que o sistema classificou corretamente considerando o número de ENs marcadas pelo Corpus Gold (CG). Por fim, a combinação entre as duas métricas anteriores resultará na Medida-F, conforme apresentado nas três fórmulas respectivamente:

$$P = \frac{\textit{Total ENs corretamente classificadas pelo sistema}}{\textit{Total ENs classificadas pelo sistema}}$$

$$A = \frac{\textit{Total ENs corretamente classificadas pelo sistema}}{\textit{Total ENs classificadas pelo CG}}$$

$$MF = \frac{2 * P * A}{P + A}$$

2.2 Técnicas para Reconhecimento de Entidades Nomeadas

Grande parte do trabalho inicial para REN empregava recursos baseados em regras escritas manualmente [DPDPL97]. Abordagens mais recentes baseiam-se em técnicas

de aprendizado de máquina, como o supervisionado, os quais são muito eficazes e amplamente usadas para resolver problemas de identificação e classificação de tais expressões linguísticas. Algoritmos de aprendizado de máquina aprendem automaticamente se partirem de uma conhecimento gerado [AKP15]. Um programa de treino, por exemplo, trabalha com esses algoritmos, de modo que seja apto a criar resultados a novos dados de teste, de acordo com um conjunto de etiquetas conhecidas. Eles geram uma função que norteia a entrada em saídas desejadas, os quais são tradicionalmente, usados em problemas de classificação.

Os métodos mais comuns de aprendizado de máquina são: Supervisionado, Não supervisionado e Semi-supervisionado. Os dois primeiros são os mais amplamente adotados, cerca de 70% e 15%, respectivamente. Algoritmos de aprendizado de máquina Supervisionado são treinados com exemplos etiquetados, usados como entrada, onde a saída desejada é conhecida. Ele recebe um conjunto de dados de entrada juntamente com a saída correta correspondente e, o algoritmo aprende comparando a sua saída real com a saída correta. Assim há a possibilidade de modificar o modelo gerado. A aprendizagem supervisionada utiliza padrões para prever os valores dos rótulos nos dados adicionais não etiquetados, através dos métodos como classificação, regressão e predição. Ela é comumente usada em aplicações onde os dados históricos predizem prováveis eventos futuros. Por exemplo, ela pode antecipar quando as transações de cartão de crédito são suscetíveis de ser fraudulentas.

Já o aprendizado não supervisionado é empregado num conjunto de dados não rotulado. O sistema não é informado da resposta correta, ou seja, o algoritmo deve descobrir o que está sendo mostrado. O objetivo é explorar os dados e encontrar uma estrutura interna. A aprendizagem não supervisonada funciona bem em dados transacionais. Por exemplo, ela pode identificar clientes com caracterísiticas semelhantes, que podem ser tratados similarmente em campanhas de marketing. Ou ainda, encontrar as principais características que diferenciam um tipo de cliente do outro. As técnicas mais comuns incluem mapas auto-organizados, mapeamento do vizinho mais próximo e agrupamento K-means. Esses algortimos também são usados para segmentar textos bem como identificar "outliers"de dados.

A aprendizagem semi-supervisionada é utilizada para as mesmas aplicações da aprendizagem supervisionada. A diferença é que ela trabalha com dados rotulados e não rotulados, na etapa de treino. O que ocorre, normalmente, é o uso de uma pequena quantidade de dados rotulados com uma grande quantidade de dados não marcados, pois esses têm baixo custo e requerem menos esforço humano para adquiri-los. Em outras palavras, o custo associado à rotulagem é muito alto para permitir um processo de treinamento totalmente rotulado. Tal tipo de aprendizagem pode ser usada em métodos como classificação, regressão e previsão. Seus primeiros exemplos incluem a identificação do rosto de uma pessoa na "web cam".

Três diferentes algoritmos de aprendizado de máquina foram escolhidos para o presente estudo, pois esses são os mais comuns para a tarefa de REN e requerem a entrada de dados rotulados. São eles: CRF, Naive Bayes e C4.5. Eles representam abordagens distintas de algoritmos, a fim de mostrar qual delas é superior ao lidar com uma estrutura de dados específica. Nas seções seguintes, são apresentadas as referidas técnicas de aprendizdo de máquina, as quais serão considerados para a avaliação.

2.2.1 Conditional Random Fields

A literatura apresenta bibliografias que utilizam Conditional Random Fields (CRF) para abordar questões de PLN, com enfoque na EI de sequências estruturadas de um corpus. O CRF é oriundo dos seguintes modelos de Markov: Hidden Markov Models (HMM) e Maximum Entropy Markov Moldels (MEMM). O HMM ou estado finito automático e probabilístico é um dos métodos mais comuns para tarefas de etiquetagem e de segmentação. Ele identifica a maioria das sequências de rótulos nas palavras, dada uma sentença. Além disso, ele é uma forma de modelo generativo, que define um conjunto de distribuição probabilística p(Y,X) onde X e Y são variáveis aleatórias, que classificam uma sequência de observação de palavras e suas sequências de rótulos correspondentes, respectivamente. Sua finalidade é aumentar a probabilidade conjunta dessa sequência de observação bem como sua sequência de estado (rótulo e palavra de observação). De acordo com a propriedade de Markov, cada rótulo y depende do seu rótulo anterior, probabilidade de transição, enquanto que cada palavra de observação x depende somente do rótulo atual, probabilidade de emissão. Então, o HMM assume a independência entre palavras e seu contexto ou a independência de outras features. Esta simplificação permite uma aprendizagem rápida e uma maximização global da probabilidade conjunta ao longo de toda a observação e a sequência de rótulos. No entanto, um modelo generativo, como HMM, enumera todas as possíveis sequências de observação.

Essa é uma tarefa intratável para a maioria dos domínios. A menos que os elementos de observação sejam representados como unidades isoladas, ou seja, independente de outros elementos numa sequência de observação. Mais precisamente, o elemento de observação, em algum dado instante, só pode diretamente depender do estado, ou rótulo, naquele momento. Isto é um pressuposto necessário para um conjunto de dados um pouco simples, contudo a maioria das sequências de observação de palavras é melhor representada por várias features interagindo e pela longa distância de dependência entre os elementos de observação.

Então, a questão da representação dos dados é um problema que merece atenção, quando se rotula dados sequenciais, ou seja, um modelo que suporte inferência tratável é necessário, no entanto, um modelo que otimize a sequência de etiquetas também é

desejável, pois é esse processo que se busca na tarefa de REN. Uma maneira de satisfazer ambos os critérios é utilizar os modelos condicionais. São modelos que definem uma probabilidade condicional p(Y|x) sobre uma sequência de rótulos, dada uma sequência de observação particular, ao invés de uma distribuição conjunta sobre o rótulo e as sequências de observação.

Os modelos condicionais são usados para etiquetar uma nova sequência de observação x, selecionando a sequência de rótulo y de modo que aumente a probabilidade condicional p(y|x). A natureza condicional de tais modelos significa que nenhum esforço é desperdiçado em modelar as observações, e é livre de ter que fazer suposições de independências injustificadas sobre essas sequências. Arbitrariamente, atributos de dados de observação podem ser capturados pelo modelo, sem o modelador ter que preocupar-se sobre como esses atributos são relatados. Nos modelos condicionais não se tenta modelar a sequência de observação. Tal modelo especifica as probabilidades da sequência de rótulo dada uma sequência de observação. Isso leva a uma grande redução no número de combinações possíveis entre features de palavras de observação e seus rótulos. Portanto, as probabilidades podem depender do modo em que as features foram definidas, bem como suas interrelações e arbitrariedades, dada uma sequência de observação. O que acrescenta mais conhecimento ao modelo.

Um dos exemplos de modelos condicionais é o MEMMs, o qual baseia-se no Princípio da Entropia Máxima [CN03]. Tal princípio exprime que, a partir de uma coleção de fatos, um modelo deve ser consistente com todos os fatos, mas, por outro lado, tão uniforme quanto possível. Seu objetivo é encontrar uma distribuição condicional de uma sequência de rótulo dada uma sequência de observação que tenha a maior entropia. Os MEMMs substituem as funções de transição e observação de HMM por uma única função, o que significa que ela modela a probabilidade sobre o próximo estado, dado o estado atual e as suas observações. Isso porque a probabilidade de uma transição entre rótulos pode depender também de observações passadas e futuras, não apenas da atual.

Por outro lado, uma fraqueza do MEMM é o problema do viés do rótulo , o qual consiste nos estados ou rótulos com menos transições de saída, ou seja, aqueles que ocorrem com menos frequência no conjunto de treino. Isso quer dizer que dada uma sequência de observação, uma palavra pode ser ignorada no modelo gerado, se houver um rótulo com poucos exemplos e que possua somente uma transição de saída, pois esse rótulo não está gerando observação e sim condicionado a tal saída. Então, seu valor probabilístico irá para o estado que é visto repetidas vezes no treino.

Dessa forma, surgiu o CRF como um modo de resolver os problemas encontrados nos modelos de Markov baseados em grafos direcionados. CRF é um modelo matemático probabilístico usado para tarefas de aprendizado de máquina, que tem o objetivo de etiquetar e segmentar dados sequenciais, baseados numa abordagem condicional [DPDPL97]. Esse modelo é considerado por muitos pesquisadores como a técnica do estado da arte

na marcação de sequências de dados e tem sido usado para modelar estruturas de ordem linear, nas tarefas de linguagem natural como part-of-speech tagging, informação sintática e o reconhecimento de ENs. Sua aplicabilidade repercute em áreas como a Medicina, Bioinformática, Visão Computacional e Biomedicina. Sua introdução foi dada por John Lafferty [LMP01] que demonstrou as vantagens dessa nova estrutura sobre Hidden Markov Models (HMMs) e Maximum Entropy Markov Models (MEMMs).

CRF possui uma forma de modelo gráfico não direcionado que define uma única distribuição linear sobre sequências de rótulos, dada uma sequência de observação. O treinamento de CRF consiste na avaliação do peso, a fim de aumentar a probabilidade condicional das seqüências de rótulos dado um conjunto de dados de treino. De acordo com Lafferty [46], X é definido como sendo uma variável aleatória sobre uma sequência de dados para serem etiquetados, Y como uma variável aleatória sobre uma sequência de etiquetas correspondentes. As variáveis X e Y podem ser representadas da seguinte forma: X = (X1, X2,...,Xn) e Y = (Y1, Y2,..., Yn). Todos os Yi componentes de Y são assumidos para variar ao longo de um alfabeto Y de rótulos finitos. Pode-se assumir, neste caso, que as dependências de Y, condicionadas sobre X, formam uma cadeia. Para uma estrutura em cadeia, a probabilidade condicional de uma sequência de rótulos pode ser expressa por p(Y|X).

De acordo com o exemplo "A PUCRS oferece vários cursos de graduação em Porto Alegre", tem-se:

X = A, PUCRS, oferece, vários, cursos, de, graduação, em, Porto_Alegre

Y1 = O, ORGANIZAÇÃO, O, O, O, O, O, O, LOCAL

Y2 = O, LOCAL, O, O, O, O, O, O, LOCAL

Y3 = O, O, O, O, O, O, O, O, LOCAL

O vetor X corresponde a representação do texto. Já a tag "O"(Outside) configura as palavras que não são ENs [TKSDM03]. Os vetores Y1, Y2 e Y3 representam as etiquetas possíveis de X. A probabilidade condicional p(Y1|X) deve ser maior que p(Y2|X) e que p(Y3|X), pois, nessa sentença, PUCRS é uma organização à medida que representa uma instituição de ensino superior e "Porto Alegre"é um local. Logo, tem-se que a melhor etiquetagem de X é o vetor Yk tal que P(Yk|X) é maior do que qualquer outro Yj com P(Yj |X) dada uma distribuição condicional P(Y|X).

Em função da modelagem mencionada, pode-se empregar Linear-Chain Conditional Random Field [70], um caso particular de CRF utilizado para encontrar a distribuição P(Y|X) e a partir dela obter a melhor etiquetagem Y a partir do vetor de entrada X. As dependências de Y condicionadas sobre X formam uma cadeia linear, Assim, para as formulações de cadeia linear de CRF convencional, uma cadeia de Markov de primeira ordem e unidimensional é assumida para representar as dependências entre as variáveis de eti-

quetas previstas, enquanto nenhuma dependência temporal é imposta entre as variáveis observadas.

A vantagem primária dos modelos de CRF sobre os modelos HMM é a sua natureza condicional, pois resulta no abrandamento de pressupostos independentes, necessários para os modelos HMM, a fim de assegurar uma inferência tratável. Ao contrário do MEMM, o qual usa um modelo exponencial por estado para prever o próximo estado, baseado no estado atual, o CRF apresenta um modelo exponencial único para a probabilidade de uma sequência inteira de rótulos, dada a observação, atribuindo pesos a esse rótulos em relação a outros, baseados nas observações correspondentes. Linear-Chain Conditional Random Field foi o método utilizado nesta tese e será explicado no capítulo 5.

2.2.2 J48 Decision Tree

Uma das mais populares técnicas de aprendizado de máquina é representada pelo algoritmo J48 Decision Tree. Segundo [WKQ+08], esse classificador posiciona-se entre um dos melhores na área de EI . Disponível no Weka, o J8 é uma implementação do algoritmo baseado em árvore de decisão C4.5 [Qui92]. As árvores de decisão destacam-se na resolução de problemas de classificação envolvendo features nominais. Por exemplo:

O J48 organiza o seu processo de classificação na forma de árvore, ou seja, ele responde a uma série de questões sobre os atributos do conjunto de dados de teste e, após, cria uma estrutura hierárquica de nodos e arestas direcionadas. Então, o J48 Decision Tree não é um algorítmo probabilístico, ou seja, ele aprende a classificar por meio de uma tarefa de indução de árvore de decisão. Uma árvore de decisão possui os seguintes componentes:

1) Um nodo raiz que corresponde ao nó principal. Normalmente, o nodo raiz utiliza um atributo ou feature mais importante, entre as classes participantes do domínio;

2) Nodos internos, formados por uma aresta de entrada e duas ou mais arestas de saída;

3) Folhas ou nodos terminais, os quais são formados por uma aresta de entrada e nenhuma de saída; e

4) Arestas, que correspondem aos valores possíveis e diferentes de cada feature.

Em uma árvore de decisão, cada nodo da folha recebe uma classe. Os nodos terminais, os quais incluem a raiz e outros nodos internos, contêm condições de teste de atributos para separar registros que possuem características diferentes. Por exemplo, um determinado nodo raiz usa o atributo temperatura corporal para separar vertebrados com alta e baixa temperatura corporal [WFHP16]. Ao se constatar que todos os vertebrados de baixa temperatura corporal não são mamíferos, é criado um nó folha com a classe não mamíferos, o qual corresponde a um dos filhos do nó raiz. Por outro lado, se os vertebrados

possuem alta temperatura corporal, poderá ser utilizado um atributo subsequente, denominado data de nascimento, a fim de distinguir os mamíferos de outros animais com alta temperatura do corpo, como os pássaros.

Após a construção de uma árvore de decisão, torna-se imediata a classificação de um conjunto de dados. A partir do nó raiz, aplica-se a condição de teste ao conjunto de registros e segue-se a ramificação apropriada, com base no resultado do teste. Essa ação conduzirá a um nó folha ou a outro nó interno, estabelecendo-se uma nova condição de teste. Desse forma, será atribuída uma classe a um dos dados do registro, a qual está associada a um nó folha. Logo, o caminho de um árvore de decisão é usado para prever o rótulo de uma classe.

Para garantir o bom desenvolvimento de uma árvore de decisão, duas importantes etapas devem ser observadas. A partir dessa construção, as informações sobre o processo de aprendizado são geradas com base em features extraídas por meio dos diversos caminhos originados pela árvore. A segunda etapa chama-se classificação. Seu funcionamento é realizado através de testes nos atributos da amostra, tanto pelo nó raíz, quanto pelos nós subsequentes, caso seja necessário. O resultado dos testes é à atribuição de uma classe a um dado do registro, por meio da propagação dos valores de atributos de uma instância pelo nó raiz, até um dos nós folhas.

2.2.3 Naive Bayes

O Naive Bayes é amplamente reconhecido como um classificador probabilístico de clara interpretação, no qual é fundamentado pelo teorema Bayesiano [SLZ13]. Como os classificadores Bayesianos são baseados nos princípios estatísticos, a presença ou ausência de uma palavra em um documento textual determina o resultado da predição. Significa que, a cada termo processado, é atribuída uma probabilidade dele pertencer a uma determinada categoria. Esse valor probabilístico é calculado a partir das ocorrências do termo no conjunto de treino, o qual já possui suas classes conhecidas. Após o resultado de todas essas probabilidades, um novo documento poderá ser classificado, de acordo com a soma das probabilidades para cada classe de cada termo, que ocorre no documento.

Particularmente, o Naive Bayes assume que o valor de uma feature específica não está relacionado com a presença ou ausência de outra feature, dada a variável de classe. Ele considera que cada uma dessas features pode contribuir independentemente da sua probabilidade, sem levar em consideração a presença ou ausência de outras features. Apesar da sua concepção ingênua, o Naive Bayes funciona bastante bem em várias situações complexas do mundo real. Consequentemente, para alguns tipos de modelos probabilísticos, esse classificador pode ser treinado de forma muito eficiente em uma configuração de aprendizagem supervisionada.

Dessa forma, suponha que cada feature X_i é condicionalmente independente de qualquer outra feature X_i, dada uma classe C e p(X_i,...,X_n |C) é o produto das probabilidades de cada termo que aparece no documento. Então, tal produto pode ser estimado como:

p(X_i,...,X_n |C)= $\prod_{i=1}^{n}$ p(X_i|C)

onde a probabilidade P(X_i|C), P(X_i|C)... P(X_n|C) pode ser estimada a partir de um conjunto de treino. Através desse cálculo, pode-se obter as probabilidades posteriores da amostra pertencentes a cada classe. Então, com base em um critério posterior máximo Bayesiano, é selecionada a classe com maior probabilidade posterior como etiqueta dessa classe. Isso porque na análise Bayesiana, a classificação final é produzida combinando a informação anterior e a sua probabilidade, a fim de formar uma probabilidade posterior, usando a regra de Bayes.

Objetivamente, o Naive Bayes:

1) Verifica a palavra chave no documento de teste e a armazena num arquivo;

2) Calcula a frequência da palavra-chave no documento de teste;

3) Calcula a probabilidade de cada palava chave no documento de teste; e

4) Classifica o documento de teste dentro das várias classes a partir da base probabilística calculada na etapa de treino.

Ao considerar relevantes esses pontos-chave, o CRF foi a abordagem escolhida para gerar o método da presente tese. As técnicas apresentadas, J48 Decision Tree e Naive Bayes, serão aplicadas na avaliação experimental descritas no Capítulo 7. Dando continuidade a pesquisa, tem-se como um dos propósitos deste trabalho, o estudo de ENs geológicas bem como as suas respectivas classes. Dessa forma, a próxima seção apresentará as considerações determinadas como importantes sobre o domínio escolhido dentro do contexto de REN.

3. TRABALHOS RELACIONADOS

O presente capítulo abordará, especificamente, os trabalhos que se relacionam com nossa pesquisa. Nesse sentido, existem muitas referências que realizam a tarefa de REN por meio de várias técnicas, domínios e idiomas, como o inglês, português, chinês e indiano. As técnicas abrangem recursos de PLN tais como: abordagens linguísticas, abordagens baseadas em técnicas de Aprendizado de Máquina, sistemas híbridos e métodos linguísticos que utilizam regras escritas manualmente por linguístas. Já entre os domínios para REN pode-se citar: medicina, biomedicina, biologia, jornalismo, notícias esportivas e geologia. A literatura, porém quase não dispõe de trabalhos de REN para o domínio de geologia [Sob12] [SMG10] quando comparado com o número de referências disponíveis, para essa tarefa, nos quatro primeiros domínios citados [JD17] [HS17] [LVC+15] [AHvdH+15] [CPCT14] [DNF+05] [FDM+05] [Gab13] [JM09] [KLW15] [Cri08] [SZZ+03] [Sux08] [Zac12]. Especialmente, neste capítulo, destacam-se três trabalhos:

O trabalho de Sobhana [SMG10] realiza uma comparação do desempenho de diferentes modelo de recuperação de informação e co-ocorrência baseado na expansão de consulta e desambiguação de nomes de locais para um corpus de Geologia. Foi criado um corpus com documentos do idioma indiano, o qual recebeu a indexação e configuração adequada para que ele fosse a entrada dos sistemas. Alguns modelos de recuperação avaliados foram: TF-IDF, BM25, InL2 e PL2. Os experimentos com o corpus de Geologia resultaram em melhores resultados quando comparados com o método baseline para a tarefa de expansão de consulta com a desambiguação de ENs do tipo Local.

O artigo de Batista [BSCB10] desenvolveu uma ferramenta denominada Hendrix como o objetivo de identificar entidades geográficas em textos da língua portuguesa bem como produzir uma sumarização geográfica. Três etapas compuseram o sistema: 1) identificar, nos textos, as entidades geográficas como nomes de serras, ruas e rios por meio de CRF; 2) retirar a ambiguidade semântica de termos geográficos, excluindo assim entidades idênticas extraídas do corpus e 3) produzir um resumo geográfico através da geração de uma lista de entidades geográficas, as quais são baseadas em uma lista de conhecimento externa, ou seja, uma ontologia.

Em [Sob12], é proposto um sistema de REN usando CRF com a ajuda da etiquetagem de ENs num corpus de geologia, o IITKGPGEOCORP. Assim, um novo conjunto de etiquetas foi criado para auxiliar a anotação do corpus. A técnica de CRF contou com a ajuda de algumas features, desenvolvidas especialmente, para extrair informações geológias como: prefixo, sufixo, POS tagging, features de dígitos e informações que circundam as palavras dos textos e suas tags. Dentro desse contexto, as seções seguintes, apresentarão tais trabalhos de forma mais detalhada.

3.1 Sistemas para REN em Domínios Específicos

De acordo com a literatura encontrada em pesquisas, há uma demanda de trabalhos que realizam REN por meio de várias técnicas diferentes, disponibilizando aos usuários ferramentas e recursos para domínios especializados. O trabalho descrito em [KLW15] propõem um método eficiente e eficaz para identificar ENs a partir do PubMed para categorias semânticas biomédicas. A abordagem proposta utiliza padrões lingüísticos para selecionar sintagmas nominais candidatos beseados em palavras-chaves e uma técnica de aprendizagem de máquina para completar a tarefa de REN.

Inicialmente, identificaram-se palavras-chave como família, produto, sistema, gene e proteína. Sendo assim, uma das condições para determinar as palavras-chave foi: numa EN, um termo à direita dela tem maior probabilidade de ser a palavra que determina a natureza da entidade do que um termo à esquerda. Por exemplo: "HBx gene". Os termos escolhidos foram filtrados pelo classificador SVM (Support Vector Machine) e sua ambiguidade foi julgada por três anotadores. O próximo passo extraiu termos candidatos relacionados às palavras-chave por meio de padrões linguísticos como, por exemplo, a presença dos termos gramaticais "is", "are"ou "as". Esses termos ajudaram na determinação das ENs.

O classificador SVM foi, novamente, utilizado para a remoção de termos incorretos. Tal classificador treinou sobre um conjuntos de textos denominado SemCat e os termos candidatos foram obtidos a partir de resumos da base de documentos científicos PubMed. O SemCat [KLW15] [TXT+05] abrange um conjunto de textos que contém entidades categorizadas semanticamente para Genômica. As features usadas pelo sistema foram prefixos, sufixos e trigramas. Por fim, os resultados experimentais demonstram que o método proposto é promissor por alcançar 93% de precisão para as classes biomédicas: Gene, Proteína, Doença, Célula e Células.

Por outro lado, Akhondi [AHvdH+15] apresenta um conjunto de abordagens que combina métodos baseadas em gramática, dicionários e expressões regulares, definidas manualmente, para extrair ENs químicas. CHEMDNER [KRL+15] foi o corpus utilizado para o desenvolvimento e avaliação do sistema de REN. Esse corpus foi anotado com os seguintes tipos de entidades: "DMSO", "Iodopyridazines", "(CH3) 2 SO", "Chebi: 28262", "ácido 2-acetoxibenzóico", "aspirina", "C4-C-NPEG9" e suas respectivas categorias: Abreviação, Família, Fórmula, Identificador, Sistemático, Trivial e Outro. Dentre os onze recursos lexicais empregados, destacam-se dicionários e bases de dados como: ChEBI, dicionário formado por pequenas entidades moleculares; ChEMBL, base de dados de moléculas bioativas com propriedades farmacológicas e UMLS, uma coleção de conceitos biomédicos agrupados em 135 tipos semânticos diferentes [MBMB01]. Utilizou-se o "Oscar", pacote de software open-source para realizar o reconhecimentos das ENs químicas, o qual é composto por diferentes tipos de modelos, tal como o modelo Bayesiano e o MEMM [JAW+11]. Todas

as ferramentas foram utilizadas com sua configuração padrão, sem treinamento adicional e ajustamento. Vários experimentos foram realizados com combinações de tais recursos e, a partir deles, alcançou-se uma Medida-F de 78% sobre os dados de treino e de teste. Assim, o sistema é capaz de fornecer informações para a maioria dos compostos encontrados.

Já Danger et.al [DPMR14] se propõem a desenvolver um módulo PPIES para Interações Proteína-Proteína (PPI). Entre as suas funcionalidades está a detecção de ENs com o reconhecimento de doze classes de entidades relevantes para esse contexto. PPI são interações formadas por proteínas nas quais fazem parte da maioria das funções biológicas de qualquer ser vivo. Dessa forma, os sistemas automáticos de extração de informação PPI são uma necessidade imediata para os biólogos. O PPIES é composto por dois módulos: um dicionário com detecção de siglas padronizadas e um classificador denominado CRF. A base de dados MIMx, formada por textos de domínio Biomédico foi utilizada como treino do sistema e o JNLPBA04 [KOT+04], o corpus de teste avaliado. As classes extraídas para esse domínio são: Proteína, DNA, RNA, Tipo de Célula, Linha Celular, Organismos, Método de Detecção de Interação, Método de Identificação de Exame, Tipo de Interção, Tipo de Agente, Função Biológica e Tecido. Já as principais features empregadas foram: notação BIO, a palavra atual que está sendo analisada, a palavra anterior à palavra que está sendo analisada, prefixo e sufixo de tamanho 3 e 4, POS Tagging e tag de chunk. Com o uso das features e das bases de dados realizaram-se muitos experimentos por meio de algoritmos de Máquina de Vetores de Suporte (SVM). Três dos oito sistemas participantes da avaliação combinaram algoritmos de Aprendizado de Máquina do tipo HMM e CRF. Os outros sistemas utilizaram MEMM, HMM e CRF isoladamente. Os resultados finais originaram 77,25%, 75,04% e 76,13% para Precisão, Recall, e Medida-F, respectivamente, superando todas as soluções atualmente disponíveis na literatura para o JNLPBA04. A conquista mais notável desse trabalho é a disponibilidade de um sistema que integra harmoniosamente um módulo contendo um dicionário confiável e um classificador CRF para identificar as ENs mais importantes do tipo PPI. Assim, o PPIES é um sistema comparável e de melhor desempenho que o atual estado da arte.

Em outro estudo, Keretena et.al [KLCS15] formularam uma técnica de representação de segmento extendida para aplicações médicas denominada Reconhecimento de Entidades Nomeadas Médicas (MNER). O MNER tem a finalidade de indentificar muitas aplicações médicas, tais como: interações entre doenças, detecção dos efeitos adversos das doenças, classificação de diagnósticos, interações entre gene-gene e proteína-proteína além dos sistemas de pergunta e resposta biomédicas. O "i2b2 2010" é o corpus utilizado para esse fim, formado por 826 relatórios anotados manualmente por um especialista, o qual contém três classes: Problema, Exame e Tratamento. Inicialmente, a técnica realiza a etiquetagem das ENs médicas por meio do POS tagging para os dados de treino. Os mesmos dados (amostras) etiquetados são usados como entrada ao algoritmo de Aprendizado de Máquina. Além das features já citadas por outros trabalhos relacionados, como prefi-

xos, sufixos e indicadores ortográficos, destacam-se as seguintes novas features utilizadas neste artigo: 1) A frequência de cada palavra a qual fornece um indicador se ela é uma EN ou não no conjunto de dados. Geralmente, as ENs têm taxas de frequência mais baixa em comparação com outras palavras. 2) A feature "é Inglêsa" é usada porque muitos termos médicos possuem palavras de outros idiomas; e 3) a feature "é Stop Word" será verdadeira se a palavra é frequentemente usada no idioma Inglês. Para avaliar a efetividade da técnica proposta, oito classificadores diferentes foram utilizados nos experimentos: Naïve Bayes, CRF, Entropia Máxima, K-NN, Random Tree, C4.5, Ada-Boost e Random Forest. Os resultados experimentais mostram uma melhoria global média de sete dos oito classificadores. O classificador k-NN apresenta uma média geral de 0,18% de Medida-F ao longo de três experimentos, enquanto que o classificador C4.5 regista uma melhoria de 9,33% na sua Medida-F.

O trabalho de Zaccara [Zac12] propõem desenvolver um ambiente para anotar e classificar ENs esportivas de textos escritos em português do Brasil. O corpus utilizado foi UOLCP2011, o qual é formado por cem notícias do Campeonato Pulista de 2011 e extraídas do site portal Universo Online. O processo de anotação e classificação do corpus foi feito de forma manual e auxiliado pelas seguintes ferramentas: primeiro, a criação de um crawler, a fim de reconhecer a área onde elementos como título, data e hora de criação, estão presentes. Segundo, o desenvolvimento do WebCorpus, para que especialistas façam a anotação e classificação do corpus sem a necessidade de conhecimento técnico para o cumprimento dessa etapa.

Os testes realizados utilizaram três métodos de aprendizagem de máquina a saber: Maximização de Entropia, Índices Invertidos e ROdIME. Esse último foi criado por Zaccara para o presente trabalho e corresponde a mesclagem dos dois métodos anteriores. O conjunto de features empregado abordou características como: alfanumérica, a palavra iniciando com letra maiúscula, a palavra escrita em caixa alta, números e a posição da palavra na sentença. Já as ENs receberam uma das seguintes classes: Pessoa, Lugar, Organização, Time, Campeonato, Estádio e Torcida.

Os resultados do ROdIME foram comparados com o modelo MTodas [Car12] e se apresentaram superiores. O ROdIME mostrou-se como um classificador competitivo, preciso e com o menor número de ENs esportivas reconhecidas erroneamente. Assim, ele torna-se uma boa opção para o processo de REN no idioma português do Brasil para o domínio esportivo. A seguir, apresentaremos as abordagens utilizadas na tarefa de REN para o domínio de Geologia.

3.2 Extração de Informação Textual na Área da Geologia

A extração de informação a partir de textos, em especial, na área da Geologia é um assunto pouco explorado. Especialmente, no que se refere ao REN por meio de CRF a partir de textos da língua portuguesa numa subárea específica. Um justificativa importante nessa questão são os poucos trabalhos apresentados pela literatura [BSCB10] [SMG10] [Sux08].

Suxiang [Sux08], por exemplo, apresenta um estudo comparativo do desempenho de diferentes modelos de recuperação de informação baseados na coocorrência e na desambiguação de nomes de locais em textos geológicos. O corpus IITKGP-GEOCORP foi criado a partir de uma coleção de relatórios e artigos científicos sobre a geologia do subcontinente indiano. O corpus é composto por cerca de 500 documentos, sendo que, em cada um deles, há em torno de 10.000 palavras. Muitos desses relatórios foram submetidos à Divisão de Ciências da Terra do Departamento de Ciência e Tecnologia do Governo da Índia.

Primeiramente, é feito o REN geológicas para a classe Local por meio de CRF. Com a ajuda de especialistas, criaram-se regras de desambiguação de ENs classificadas como Local. Um classificador baseline foi preparado através dos dados anotados. Após a identificação das ENs, criou-se um grafo de coocorrência baseado no método de expansão de consulta das entidades sementes para desambiguação das entidades do tipo Local. O grafo de coocorrência representa nomes de locais geográficos, relacionamentos e resolve o problema da ambiguidade entre esse tipo de entidade. A próxima etapa foi a consulta de termos geológicos a partir do grafo de coocorrência, criado anteriormente. Essa consulta é feita por um algortimo o qual percorre o nó raiz do grafo de coocorrência, explora os nodos visinhos e armazena a ordem da trajetória de vértices visitados no grafo. Cada vértice corresponde a uma EN.

Dois experimentos foram feitos para comparar o desempenho dos seguintes modelos de recuperação de informação: TF-IDF, BM25, DFR-BM25, InL2, PL2, IFB2, BB2, InexpC2 e InexpB2. O primeiro identificou ENs geológicas ambíguas por meio de coocorrência e, o segundo extraiu as ENs geológicas ambíguas através da consulta por termos geológicos. Os resultados apresentados mostraram que a coocorrência baseada na expansão de consulta melhorou significativamente o desempenho da recuperação de informações geológicas quando comparado com o classificador baseline. Verificou-se que a média de precisão de um sistema com expansão de consulta é superior quando comparada com um sistema baseline que não usa a expansão da consulta. Os resultados experimentais do sistema com a expansão da consulta mostram uma melhoria de 8,9% quando comparado com o baseline.

Já Sobhana, Mitra e Ghosh [SMG10] apresentaram um sistema de REN para textos de Geologia usando CRF. O sistema faz uso de diferentes informações contextuais das palavras com uma variedade de features que são úteis para predizer as várias classes de ENs.

O IITKGP-GEOCORP foi o corpus desenvolvido para ser aplicado neste sistema, o qual é formado por uma coleção de relatórios e artigos científicos sobre geologia do subcontinente Indiano. Muitos desses relatórios são submetidos à Divisão de Ciências da Terra do Departamento de Ciência e Tecnologia do Governo da Índia. Cerca de duzentos documentos o compoem, sendo que, cada documento, contém em torno de 10.000 palavras, as quais representam diferentes aspectos da geologia. Os documentos geológicos apresentam descrição textual de mapas, fonemas geológicos, imagens e mapas de espaço geográfico na forma de referências espaciais, referências da terra ou do solo e informação temporal. Seu vocabulário consistiu de vários termos geológicos conhecidos, além de outros termos raros. Os títulos de alguns destes documentos são:

1) Processos e Morfologias Costeiras do Delta do Godovari;

2) Retração das Geleiras Himalaianas: Indicador de Mudanças Climáticas;

3) Metamorfismo do Anortosito Oddanchatram, Tamil Nadu, Sul da Índia;

4) Magmatismo K-T e tectônica de bacia em Rajasthan Ocidental, Índia: Resultados da tectônica extensional e não da atividade da pluma Reunion;

5) Geoterma crustal (ou Gradiente Geotérmico Crustal) no sul da Província Basáltica Deccan, Índia: O Moho é tão frio quanto os crátons adjacentes;

6) Erosão e sedimentação em Kalpakkam (N Tamil Nadu, India) provocado pelo tsunami de 26 de dezembro de 2004; e

7) Distribuição granulométrica, morfoscopia e química elementar de sedimentos em suspensão da Geleira Pindari, Kumaon Himalaia, na Índia.

O corpus foi etiquetado com ENs, as quais ajudaram a identificar entidades como localização, pessoa e organização. Além disso, esse corpus foi anotado manualmente com as seguintes classes: País, Estado, Cidade, Região, Montanha, Aldeia, Ilha, Rio, Vila, Mineral, Organização, Medidas de Escala, Ano, Pessoa, Falha, Rocha e Tempo. As features auxiliaram o CRF a determinar a classe a que pertence sua respectiva EN. As principais features estabelecidas para essa tarefa de REN foram criadas com base na combinação de possibilidades diferentes em relação à palavra que será analisada e o seu contexto, tais como:

- Prefixo e Sufixo que correspondem a uma seqüência dos primeiros ou dos últimos caracteres de uma palavra, os quais não podem ser um prefixo ou um sufixo com significado línguístico isolado. Estas features são usadas de dois modos diferentes: o prefixo ou o sufixo da palavra de tamanho fixo e das palavras envolvidas ou que fazem parte desse contexto.

- A palavra no contexto, isto é, as palavras anteriores e posteriores à palavra que está sendo analisada e

- Features Numéricas formadas por dígitos binários de acordo com a presença e/ou o número dos dígitos em um token. Tais features auxiliam no reconhecimento de diversas ENs como expressões de tempo, de data, porcentagens, etc.

Por outro lado, o trabalho de Batista [BSCB10] elaborou a ferramenta Hendrix - Entity Name Desambiguator and Recognizer for Information Extraction - cujo objetivo foi o de extrair entidades geográficas de documentos em português e produzir o seu resumo geográfico. Três etapas compuseram o processo:

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

2a) Desambiguar significados geográficos, ou seja, definir a semântica das entidades geográficas, eliminando nomes idênticos aos extraídos dos textos. Para cumprir este propósito, utilizou-se uma base de conhecimento externa, a Geo-Net-PT [CSM05];

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

Três módulos principais compõem o HENDRIX. Primeiro, um módulo baseado em CRF, implementado pela ferramenta Minorthird [Coh], para extrair nomes de entidades geográficas. O segundo módulo, denominado PAREDES, foi criado para análise e referenciação dos nomes das entidades encontradas pelo Minorthird e um terceiro, o PAGE, que faz a extração de EN em um grande corpus junto com o HENDRIX.

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

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

No período de teste, a Coleção Dourada, foi modificada de modo que mantivesse, apenas, as anotações para entidades geográficas. No entanto, os testes mostraram que,

por exemplo, quando o modelo extrai a entidade Portugal, ele a identifica sempre como uma entidade geográfica, quando este pode se referir a um termo não geográfico (como, no texto, fazer alusão ao governo de Portugal), elevando assim o número de falsos positivos.

O HENDRIX participou do evento de avaliação de sistemas de perguntas e respostas, O GikiCLEF 09, na edição de 2009. Seu objetivo é fazer a avaliação de sistemas que utilizam a Wikipedia para buscar documentos que contém a resposta a uma determinada pergunta ou uma informação necessária. Um único modelo de CRF foi desenvolvido para a participação neste evento, a fim de reconhecer além de lugares, organizações, eventos e pessoas. Compuseram a fase de treino do modelo as CD do Primeiro HAREM e do evento do Mini-HAREM. Já a fase de teste foi formada, somente, pela CD do Segundo HAREM, alcançando 64% de Precisão e 45% de Abrangência. Em relação aos resultados, percebe-se que o desempenho para a categoria LOCAL diminuiu e muitas entidades foram corretamente identificadas, porém classificadas com a categoria errada. Conforme [BSCB10], conclui-se, então, que deveria ter sido treinado um modelo independente para cada uma das categorias e, consequentemente, originado resultados mais satisfatórios.

Dando continuidade a essa tese, o próximo capítulo apresentará o domínio proposto, a subárea definida, bem como as Entidades Geológicas e suas classes.

4. ESTUDO DO DOMÍNIO

Dentro dos domínios de pesquisa estudados para a tarefa de REN, destaca-se o de Geologia, visto que há uma carência de trabalhos que envolvam esses tipos de EN, além da falta de ferramentas automatizadas, que realizem a extração de tais informações, principalmente para textos do português. Adicionalmente, a adequada identificação e classificação de EN sob o domínio da Geologia, representa um grande desafio aos pesquisadores de PLN. Além da deficiência de informações semânticas geológicas e a necessidade de capturá-las automaticamente, ocorre a falta de uma base de dados. Desta forma, para a realização do objetivo da tese, é fundamental a construção de um conjunto de textos de referência, ou seja, um corpus.

Especificamente, no domínio de Geologia, verificou-se que ele é formado por várias subáreas, como Sedimentologia, Geocronologia, Petrologia, Geologia Sedimentar e Estratigrafia. Por isso, foi essencial delimitar uma subárea, devido à grande quantidade de EG no domínio em questão, bem como a obtenção de uma avaliação mais especializada na tarefa de REN.

O presente capítulo tem o objetivo de contextualizar esses assuntos da seguinte forma: apresentar a subárea Bacia Sedimentar Brasileira, os tipos de bacias que comportam essa subárea, suas origens e evoluções na história geológica da Plataforma Sul-brasileira, as suas portadoras de hidrocarbonetos, assim como sua importância econômica. As demais seções descreverão as classes dentro da subárea definida, construção do corpus, processo de anotação, os guias que auxiliaram os especialistas na classificação dos textos, o IdENGeo, a métrica de confiabilidade e a discussão da anotação.

4.1 Bacias Sedimentares Brasileiras

Segundo uma concepção geográfica, as bacias sedimentares são definidas como áreas de grande extensão (pelo menos 10.000 Km^2), caracterizada pelo acúmulo espesso de sedimentos e rochas sedimentares, por um grande período de tempo geológico [Mar05]. Esse acúmulo denso deve-se à combinação de sedimentação e subsidência da crosta, a qual é induzida, parcialmente, por sedimentos e principalmente por mecanismos tectônicos, como queda de uma estrutura ou superfície [GJ13].

Por outro lado, numa concepção geodinâmica, uma bacia sedimentar consiste na ação de mecanismos de subsidência interligados, de acordo com o mesmo sistema tectônico. Este, por sua vez, é responsável pelo reconhecimento no arcabouço estratigráfico e na descrição do seu regime tectônico [Mar16]. O arcabouço estratigráfico compreende o estudo da origem e da sucessão das camadas de rochas de uma região, no tempo e no es-

paço [VWPM+88]. A Bacia de Campos [GM90] e a Bacia do Espinhaço [Mar00] estão entre os exemplos de bacias sedimentares brasileiras que adotam uma concepção geodinâmica.

Conforme [MRB+07], as bacias sedimentares brasileiras estão dispostas de acordo com o seu contexto tectônico e com sua afinidade geológica, ou seja, o preenchimento sedimentar magmático. Resumidamente, elas estão organizadas nos seguintes grandes grupos: Sinéclises Paleozóicas, Bacias Meso-Cenozóicas da Margem Distensiva, Bacias Meso-Cenozóicas da Margem Transformante, Riftes Mesozóicos Abortados e Bacia de Antepaís Andino (de acordo com classificação proposta pela Petrobras, 2007). As Sinéclises Paleozóicas correspondem a estruturas geológicas, dobras, falhas, fraturas, que ocorreram na Plataforma Continental Sul-Americana, com amplitude regional de milhares de Km^2. Elas são geralmente densas e com camadas sedimentares, formadas ao longo de vários períodos geológicos.

As bacias Meso-Cenozóicas da Margem Distensiva incluem aquelas situadas do nordeste da margem brasileira, bacia Potiguar [MMDR04], até o limite ao sul da bacia de Pelotas. Entre outras características, pode-se destacar a acomodação sedimentar na fase Rifte, ou seja, a existência de rocha sedimentar em fissuras da superfície terrestre. De acordo com [AP73], a margem leste brasileira desenvolveu-se em quatro grande estágios: pré-fifte, rifte, marinho restrito e marinho aberto. O pré-fifte ocorre na bacia Sergipe-Alagoas e é representado por sedimentos, geralmente, avermelhados, deposicionados por meio fluvial, eólico e lacustre.

O estágio rifte é formado por natureza diversa ao longo da margem leste. Ele comporta uma seção de folhelhos, relevantes como rochas geradoras de hidrocarbonetos, na Bacia de Campos. Os riftes possuem também conglomerados, arenitos e coquinas. Já o estágio marinho restrito caracteriza-se, entre outros, pela presença de evaporitos, o qual ocorre nas bacias de Santos, Campos e do Espírito Santo. Por fim, o marinho aberto consiste de depósito de carbonatos e de sais encontrados nas bacias de Pelotas, bacia de Santos, entre outras.

Diferentemente das bacias descritas no parágrafo anterior, as bacias Meso-Cenozóicas de Margem Transformante possuem um mecanismo, responsável pela ruptura de rochas e de movimentos de subsidência. Tal mecanismo ocorre em vários lugares e são formados pelo mais alto grau de deposição em uma bacia geológica.

As bacias sedimentares que pertencem ao grupo Riftes Mesozóicos Abortados possuem amplitude temporal e preenchimento sedimentar bastante variável. Algumas delas incluem um estágio pré-rifte e um pós-rifte em seu registro geológico. A bacia do Tucutu, por exemplo, possui preenchimento sedimentar no intervalo temporal Neojurássico-Eocretáceo e é composta por basalto eojurássico.

Por fim, as Bacias de Antepaís Andino são aquelas que possuem depósitos sedimentares a partir do Eopaleozóico. A única representante no Brasil é a bacia do Acre,

a qual possui uma característica marcante na Era Meso-Cenozóica, ou seja, uma grande quantidade de sedimentos terrígenos alojados a ela.

Os grupos aqui apresentados relataram características relevantes das bacias sedimentares. Sua importância econômica é considerável, pois, a partir delas, surgem combustíveis fósseis como carvão mineral, folhelhos oleígenos ou betuminosos, gás natural e petróleo. Através do processo exploratório nas bacias sedimentares, pode-se identificar que algumas rochas sedimentares presentes nessas bacias, são consideradas reservatórios de petróleo e de gás. Por exemplo: bacia de Campos, bacia do Solimões, bacia Potiguar e bacia Sergipe-Alagoas.

Dentre os combustíveis fósseis existentes no planeta, destaca-se o petróleo, por este ser o recurso natural mais abundante no mundo. O petróleo é responsável por prover indústrias, automóveis, além de outras várias funcionalidades, como fornecer energia elétrica. Logo, trata-se de um dos óleos minerais mais importantes, derivado da matéria orgânica decomposta e convertida, primeiramente, em querogênio. O querogênio consiste na porção insolúvel da matéria orgânica, que é alterada por ações geológicas, causadas pela água e pelo vento, por exemplo[1] [2].

Além dos recursos orgânicos, as bacias sedimentares são compostas por rochas inorgânicas, que dão origem a rochas como arenito e calcário. Estas rochas são bastante empregadas na construção civil. Assim, ao analisar todos esses aspectos que envolvem o contexto das bacias sedimentares brasileiras, observa-se que estudá-las é fundamental na pesquisa de recursos orgânicos utilizados como fontes de energia. Especialmente, na avaliação do potencial exploratório de petróleo.

4.2 Determinação das Entidades Geológicas e suas Classes

Após a escolha da subárea, realizou-se um estudo das EG e de suas respectivas classes, através de leituras em livros e outros documentos científicos, com o objetivo de conhecer as EG que envolvem o assunto Bacia Sedimentar Brasileira. Paralelamente, sob a orientação de geólogos e professores da referida subárea, estabeleceram-se treze classes geológicas: Eon, Era, Período, Época, Idade, Rocha Sedimentar Orgânica, Rocha Sedimentar Siliciclástica, Rocha Sedimentar Carbonática, Rocha Sedimentar Química, Bacia Sedimentar, Unidade Litoestratigráfica, Contexto Geológico de Bacia e Outro. As classes estão descritas sequencialmente e organizadas em grupos, seguidas de exemplos, conforme [CFGF13] [HK99] [OOG16]

[1] O querogênio é constituído com base em lipídios, proteínas e carboidratos dos seres vivos. A partir desses, ele se transforma em petróleo, gás natural ou grafite.

[2] https://www.spec2000.net/11-vshtoc.htm

- **Tempo Geológico**

1. Eon: Maior subdivisão de tempo dentro da Escala de Tempo Geológico, representadas por Hadeano, Arqueano ou Arcaico (termo usado em Portugal), Proterozoico e Fanerozoico. Ex.: Litologicamente, é representado por rochas graníticas e gnáissicas, com núcleos granulíticos e charnoquíticos, arqueanos a proterozoicos.

2. Era: Corresponde a subdivisão de Eon. São Eras: Cenozoico, Mesozoico, Paleozoico. Obs.: Para os Eons Arqueno e Proterozoico, há subdivisões denominadas Eras (Eoarqueano, Paleoarqueano, Mesoarqueano e Neoarqueano) e Paleoproterozoico, Mesoproterozoico e Neoproterozoico, de acordo com a International Chronoestratigraphy Chart 2016 [OOG16].

Ex.: Este complexo de rochas vulcânicas de maior densidade modificou a dinâmica deposicional dos sedimentos Cenozoicos.

3. Período: É a subdivisão de uma Era. São eles: Quaternário, Neogênico, Paleogênico, Cretácico (Cretáceo), Jurássico, Triássico, Permiano, Carbonífero, Devoniano, Siluriano, Ordoviciano, Mississipiano e Pensilvaniano, esses dois últimos, apenas para a América do Norte. Ex.: Em torno de 180 Ma (Jurássico): diques e derrames de composição toleítica.

4. Época: Subdivisão do Período na Escala do Tempo Geológico. Alguns exemplos: Holocênico (Holoceno), Pleistocênico (Pleistoceno), Pliocênico (Plioceno), Miocênico (Mioceno), Oligocênico (Oligoceno), Eocênico (Eoceno), Paleocênico (Paleoceno), Cretácico (Cretáceo) Superior, Cretácico (Cretáceo) Inferior, Jurássico Superior, Jurássico Médio, Jurássico Inferior, entre outros. Ex.: Durante o Oligoceno, a deformação é pequena quando comparada aos outros períodos de deformação.

5. Idade: Subdivisão de Época. Alguns exemplos: Pleistocênico (Pleistoceno) Superior, Pleistocênico (Pleistoceno) Médio, Calabriano, Gelasiano, entre outros. Ex.: maior incidência entre 80 Ma1 e 90 Ma (Santoniano/Turoniano): – predominam intrusões de composição básica a intermediária.

- **Rochas Sedimentares**

6. Rocha Sedimentar Siliciclástica: Origina-se de fragmentos de rochas ígneas, metamórficas ou sedimentares, transportados e depositados para, posteriormente, formar uma rocha sedimentar Siliciclástica. Alguns exemplos: arenito, argilito, siltito, conglomerado, folhelho, diamictito, varvito, etc. Ex. Os arenitos da Formação Juruá são constituídos por minerais provenientes de rochas-fonte situadas ao Norte da Bacia do Solimões, transportados por um sistema de paleodrenagens pleistocênica.

7. Rocha Sedimentar Carbonática: Formada, predominantemente, por carbonato de cálcio e/ou por fragmentos de organismos (bioclastos), bem como pela interação entre o metabolismo de microorganismos e as partículas sedimentares presentes no ambiente deposicional. Alguns exemplos: calcário, dolomito, etc. Ex.: O calcário é cinza claro e apresenta proporções variáveis de fragmentos detríticos que podem chegar a 40 % da rocha.

8. Rocha Sedimentar Química: Formada por precipitados químicos: sais, carbonatos ou sulfatos. Por exemplo: evaporitos, fosforitos, Ironstones. Ex.: Na região da Fazenda Ressaca ocorrem fosforitos associados à porção superior desta formação.

9. Rocha Sedimentar Orgânica: Origina-se dos restos de fragmentos dos organismos vivos, a qual está relacionada à preservação de matéria orgânica. Exemplo: carvão, etc. Ex.: Apenas recentemente ocorreu alguma recuperação, com a elevação dos preços e o maior consumo de Carvão no complexo termoelétrico de Tubarão-SC.

- **Outras Classes**

10. Bacia Sedimentar Brasileira: São grandes áreas de sedimentação, ou seja, deposição de sedimentos (agregados de matéria orgânica e/ou mineral), formada por rochas sedimentares e, eventualmente, por rochas magmáticas. Sua formação foi a partir do Paleozóico. São elas: Bacia do São Francisco, Bacia do Espírito Santo, Bacia de Campos, Bacia do Paraná, entre outras. Ex.: Guerra (1989) estudou a influência da sobrecarga do Banco Vulcânico de Abrolhos sobre a estruturação halocinética da Bacia do Espírito Santo.

11. Contexto Geológico de Bacia: É a classificação relacionada aos eventos geológicos (espacial e temporal), ou seja, são os estágios relacionados à Tectônica, Sedimentação e Magmatismo. Ex.: Intracratônica ou Sinéclise, Rifte, Drifte e Margem Passiva. Ex.: Sequência Rifte, constituída unicamente pela Formação Abaiara, de idade neocomiana, formada por sucessão de arenitos descontínuos lateralmente intercalados em folhelhos calcíferos de coloração variegada.

12. Unidade Litoestratigráfica: compreende três componentes estratigráficos de acordo com o Código de Nomenclatura Estratigráfica de 1986, [PCA+86]: Formação, Grupo e Membro. A Formação Estratigráfica consiste na unidade principal da litoestratigrafia. É constituída por um corpo rochoso, o qual possui relativa homogeneidade litológica. Em geral, sua forma é tabular, com continuidade lateral e, representado graficamente, na superfície terrestre ou em subsuperfície. Uma formação pode conter um ou mais tipos de rochas. Ainda podem constituir uma formação elementos complementares, como estruturas sedimentares e fósseis. Exemplos: Formação Irati, For-

mação Abrolhos, Formação Lagoa Feia, Formação Coqueiros, Formação Pendência, etc. Ex.: A Formação Abrolhos de idade cenozoica é caracterizada por uma associação litológica complexa que engloba rochas (vulcânicas) básicas. Já o segundo componente é constituído por duas ou mais formações contíguas associadas que tenham, em comum, propriedades litológicas distintas e diagnósticas. Alguns exemplos: Javari, Tapajós, Curuá, entre outros. O terceiro componente, Membro Estratigráfico, representa a subdivisão litológica de uma formação. Consiste-se de uma entidade que possui características litológicas próprias, as quais permitem diferenciá-las das partes adjacente da formação. Ex.: Arari, Fazendinha, Ururiá, etc.

13. Outro: São as EG que não se enquadram em nenhuma das classes anteriores e pertencem à subárea Bacia Sedimentar Brasileira. Ela é uma classe utilizada apenas para os casos em que o especialista achar muito relevante anotá-la, pois o foco está nas classes definidas anteriormente.

A escolha dessas classes deve-se ao fato de que elas expressam importantes relações entre as EG de acordo com a subárea definida. Por exemplo, na sentença “A bacia do Rio do Peixe tem como substrato rochas sedimentares cretáceas dos grupos Bauru e Cuiá e localizadas ocorrências de basaltos da Formação Serra Geral", a EG “bacia do Rio do Peixe" trata-se de uma bacia sedimentar, pois sua estratigrafia compreende os grupos Bauru e Caiuá, os quais, obrigatoriamente, contêm a Formação Serra Geral, além de membros sedimentares. Assim, após a definição das classes, foi elaborada a construção do corpus, que consiste na seção que será detalhada a seguir.

5. CONSTRUÇÃO DO CORPUS

Inicialmente, fez-se a leitura de vários trabalhos científicos para a identificação de EG relacionadas à subárea Bacia Sedimentar Brasileira. Após, selecionou-se semimanualmente, um conjunto de textos para o domínio de Geologia. Esses textos são formados por teses, dissertações, artigos e boletins de Geociências da Petrobras no idioma português do Brasil. As EG pesquisadas foram: termos geológicos de acordo com a tabela Cronoestratigráfia [UGS], nomes de rochas sedimentares [HK99], nomes de bacias sedimentares brasileiras [Mar05] [BSVG03], os estágios relacionados à Tectônica, Sedimentação e Magmatismo e unidades estratigráficas. Dentre os serviços 'on-line' utilizados para a formação do corpus estão bibliotecas digitais, como Portal de Periódicos da Capes, Scielo, ACM Digital Library, IEEE Xplore, além do Google Scholar.

Obedeceram-se três critérios para a construção do corpus: relevância, sincronicidade e homogeneidade. O primeiro critério teve o cuidado de coletar textos teoricamente importantes dentro da subárea definida e respeitando o domínio estabelecido. Já o segundo estabeleceu um ciclo de tempo definido para a seleção dos textos, o que ocorreu num período de seis meses, para essa tese. Por fim, a homogeneidade foi estabelecida, principalmente, para não misturar textos com outros elementos, como imagens, tabelas e gráficos.

Nesse sentido e com o objetivo de gerar um corpus de leitura e avaliação, foram retirados, semiautomaticamente todos os 'abstracts', figuras, legendas, tabelas, gráficos, fórmulas e referências bibliográficas. No caso de teses e dissertações, excluíram-se também sumários, apêndices e anexos para que fique um conjunto de dados formado apenas pelo texto propriamente dito. Após a eliminação de todos os referidos elementos e para garantir a qualidade do corpus proposto, realizou-se uma revisão manual texto à texto.

Com a seleção de cinquenta textos, passou-se para a etapa de geração do corpus de referência, ou seja, o conjunto de textos que será comparado com as EG geradas pelos classificadores. Para isso, contou-se com o trabalho de nove geólogos, entre eles professores, doutorandos e alunos de graduação do curso de Geologia da UNISINOS do 6º semestre. Esses últimos anotadores já possuem conhecimento na subárea Bacia Sedimentar. Os textos receberam a etiquetagem morfológica POS tagging por meio da ferramenta OpenNLP e foram segmentados em sentenças, a fim de tornar mais eficiente o seu processamento, ao ser aplicada a técnica de aprendizado de máquina para o REN. Dessa forma, construiu-se o GeoCorpus, o qual funcionará como entrada nos classificadores de EG na forma de vetor. Dando continuidade à parte prática da tese, será apresentado o processo de anotação.

5.1 Processo de Anotação

Realizada a seleção e limpeza dos textos, partiu-se para o processo de anotação das EG presentes no GeoCorpus. O objetivo desta seção é descrever os recursos aplicados no GeoCorpus, que o tornaram apto de ser processado nos classificadores. Quatro etapas descrevem o processo de anotação: a primeira corresponde ao Guideline, ou seja, às orientações transmitidas aos especialistas em geologia para a realização das EG. A anotação seguiu os critérios de [Cri08], que consiste nas etapas de REN: identificação e classificação das EN.

A segunda consistiu da construção de uma ferramenta de anotação e classificação das EG, a fim de facilitar o trabalho dos anotadores. Já a terceira etapa correspondeu ao cálculo do coeficiente Kappa para avaliar a concordância da anotação realizada anteriormente. Finalmente, a discussão da anotação encerra esse processo como um todo, através das análises de casos ambíguos de EG classificadas. Esses casos foram passíveis de discussão, pois são anotações interessantes, tanto a nível morfológico, quanto à nível semântico.

5.1.1 Guideline

Tradicionalmente definido como guia de anotação, o Guideline é um documento que contém diretrizes. Especificamente para o presente trabalho, as diretrizes correspondem ao processo de anotação das EG em textos da Língua Portuguesa, no qual referem-se, exclusivamente, à Bacias Sedimentares Brasileiras. Seu objetivo é orientar os especialistas em Geologia na tarefa de classificação dos textos. Para uma melhor organização do documento, o Guideline foi dividido em quatro seções: Introdução, REN, Processo de anotação e IdENGeo.

A Introdução explica o significado do Guideline, sua finalidade dentro do contexto e como ele está divido. A seção de REN é formada pela sua definição e exemplos, com o intuito de situar o geólogo na área de Processamento de Linguagem Natural, dentro do âmbito da Ciência da Computação. Os desafios e complexidades encontrados no REN e as restrições, as quais devem ser seguidas durante a tarefa de classificação. Tais restrições determinaram que, devido a escolha do domínio da Geologia para o REN, a EN de nosso interesse refere-se a Entidades Geológicas (EG), que consistem em termos específicos no texto, desde que esses façam parte de uma subárea geológica. Então, para fins de delimitação do escopo deste trabalho, a subárea definida é Bacia Sedimentar Brasileira, pois a quantidade de EG como um todo, no referido domínio, é demasiado ampla e para a obtenção de resultados mais específicos na sua avaliação.

No Processo de anotação, explica-se em que consiste a tarefa de classificação, as características dos textos, as classes geológicas que abrangem a subárea definida e os passos que devem ser realizados para a anotação dos textos. Dessa forma, eles são constituídos pelas duas etapas de REN: identificação e classificação das EG. Numa fase inicial, estabeleceram-se algumas EG que os anotadores teriam que identificar bem como clássificá-las, descritas na Seção 4.2. A identificação consistiu em delimitar as EG, ou seja, onde ela inicia e termina. Isso vai depender muito do contexto em que ela está inserida no texto e também da interpretação dos anotadores. Paralelamente, a classificação teve por objeto determinar a classe semântica correspondente àquela EG identificada.

A segunda etapa merece destaque, pela complexidade devido à ambiguidade das palavras. Significa que uma mesma EG pode ser classificada com mais de uma classe ao se considerar o contexto e o domínio que ela está inserida. Por exemplo, na sentença: “A bacia do Paraná faz parte do estado do Paraná" , a primeira EG é classificada como Bacia Sedimentar e a segunda como Lugar. Cronologicamente, o processo de anotação envolveu os seguintes passos:

1º) Selecionar os termos que referem-se a uma EG no texto; 2º) Atribuir uma classe a EG;

3º) Verificar a delimitação da EG, ou seja, uma ou mais palavras que formam uma EG já marcada no texto e corrigi-la caso necessário; e

4º) Averiguar a classificação de uma EG identificada e modificá-la se for preciso.

Estabeleceram-se os dois últimos passos, porque o GeoCorpus foi inserido num modelo de classificação, o qual desenvolveu-se para um experimento inicial [AV14]. Tal modelo possui várias classes de Geologia e não se restringiu a uma subárea específica. O Guideline completo encontra-se no Apêndice A. A seguir, será apresentado o recurso desenvolvido para facilitar o processo de anotação.

5.1.2 IdENGeo

O IdENGeo é uma ferramenta de marcação de Entidades Nomeadas do domínio de Geologia, que objetiva auxiliar os anotadores na identificação e na classificação das EG, tornando a tarefa de anotação o mais intuitiva e simplificada possível. Os arquivos de texto que receberão a anotação devem estar no formato xml, do contrário a ferramenta não os reconhecerá. Essa ferramenta possui uma interface gráfica que permite ao usuário a visualização e a edição/adição de informações relevantes para a tarefa de anotação. Dentre as funcionalidades do IdENGeo temos:

- **Área de edição**: painel em que o usuário visualiza o texto de entrada a ser marcado com as EG;

- **Menu de filtros**: menu de funções de filtros que servem para facilitar a visualização das EG classificadas no texto. Esse menu é constituído pelos botões “Desmarcar Tudo", “Marcar Tudo" e a lista de botões com as 13 classes geológicas ilustradas em cores diferentes. A aplicação dos filtros possibilita: a visualização de todas as EG já classificadas no texto (botão “Marcar Tudo") e a visualização das EG por classe (botão “Desmarcar Tudo" seguido dos botões correspondentes a uma ou mais classes de interesse); e

- **Grupos de ações**: quatro grupos de ações localizados abaixo da área de edição que compreendem as seguintes funções: 1) Novo texto: função de seleção do novo texto a ser anotado e identificação do seu anotador; 2) Atualizar texto: função de seleção de um texto com a anotação ainda não concluída e, assim, dar continuidade a mesma; 3) Marcação de texto: função de habilitar o menu de classificação das EG; 4) Salvar texto: função de salvar o texto anotado.

A Figura 5.1 ilustra a interface gráfica do IdENGeo. Nela, um texto inicial foi carregado na Área de edição, bem como as EG já marcadas nas cores correspondentes à cada classe geológica do menu filtro. Além disso, o anotador pode iniciar um nova anotação das EG ou ainda continuar a marcação de um texto ainda não finalizado através dos Grupos de ações. Nesse contexto, o anotador realizará a classificação das EG seguindo o processo de anotação descrito na Seção 5.1. Para realizar a marcação de uma EG ainda não classificada deve-se selecionar o trecho do texto que expressa a EG e clicar no botão “Adicionar Marcação". Após, deve-se selecionar a referida classe da EG a partir do menu de classes, seguido do botão “OK". Cabe salientar que, o menu de classificação seguiu a organização de classes por grupos, conforme apresentado na Seção 4.2. Caso o anotador necessite remover a classe escolhida, deve utilizar o botão “Remover Marcação".

Finalizada a classificação, são gerados arquivos xml, que apresentam a forma de anotação das EG nos textos. Tais arquivos servirão de entrada para os classificadores de REN.

5.1.3 Anotação do IdENGeo

Conforme descrito na seção anterior, a saída do IdENGeo é formada por arquivos de texto no formato xml. Cada EG identificada é delimitada pelas tags que representam

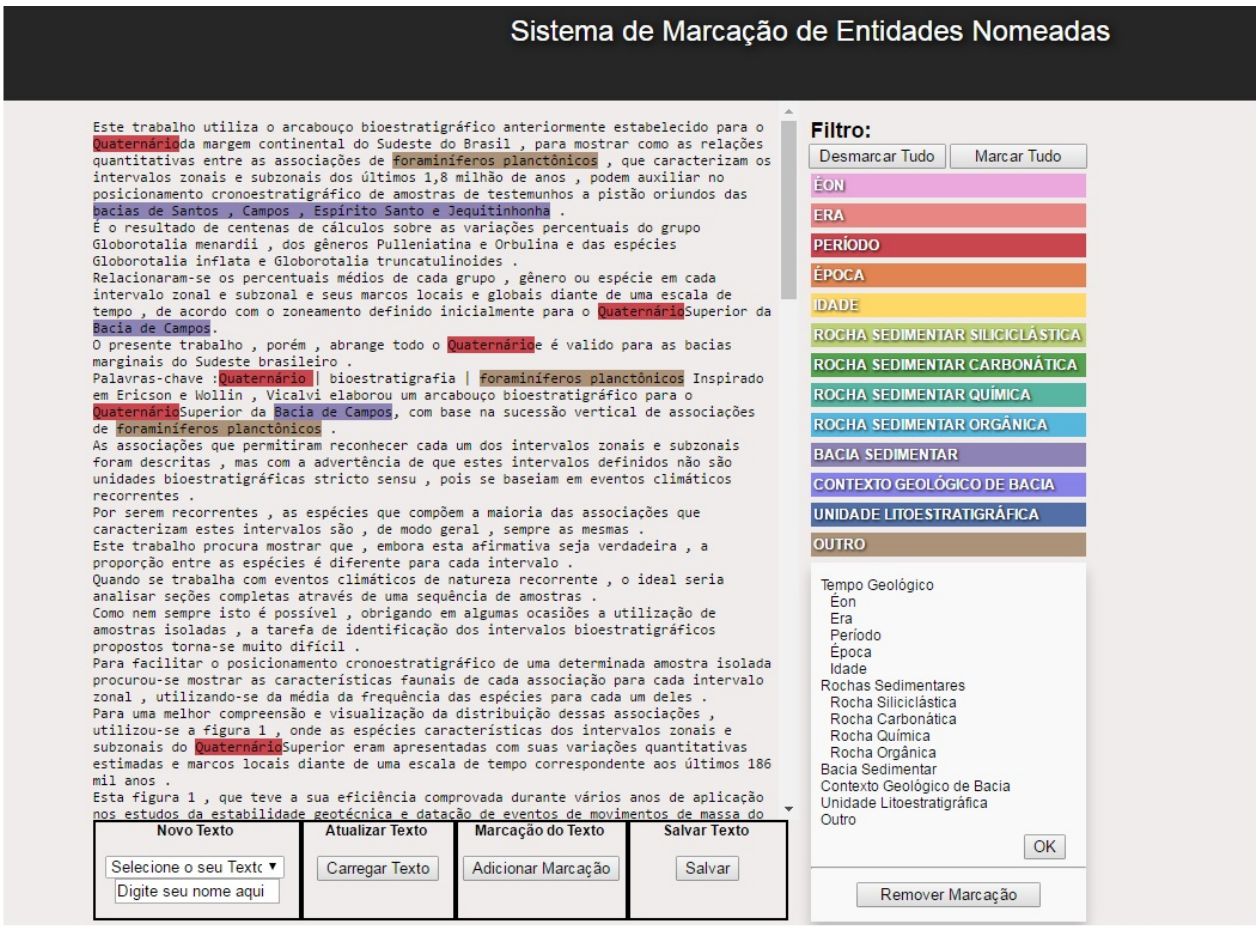


Figura 5.1 – Classificação das Entidades Geológicas no IdENGeo

início e fim de uma Entidade Nomeada, ou seja: <EG> e </EG>. A classe das EG estão contidas na tag que marca o início dessas EG. Por exemplo: <EG CATEG="BaciaSedimentar"> Bacia do Paraná </EG>. A Figura 5.2 ilustra o formato dos arquivos de anotações gerado pelo IdENGeo.

```
<doc>
A caracterização das zonas de cisalhamento na porção meridional do Cinturão Ribeira, PR, bem como seu reconhecimento em
profundidade, constitui um grande desafio, pois a falta de informação de subsuperfície dificulta a avaliação destas
estruturas crustais <EG CATEG="Era"> neoproterozoicas </EG>.
A influência destas feições na instalação e evolução da <EG CATEG="BaciaSedimentar"> Bacia do Paraná </EG>, durante o
<EG CATEG="Eon"> Fanerozoico </EG>, é amplamente aceita pela comunidade geocientífica.
O presente trabalho aplica métodos de realce de anomalias (análise qualitativa) e de estimativas de profundidade de fontes
(análise semiquantitativa), aos dados aeromagnéticos sobre a <EG CATEG="Outro"> Zona de Cisalhamento Lancinha (ZCL) </EG> no
Estado do Paraná, visando verificar seu arranjo espacial em subsuperfície.
</doc>
```

Figura 5.2 – Formato dos arquivos de saída do IdENGeo.

5.1.4 Coeficiente Kappa

O coeficiente Kappa é uma medida estatística de conformidade para análise do discurso [Coh68]. É um teste apropriado para verificar o grau de concordância que ocorre

entre anotadores ao atribuírem um rótulo a um ítem, dado um conjunto de classes. O cálculo do coeficiente Kappa (K) leva em consideração a possibilidade de concordância entre as anotações. K depende do número de anotadores, do número de ítens que serão classificados e do número de classes que poderão ser atribuídas aos ítens [VG05].

O Kappa desta tese foi realizado com base na classificação das EG e envolveu a anotação de três especialistas. Fizeram parte do cálculo do Kappa, somente os tokens identificados como EG pelos três geólogos. Dois textos do corpus foram utilizados para a obtenção do valor de K, o que constitui 122 EG. Para fins de demonstração do cálculo do coeficiente K, utilizou-se a Tabela 5.1.4, formada por dez EG, que foram selecionadas a partir dos dois textos considerados.

Tabela 5.1.4 - Cálculo do coeficiente Kappa.

Classes / Entidades Geológicas	Éon	Era	Período	Época	Idade	Rocha Sedimentar Siliciclástica	Rocha Sedimentar Carbonática	Rocha Sedimentar Química	Rocha Sedimentar Orgânica	Bacia Sedimentar	Contexto Geológico de Bacia	Unidade Litoestratigráfica	Outro	S
Neoproterozoicas	1	2	0	0	0	0	0	0	0	0	0	0	0	0,33
Bacia do Paraná	0	0	0	0	0	0	0	0	0	3	0	0	0	1,00
Fanerozoico	3	0	0	0	0	0	0	0	0	0	0	0	0	1,00
diamictitos	0	0	0	0	0	3	0	0	0	0	0	0	0	1,00
Formação Salitre	0	0	0	0	0	0	0	0	0	0	0	3	0	1,00
Grupo Una	0	0	0	0	0	0	0	0	0	0	0	3	0	1,00
Formação Serra Geral	0	0	0	0	0	0	0	0	0	0	0	3	0	1,00
cisalhamento	0	0	0	0	0	0	0	0	0	0	3	0	0	1,00
Bacia do Irecê	0	0	0	0	0	0	0	0	0	3	0	0	0	1,00
Ediacarano	0	0	3	0	0	0	0	0	0	0	0	0	0	1,00
N=10	4	2	3	0	0	6	0	0	0	6	3	9	0	Z=9,33

Dessa forma, o coeficiente Kappa é definido como:

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

Onde P(A) é a proporção de vezes que os anotadores concordaram e P(E) é o valor estimado de concordância entre eles. Quando há uma concordância entre todos os anotadores, K é igual a 1. Caso não haja condordância, K=0. Ao analisar o conteúdo de um texto, de modo que se possa resolver o problema da confiabilidade, K > 0,8 é adotado para determinar boa confiabilidade [Kri12]. Já 0,68 <= K < 0,8 gera uma confiabilidade passível de análise para cada caso anotado.

De acordo com as fórmulas seguintes, considera-se:

S_i significa a concordância para a descrição i;

m é o número de classes;

C é o número de anotadores;

N é o número de itens a serem classificados;

NC é o número total de anotações e

Z é a soma dos valores de S.

Logo:

$$S_i = \frac{1}{C(C-1)} * \sum_{j=1}^{m} n_{ij}(n_{ij} - 1)$$

$$S_1 = \frac{1}{3(2)} * [1(0) + 2(1) + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0]$$

$$S_1 = \frac{1}{6} * 2 = 0,33$$

$$S_2 = \frac{1}{3(2)} * [0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 3(2) + 0 + 0 + 0]$$

$$S_2 = \frac{1}{6} * 6 = 1$$

$$Z = \sum_{i=1}^{N} S_i$$

$$P(A) = \frac{Z}{N} = \frac{9,33}{10} = 0,933$$

$$NC = N * C = 10 * 3 = 30$$

$$P(E) = \sum (\frac{Ci}{NC})^2$$

$$P(E) = (\frac{C1}{NC})^2 + (\frac{C2}{NC})^2 + (\frac{C3}{NC})^2 + (\frac{C4}{NC})^2 + (\frac{C5}{NC})^2 + (\frac{C6}{NC})^2 + (\frac{C7}{NC})^2 + (\frac{C8}{NC})^2 + (\frac{C9}{NC})^2 + (\frac{C10}{NC})^2 + (\frac{C11}{NC})^2 + (\frac{C12}{NC})^2 + (\frac{C13}{NC})^2$$

$$P(E) = (\frac{4}{30})^2 + (\frac{2}{30})^2 + (\frac{3}{30})^2 + (\frac{0}{30})^2 + (\frac{0}{30})^2 + (\frac{3}{30})^2 + (\frac{0}{30})^2 + (\frac{0}{30})^2 + (\frac{0}{30})^2 + (\frac{6}{30})^2 + (\frac{3}{30})^2 + (\frac{9}{30})^2 + (\frac{0}{30})^2$$

$$P(E) = 0,182$$

Por fim:

$$k = \frac{P(A) - P(E)}{1 - P(E)}$$

$$k = \frac{0,933 - 0,212}{1 - 0,212}$$

$$k = 0,91$$

Assim, o Kappa resultante para os exemplos da Tabela 5.1.4 foi de 0,91 de concordância. Ao considerarmos todas as EG compreendidas nos dois textos, o coeficiente Kappa resultou em 0,88 de concordância, o que significa uma interpretação “quase perfeita” (Tabela 5.1).

Tabela 5.1 – Interpretação do Kappa

Kappa	**Concordância**
<0	Insignificante
0.01 - 0.20	Fraca
0.21 - 0.40	Razoável
0.41 - 0.60	Moderada
0.61 - 0.80	Substancial
0.81 - 0.99	Quase perfeita

5.1.5 Resultado da Anotação

Com o término das anotações dos textos que compõem o GeoCorpus, computou-se o valor de EG marcadas por classe bem como o total de EG anotadas. Dessa forma, o GeoCorpus é formado por 19 textos, 6.116 sentenças, 163.790 palavras e 5.532 EG (Tabela 5.2), distribuídas nas seguintes classes, ilustradas na Tabela 5.3:

Tabela 5.2 – Descrição do corpus de Geologia

	Número de Sentenças	Número de Palavras	Número de EG
GeoCorpus	6.116	163.790	5.532

Tabela 5.3 – Valores das Entidades Geológicas anotadas por classe.

Classes	**Entidades Geológicas**
Eon	288
Era	326
Período	637
Época	650
Idade	796
Rocha Sedimentar Siliciclástica	743
Rocha Sedimentar Carbonática	240
Rocha Sedimentar Química	5
Rocha Sedimentar Orgância	22
Bacia Sedimentar Brasileira	243
Contexto Geológico de Bacia	262
Unidade Litoestratigráfica	581
Outro	739
Total	**5532**

5.1.6 Discussão da Anotação

A aplicação das duas etapas de REN pelos especialistas implicou em anotações bem interessantes para a geração do modelo de classificação. O objetivo desta seção é descrever casos relevantes no corpus que se consolidou, os quais foram inteligidos como importantes, tanto para o processamento da linguagem, como para o REN. Analisaram-se muitas EG identificadas, suas delimitações e classificações, tendo especial importância a interpretação de cada geólogo.

Inicialmente, constatou-se que as classes Idade, Rocha Sedimentar Siliciclástica, Época, Período e Unidade Litoestratigráfica foram as mais frequentes no corpus. Isso porque a maioria dos textos abordou EG pertencentes a essas classes, envolvendo, entre outras caracterísiticas, aspectos sedimentológicos. Destacou-se o número considerável de EG da classe Unidade Litoestratigráfica, pois ela está intimamente relacionada à classe Bacia Sedimentar Brasileira, a qual norteia a subárea em estudo. Observou-se também que a classe Outro apresentou muitos casos, que os especialistas julgaram relevantes para a subárea Bacia Sedimentar Brasileira. Como por exemplo as EG “organismos fósseis", “ostracodes" e “conchostráceos". Já as classes Rocha Sedimentar Química e Rocha Sedimentar Orgânica tiveram pouca ocorrência nos textos que compõem o corpus.

Em geral, os anotadores observaram que os pontos de dificuldade da tarefa foram a delimitação das palavras que formam uma EG e a sua ambiguidade. Por exemplo, no trecho da sentença: “a respeito do rico acervo paleontológico das formações Brejo Santo, Crato e Romualdo", o anotador identificou três EG (“Brejo Santo", “Crato" e “Romualdo") para a classe Unidade Litoestratigráfica e não incluiu a palavra “formações", a qual se refere às três EG. Em contrapartida, no trecho “Sugere-se que os sedimentos da Formação Brejo Santo teriam sido depositados", o anotador identificou a EG “Formação Brejo Santo" incluindo a palavra “Formação".

Outra questão relatada pelos anotadores refere-se ao aspecto morfológico das palavras dispostas nos textos, ou seja, à forma em que o termo geológico está inserido na sentença. Significa que, quando uma expressão é constituída por um substantivo seguida de um adjetivo, este último não configura uma EG, pois ele caracteriza um substantivo. Por exemplo: “Assim, no Espinhaço Meridional os sedimentos paleoproterozóicos têm expressão reduzida, predominando os mesoproterozoicos", o anotador não classificou “paleoproterozóicos" como Era, porque essa palavra exerce a função de adjetivo e não de uma EG.

Nesse contexto, evidencia-se que “a tarefa de REN exige uma clarificação das bases semânticas e pragmáticas do processamento de linguagem natural, e que estas não são necessariamente consensuais ou explícitas. Por isso, é imprescindível a definição do conceito de EG para a sua delimitação e operacionalização no processo de anotação”Logo,

no que se refere às discussões sobre as anotações das EG, constatou-se que as suas classificações estavam extremamente relacionadas com a interpretação de cada anotador.

6. MODELAGEM DO MÉTODO

O presente capítulo descreve o método proposto para o reconhecimento de EG em textos da língua portuguesa, dentro da subárea Bacia Sedimentar Brasileira. Condicional Random Fields (CRF) foi a técnica de Aprendizado de Máquina empregada para gerar o modelo de classificação. A partir desse modelo, é possível extrair informações relacionadas à referida subárea. Dentre essas informações, destacam-se: tempo geológico, tipos de rochas sedimentares, bacias sedimentares brasileiras, eventos espaciais e temporais bem como componentes litoestratigráficos. A seção seguinte apresentará o CRF de modo genérico. As demais seções descrevem, especificamente a estrutura do método aplicada neste pesquisa, desde o seu pré-processamento até a conclusão da etapa de teste, a elaboração das features, os vetores de entrada inseridos no método, o modelo obtido e a validação das EG extraídas a partir do modelo gerado.

6.1 Arquitetura do Método

Genericamente, o método baseado em CRF caracteriza-se por ser uma técnica de aprendizado supervisionado, o que consequentemente gera duas etapas: treino e teste. Uma característica importante durante o desenvolvimento dessas etapas é o tipo de amostragem que será aplicado no processo. Significa que o corpus poderá ser dividido em uma proporção exata, uma parte para treino e outra para teste, ou o conjunto de textos será separado randomicamente para as duas etapas. As Figuras 6.1 e 6.2 elucidam a arquitetura do método para a criação e aplicação do modelo de CRF, respectivamente.

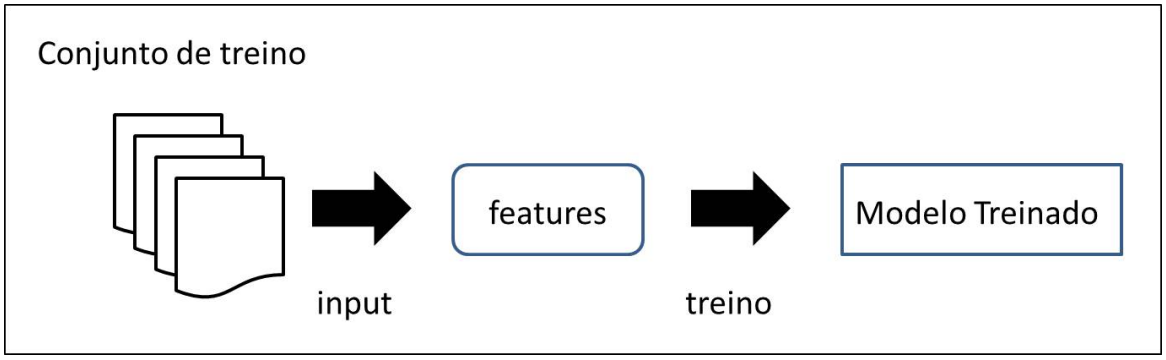


Figura 6.1 – Descrição do CRF na etapa de treino.

Inicialmente, a etapa de treino recebe como entrada um corpus de referência pré-processado, ou seja, as EG anotadas por especialistas e um ajuste na formatação dos textos, para que haja compatibilidade com o sistema. Após o pré-processamento das sentenças, um vetor com o corpus de referência é dado como entrada ao CRF, juntamente com

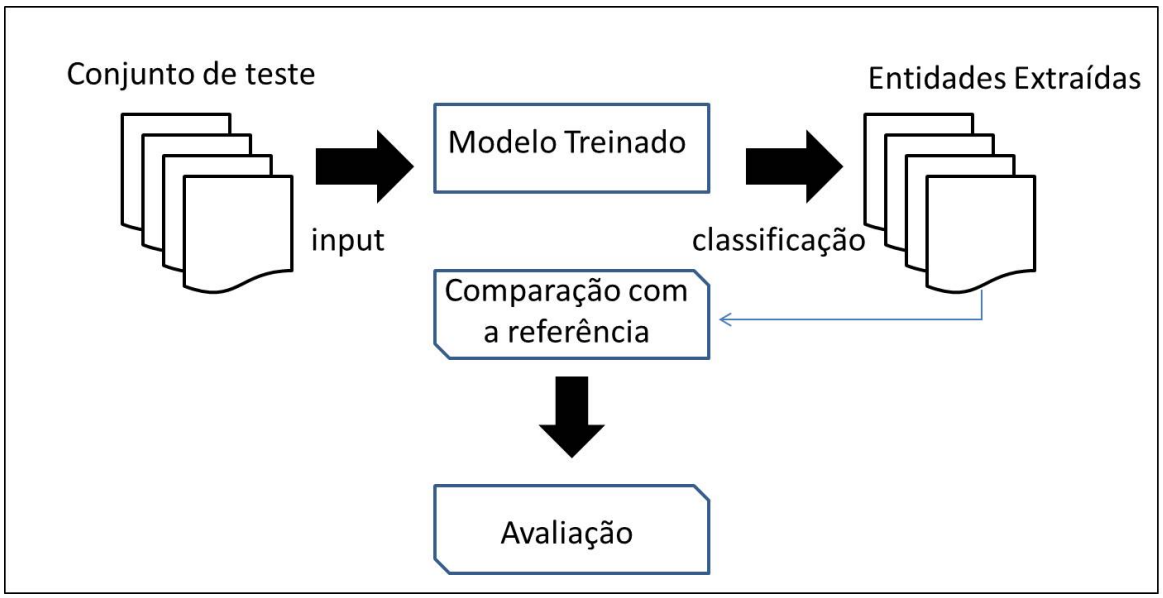


Figura 6.2 – Descrição do CRF na etapa de teste.

um conjunto de features. As features possuem duas finalidades: primeiro, caracterizar palavra por palavra no corpus e segundo, orientar o CRF na identificação e classificação correta das EG. A seguir, é gerado o modelo treinado, que corresponde a uma matriz de pesos. Tal matriz é formada por valores probabilísticos, os quais são atribuídos às features a partir das palavras que compõem os textos. Os valores formadores da matriz compreendem o intervalo entre zero e um, uma vez que esses correspondem a pesos probabilísticos.

Já, na etapa de teste, o sistema recebe como entrada um corpus diferente do aplicado na etapa anterior e o mesmo conjunto de features da etapa de treino. Em seguida, o modelo de CRF gerado é executado no corpus de teste e é avaliado o seu desempenho. Finalmente, a avaliação compara as EG marcadas no corpus de referência com as EG extraídas pelo classificador. As próximas seções descrevem a aplicação do CRF especificamente para esta pesquisa.

6.2 Geração dos Vetores de Entrada

Inicialmente, a geração dos vetores de entrada implicou no pré-processamento do GeoCorpus, o qual ocorreu da seguinte forma: findada a anotação das EG pelos especialistas, os textos em xml foram executados num programa de conversão. Seguem as etapas que o descrevem:

1º) Correção da saída do IdENGeo, ou seja, o conversor percorre todas as marcações de final de EG (</EG>) e remove os espaços que vem após ela. Tal ação é realizada para que o arquivo de entrada esteja no mesmo formato de saída do CRF. Assim, será possível avaliar o classificador na etapa de teste;

2º) Mudança no formato dos arquivos de textos de xml para txt, a fim de executá-los no OpenNLP. A biblioteca OpenNLP foi utilizada por dois motivos: primeiro, segmentar os textos em sentenças, a fim de que o CRF os processe com eficiência. Segundo, etiquetar as palavras por meio de Part-of-Speech (POS) tagging [Sch94]. O POS Tagger consiste numa marcação morfológica das palavras, a fim de auxiliar o classificador no aprendizado das EG.

3º) Aplicação da nova versão do GeoCorpus no OpenNLP;

4º) Correção dos caracteres ilegíveis inseridos por essa biblioteca. Por exemplo: como esse corpus possui uma marcação que delimita o início e fim das EG, na forma de, <EG> Bacia do Paraná </EG>, o OpenNLP o transforma em caracteres sem significado. Então o conversor identifica o início e fim de EG e corrigi essas inconsistências.

Por fim, o conversor gera uma pasta com a data e hora da execução, no formato YYYY-MM-DD. O resultado são arquivos com marcações de EG e POStags, os quais são usados tanto no treino, quanto no teste do CRF.

Concluído o pré-processamento, identificaram-se as EG. O processo de identificação foi realizado com o auxílio da anotação “IO" [dABV15] o qual denota a tag “I" (“Inside") a todas as palavras identificadas como EG. Já a tag “O" (“Outside") configura as palavras que não são EG.

Dessa forma, dois vetores são utilizados na etapa de treino. O primeiro vetor (Figura 6.3) é constituído de:

- etiquetagem POS tagging a cada palavra do texto,

- notação IO e

- categorização das EG no corpus de referência. Essas categorias foram definidas no capítulo anterior.

O segundo vetor (Figura 6.4) é formado por um conjunto de features, o qual será apresentado na próxima seção.

Já na etapa de teste (Figura 6.5), o vetor de entrada é formado pelo:

- vetor de POS tagging nas palavras que compõem os textos e

- o mesmo vetor de features aplicado na etapa de treino.

6.3 Construção das Features

A fim de auxiliar na identificação das EG, é gerado um vetor de features. As features têm o objetivo de caracterizar todas as palavras do corpus, bem como direcionar o CRF na classificação correta das EG. O vetor de features foi desenvolvido a partir do

Palavras	POS tagging	Tag "O" ou Classe
Campo	próprio	
de	preposição	OUTRO
Pampo	próprio	
localizado	verbo	"O"
na	preposição	"O"
Bacia	próprio	
de	preposição	BACIA SEDIMENTAR
Campos	próprio	
,	pontuação	"O"
tem	verbo	"O"
como	preposição	"O"
principal	adjetivo	"O"
reservatório	nome	"O"
fácies	nome	"O"
de	preposição	"O"
calcarenitos	nome	ROCHA SEDIMENTAR CARBONÁTICA
e	conjunção	"O"
calcirruditos	nome	ROCHA SEDIMENTAR CARBONÁTICA

Figura 6.3 – Primeiro Vetor da Etapa de Treino

estudo da subárea Bacia Sedimentar Brasileira[1] [MRB+07] [Fol80], [HK99] juntamente com os trabalhos relacionados à técnica de aprendizado de máquina [SZZ+03] [ZD15] [LPN17] [HS17].

Por uma questão de organização, o conjunto de features foi separado em Features não Geológicas e Features Geológias. As Features Não Geológicas são features gerais, ou seja, podem ser aplicadas em qualquer domínio. Sua formação é caracterizada por: POS tagger, a própria palavra, a presença de letras maiúsculas e/ou minúsculas, símbolos, números ou letras nas palavras e baseada na sequência frasal. Já as Features Geológicas constituem-se de características específicas para o domínio Bacia Sedimentar Brasileira. Elas foram determinadas em conjunto com especialistas e nomeadas de: Prefixo, Sufixo e Gazetteer. Os prefixos utilizados foram: Plio, Ceno, Neo, Meso, Eo, Pre, Protero, Pleisto, Plisto, Paleo, Oligo, Fanero, Holo, Fito e Mio. Já os sufixos compreendem: ceno, geno, áceo, ífero, ico, oico, iano, rico, nico, ário, perma, permas, ácea, áceas, ários, ita, itas e ito. O Gazetteer é formado por nomes geológicos que compõem a Tabela Cronoestratigráfica, os quais correspondem às EG pertencentes ao grupo tempo geológico. A Tabela 6.1 descreve o conjunto de features e o Apêndice B apresenta um exemplo do vetor de features.

[1] Tabela Cronoestratigráfica Internacional, 2016.

Palavras	Tag	PrevW	Suf
Campo	próprio	Null	False
de	preposição	Campo	False
Pampo	próprio	de	False
localizado	verbo	Pampo	False
na	preposição	localizado	False
Bacia	próprio	na	False
de	preposição	Bacia	False
Campos	próprio	de	False
,	pontuação	Campos	False
tem	verbo	,	False
como	preposição	tem	False
principal	adjetivo	como	False
reservatório	nome	principal	False
fácies	nome	as	False
de	preposição	fáceis	False
calcarenitos	nome	de	True
e	conjunção	calcarenitos	False
calcirruditos	nome	e	True

Figura 6.4 – Segundo Vetor da Etapa de Treino

6.4 Modelo Extraído

A partir dos dois vetores de entrada utilizados na etapa de treino, é gerado o modelo probabilístico CRF. De forma simplificada, o primeiro vetor contém: POS tagging, notação IO e treze categorias. O segundo, compreende um conjunto de vinte features. Evidentemente, o vetor de features é fator determinante para a derivação de um bom modelo, pois o emprego de features eficazes, dará maior probabilidade ao sistema de classificar as EG com maior precisão e abrangência. O resultado do modelo consiste numa matriz de pesos com valores probabilísticos a cada feauture.

A partir da matriz é possível avaliar o modelo na etapa de teste. Significa que o mesmo vetor de features do treino aplicado a um novo conjunto de textos valida o modelo de classificação. Assim, o CRF torna-se apto a identificar e classificar corretamente as palavras candidatas a EG em textos não etiquetados. A sentença-exemplo: “Na Bacia do São Francisco, observa-se a Formação Jequitaí", elucida a saída do classificador a partir do modelo gerado, ilustrado na Figura 6.6.

Palavras	POS tagging	PrevW	Suf
Na	artigo	Null	False
Bacia	próprio	Na	False
do	preposição	Bacia	False
São	próprio	do	False
Francisco	próprio	São	False
,	pontuação	Francisco	False
observa	verbo	,	False
-	pontuação	observa	False
se	pronome	-	False
a	artigo	se	False
Formação	próprio	a	False
Jequitaí	próprio	Formação	False
.	pontuação	Jequitaí	False

Figura 6.5 – Vetor na Etapa de Teste

6.5 Validação Cruzada

O GeoCorpus foi submetido a um método de validação cruzada denominado Cross Validation. Tal método consiste em validar o conjunto de dados, através do cálculo da média de várias estimativas de iterações. Significa que os dados validados são legítimos, eficazes e correspondem a partes diferentes do corpus. Na prática, essa técnica corresponde a um número randômico, o qual depende da divisão dos dados de entrada em “y" “folds" ou iterações. A partir dessas, y-1 partes são empregadas para a etapa de treino de um preditor, o qual é testado nos subconjuntos restantes ou dados de teste, para a validação do modelo.

O divisão do “data set" pode ser realizada de várias formas. Três delas são as mais empregadas: “Holdout" , “K-fold" e “Leave-one-out". A primeira, “Holdout" tem o objetivo de particionar o conjunto total de dados em dois subconjuntos: um para treinamento e outro para teste. O subconjunto de treino realiza o cálculo dos parâmetros, em contrapartida o de teste apresentará a legitimidade dos dados. A proporção comum para se dividir o subconjunto de dados é dois terços para treino e um terço para teste. No entanto, esse “data set" pode ser separado em proporções diferentes dessas. Realizado o particionamento, o modelo é gerado, o conjunto de teste é aplicado e o erro de predição será estimado. “Holdout" é utilizado para uma grande quantidade de dados, pois com um “data set" pequeno, o erro estimado poderá resultar em muita variação, na predição.

Tabela 6.1 – Conjunto de Features

Nome das Features	**Descrição das Features**
1) Word	A palavra em exercício (p), ignorando letras maiúsculas e minúsculas.
2) Tag	A marcação POS Tagging da palavra (p), indicando sua classificação morfológica. Exemplo: substantivo, adjetivo, verbo, advérbio, etc.
3) Cap	A ocorrência de letras maiúsculas e minúsculas da palavra (p), ou seja, “min"quando houver somente minúsculas, “max"na existência de apenas maiúsculas, ou “maxmin"quando ocorrer minúsculas e maiúsculas.
4) Ini	Indica se a primeira letra da palavra (p) for maiúscula ou minúscula.
5) Simb	Denota se a palavra (p) possui símbolos, números ou somente letras.
6) PrevW	A palavra anterior (p-1), sem considerar letras maiúsculas e minúsculas.
7) PrevT	A marcação POS Tagging da palavra (p-1).
8) PrevCap	A ocorrência de letras maiúsculas e minúsculas da palavra (p-1).
9) Prev2W	A palavra que antecede a anterior (p-2), ignorando letras maiúsculas e minúsculas.
10) Prev2T	A marcação POS Tagging da palavra (p-2).
11) Prev2Cap	A ocorrência de letras maiúsculas e minúsculas da palavra (p-2).
12) NextW	A palavra posterior (p+1), sem considerar letras maiúsculas e minúsculas.
13) NextT	A marcação POS Tagging da palavra (p+1).
14) NextCap	A ocorrência de letras maiúsculas e minúsculas da palavra (p+1).
15) Next2W	A palavra posterior (p+2), sem considerar letras maiúsculas e minúsculas.
16) Next2T	A marcação POS Tagging da palavra (p+2).
17) Next2Cap	A ocorrência de letras maiúsculas e minúsculas da palavra (p+2).
18) Pref	A presença de prefixo nas palavras do texto. Exemplos: Plio, Ceno, Holo, etc.
19) Suf	A ocorrência de sufixo no final de cada palavra do texto. Exemplos: geno, oico, iano, etc.
20) Gaz	A presença de EG com base em dois gazetteers: um pertencente ao grupo tempo geológico e outro, a algumas rochas sedimentares.

Já o método “K-fold" estabelece a divisão de todo o conjunto de dados em “z" subconjuntos de mesmo tamanho. Ao término de tal particionamento, um subconjunto será a entrada para o treino e o restante para o teste. Por fim, calcular-se-á a acurácia do modelo gerado. Esse processo é execuado “n" vezes, a qual é chamada de folds. Sua condição implica em que todos os subconjuntos criados sejam combinados, alternando os de treino com os de teste. Findada as iterações, determina-se a taxa de erro do classificador, ou seja, o número de exemplos classificados incorretamente pelo número total de exemplos. A taxa

Palavras	Classificação
Na	
Bacia do São Francisco	BACIA SEDIMENTAR
,	
observa-se	
a	
Formação Jequitaí	UNIDADE LITOESTRATIGRÁFICA
.	

Figura 6.6 – Exemplo de saída do CRF

de erro compara o corpus de referência com a classe atribuída pelo sistema classificador. Assim, obtém-se a média mais confiável sobre a legitimidade do modelo para representar os dados de referência.

O terceiro método, “Leave-one-out” utiliza a amostra de dados original. A partir dela, ocorre a divisão dos dados em uma amostra de treinamento e uma de validação. Tal processo é repetido de modo que, a cada observação na amostra, os dados de validação são empregados apenas uma vez. É o mesmo processo que ocorre com o método “K-fold” , onde “n” é o número de iterações da amostra de referência. Considerando uma amostra de tamanho “x”, a taxa de erro da amostragem resulta na soma dos erros em cada iteração divido por “x”. “Leave-one-out” é computacionalmente caro, pois exige várias repetições de treinamento. Frequentemente, ele é utilizado em pequenas amostras.

Nesse contexto, para avaliação e validação do modelo treinado de CRF são utilizados o método de validação cruzada Cross-validation "K-fold” e as métricas de Precisão, Abrangência e Medida-F. A Seção 2.1.1 explanou tais métricas.

7. PROCESSO DE AVALIAÇÃO

O processo de avaliação, fundamentalmente, caracterizou-se por ser experimental. Nesse sentido, importantes etapas foram analisadas, como a geração e o estudo dos resultados obtidos, somados a uma análise das features empregadas. Entre os recursos envolvidos nesse processo, destacam-se as bibliotecas OpenNLP e Mallet, que auxiliaram na etiquetagem, configuração e classificação do corpus, além do pacote de software Weka. A avaliação experimental considerou que as EG extraídas foram reputadas com base na anotação de referência descrita nos Guidelines (Seções 5.1 e 5.1.1). O objetivo da avaliação experimental é:

1) aplicar o método proposto para identificação e classificação de EG;

2) realizar um estudo comparativo entre CRF, J48 Decision Tree e Naive Bayes; e

3) verificar as dificuldades encontradas no processo de EI em textos do Português, no que tange a etiquetagem de sequências estruturadas.

Este Capítulo está organizado da seguinte forma: a primeira Seção descreve os critérios que envolveram o processo de avaliação. A apresentação dos resultados é exibida na segunda Seção 7.2. A Seção 7.3 discutirá tais resultados, analisando entre os atributos considerados, as features empregadas e as classes de maior impacto no domínio proposto. Por fim, a última seção analisará os erros que projetaram os resultados.

7.1 Critérios de Avaliação

Na avaliação experimental proposta, foi considerado o conjunto de vinte features para os três classificadores: CRF, J48 Decison Tree e Naive Bayes. Conforme detalhado anteriormente, as features estão organizadas em features Geológicas (Prefixo, Sufixo e Gazetteer) e Não Geológicas (Word, Tag, Cap, Ini, Simb, PrevW, PrevT, PrevCap, Prev2W, Prev2T, Prev2Cap, NextW, NextT, NextCap, Next2W, Next2T e Next2Cap) (Seção 6.3). Essas features foram implementadas de acordo com os recursos utilizados pelos classificadores. O CRF, por exemplo, anexou as bibliotecas Mallet[1] , NLTK (Natural Language Toolkit)[2] e OpenNLP[3] na implementação do seu algoritmo. Já os classificadores J48 Decison Tree e Naive Bayes tiveram como suporte para a realização do REG (Reconhecimento de Entidades Geológicas) o sistema Weka .

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

7.1.1 Weka

O pacote de software Weka foi a ferramenta utilizada para trabalhar com os classificadores J48 Decison Tree e Naive Bayes. Escrito na linguagem Java, o Waikato Environment for Knowledge Analysis (Weka) tem o objetivo de associar algoritmos, formados por diferentes abordagens, como regressão, associação e classificação. Todos porém, com o propósito de realizar aprendizado de máquina. O Weka procede à análise computacional e probabilística ao data set, o qual é dado como entrada ao algoritmo associado a ele. Ele realiza técnicas de mineração de dados e, indutivamente, realiza soluções de machine learning, a partir dos modelos gerados por esses algoritmos. Dentre as extensões de arquivos que o Weka aceita como entrada estão a arff e a csv.

Nesse contexto, o J48 Decision Tree e o Naive Bayes tiveram como input um arquivo denominado all.csv, o qual é formado pelas features e pela anotação de todas as palavras do corpus de referência. Ambos arquivos são os mesmos usados com o classificador CRF. O all.csv possui as seguintes características: as EG que o compõem são anotadas com uma das treze classes geológicas determinadas na Seção 4.2. Já as palavras que não são EG são anotadas com a classe Outside, indicando que elas não pertencem a nenhuma das classes geológicas. Esse "csv"é formado por vinte e uma colunas. Cada coluna corresponde a uma feature, e a última coluna corresponde a classe. Cada linha desse arquivo representa uma palavra ou um caracter do corpus. Essa anotação tem o objetivo de orientar os dois classificadores para a tarefa de identificação e classificação das EG. A Tabela 7.1 apresenta um exemplo de configuração de input para o J48 e Naive Bayes no Weka. A configuração completa encontra-se no Apêndice C.

Tabela 7.1 – Configuração do arquivo de input dos classificadores J48 e Naive Bayes.

suf	**next Cap**	**word**	**gaz**	**prev Cap**	**cap**	**prevT**	**prevW**	**ini**	**pref**	**tag**	**class**
ário	min	calcários	gaz	min	min	prp	por	min	null	prp	Roc Sed Carbonáticas

7.2 Apresentação dos Resultados

Nesta seção são apresentados os resultados para a tarefa de reconhecimento de EG, conforme os critérios descritos anteriormente. Os experimentos realizados exprimem os valores obtidos para cada classe geológica. Seus resultados são apresentados a partir das métricas Precisão, Abrangência e Medida-F, em relação às EG anotadas no GeoCorpus (Seção 2.1.1.). O primeiro experimento aplicou o modelo obtido pelo classificador CRF (Tabela 7.2). Em geral, o CRF apresentou os melhores resultados em relação aos outros

classificadores. Dentre as treze classes geológicas, as classes Idade, Bacia Sedimentar e Unidade Litoestratigráfica destacaram-se pelos valores mais elevados para Medida-F. A partir dessa métrica, analisaram-se os resultados com as outras duas grandezas.

Observa-se que o CRF se sobressai na Precisão, ou seja, as EG classificadas foram mais altas em cinco classes: Eon, Era, Período, Época e Bacia Sedimentar. Já para a Abrangência as classes com valores mais elevados foram Bacia Sedimentar e Contexto Geológico de Bacia.

Tabela 7.2 – Resultados com o classificador CRF

Categorias	**EG**	**Precisão**	**Abrangência**	**Medida-F**
EON	288	98,73%	53,82%	69,66%
ERA	326	74,43%	49,10%	59,17%
PERIODO	637	83,81%	40,62%	54,72%
ÉPOCA	650	92,21%	44,98%	60,46%
IDADE	796	85,63%	34,62%	49,31%
ROC SED SILICICLÁSTICAS	743	80,31%	48,37%	60,38%
ROC SED CARBONÁTICAS	240	79,81%	27,30%	40,69%
ROC SED QUÍMICAS	5	00,00%	00,00%	00,00%
ROC SED ORGÂNICAS	22	100,00%	09,09%	16,67%
BACIA SEDIMENTAR	243	76,34%	61,73%	68,26%
CONTEXTO GEOLÓGICO DE BACIA	262	33,57%	41,35%	37,05%
UNIDADE LITOESTRATIGRÁFICA	581	73,64%	59,95%	66,10%
OUTRO	739	53,31%	29,52%	38,00%
OUTSIDE CRF		97,86%	99,64%	98,74%
Média Aritmética Ponderada	-	76,78%	43,27%	54,33%

Em relação ao J48 Decision Tree os melhores valores de Medida-F foram para as classes: Contexto Geológico de Bacia, Rochas Sedimentares Carbonáticas, Rochas Sedimentares Siliciclásticas e Outro (Tabela 7.3). Destacando-se as duas últimas classes citadas, pois nelas ele desempenhou-se melhor com todas as métricas. Dentro dessa análise, o J48 obteve os melhores valores de Precisão nas classes Idade, Rocha Sedimentar Siliciclástica, Unidade Litoestratigráfica e Outro. O J48 porém, não conseguiu identificar nenhuma EG do corpus pertencente às classes Época, Rocha Sedimentar Química e Rocha Sedimentar Orgânica, consequentemente ele não as classificou.

Já o Naive Bayes teve o seu melhor desempenho de Medida-F na classe Eon com 70,10% (Tabela 7.4). As outras classes com o melhor desempenho para Medida-F foram Era, Período e Época. Assim como o CRF alcançou 100% de Precisão para Rocha Sedimentar Orgânica, o Naive Bayes também obteve o mesmo percentual nessa classe para a mesma métrica. Da mesma forma, as classes Rocha Sedimentar Carbonática e Contexto Gológico de Bacia obtiveram, isoladamente, os melhores valores de Precisão. Em termos de Abrangência, destaca-se a classe Idade com 84,10%. Tanto o CRF quanto o

Tabela 7.3 – Resultados com o classificador J48.

Categorias	EG	Precisão	Abrangência	Medida-F
EON	288	62,50%	1,70%	3,40%
ERA	326	55,20%	4,90%	8,90%
PERIODO	637	69,20%	1,40%	2,70%
ÉPOCA	650	0,00%	0,00%	0,00%
IDADE	796	100,00%	1,30%	2,70%
ROC SED SILICICLÁSTICAS	743	93,70%	59,20%	72,50%
ROC SED CARBONÁTICAS	240	81,30%	50,60%	62,40%
ROC SED QUÍMICAS	5	0,00%	0,00%	0,00%
ROC SED ORGÂNICAS	22	0,00%	0,00%	0,00%
BACIA SEDIMENTAR	243	70,10%	15,80%	25,80%
CONTEXTO GEOLÓGICO DE BACIA	262	80,60%	34,30%	48,10%
UNIDADE LITOESTRATIGRÁFICA	581	85,70%	0,50%	1,00%
OUTRO	739	79,80%	49,30%	60,90%
Média Aritmética Ponderada	-	71,53 %	19,83 %	25,50 %

Naive Bayes não conseguiram classificar as EG pertencentes à classe Rocha Sedimentar Química.

Tabela 7.4 – Resultados com o classificador Naive Bayes.

Categorias	EG	Precisão	Abrangência	Medida-F
EON	288	91,10%	57,00%	70,10%
ERA	326	68,50%	61,40%	64,70%
PERIODO	637	58,80%	62,60%	60,60%
ÉPOCA	650	57,80%	78,50%	66,60%
IDADE	796	26,40%	84,10%	40,20%
ROC SED SILICICLÁSTICAS	743	85,90%	40,30%	54,80%
ROC SED CARBONÁTICAS	240	91,70%	8,60%	15,70%
ROC SED QUÍMICAS	5	0,00%	0,00%	0,00%
ROC SED ORGÂNICAS	22	100,00%	9,10%	16,70%
BACIA SEDIMENTAR	243	49,40%	35,60%	41,40%
CONTEXTO GEOLÓGICO DE BACIA	262	81,60%	24,70%	37,90%
UNIDADE LITOESTRATIGRÁFICA	581	40,30%	87,40%	55,10%
OUTRO	739	68,30%	18,60%	29,30%
Média Aritmética Ponderada	-	61,44%	55,34%	49,47%

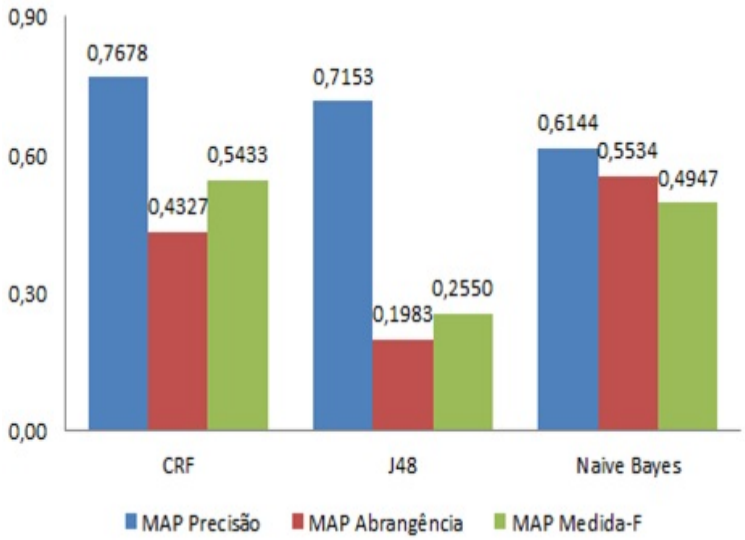
A última linha das Tabelas 7.2, 7.3 e 7.4 apresentam a Média Aritmética Ponderada (MAP) dos acertos das classes geológicas. No contexto dessa tese, o objetivo da MAP é verificar qual classificador apresentou melhor desempenho ao considerar o peso de cada classe. Isso porque há uma tendência de comportamentos diferentes quando as variáveis são divididas em grupos e, quando os grupos são combinados, podendo gerar resultados distintos.

O gráfico, Figura 7.1 apresenta a MAP para as três métricas. Observa-se que o CRF resultou nos maiores valores de MAP para Precisão em relação aos outros classificadores. Além disso, ele foi conciso num maior número de classes, ou seja, cinco classes contra quatro do J48 Decision Tree e duas do Naive Bayes, respectivamente.

Diferentemente da conclusão obtida na Precisão, o Naive Bayes foi melhor classificador na Abrangência, para esse tipo de média. Enquanto o CRF e o J48 Decision Tree foram mais abrangentes em duas e três classes, o Naive Bayes teve mais âmbito em seis classes. Isso quer dizer que, em relação a todo o corpus, ele classificou corretamente um maior número de EG.

Em função dos resultados anteriores da MAP, o CRF apresentou o melhor valor de Medida-F de MAP em relação aos outros classificadores, 0,5433. Seu valor de MAP para Precisão também foi o mais alto: 0,7678 contra 0,7153 e 0,6144. Consequentemente, dentre as EG do conjunto de classes que o CRF classifica, ele classificou o maior número de classes corretamente, quando comparado com os outros dois algoritmos.

Figura 7.1 – Gráfico da Média Aritmética Ponderada para Precisão, Abrangência e Medida-F.



7.3 Discussão dos Resultados

Nesta Seção será apresentada a discussão dos resultados sobre o processo de REG bem como a sua comparação com J48 Decision Tree e o Naive Bayes. A partir dos resultados explanados na Seção anterior, destaca-se o valor de Medida-F de MAP para o CRF. Dentre as melhores classes alcançadas pelo por ele, salienta-se Bacia Sedimentar

Brasileira, uma vez que ela é o ícone para a subárea em questão. A classe Rocha Sedimentar Orgânica teve EG classificadas por esse modelo, mesmo com poucos exemplos identificados. O Naive Bayes e o CRF não classificaram Rocha Sedimentar Química pelos poucos exemplos de EG para essa classe no corpus de referência. Com apenas 5 EG na referida classe, os pesos probabilísticos calculados por esses algoritmos os fizeram gerar um modelo que as classificou como Outside, ou seja, não foram consideradas EG. O J48 Decision Tree e o Naive Bayes alcançaram resultados melhores que o CRF em outras classes, todavia o J48 obteve resultado zero em três classes. Consequentemente, esse valor prejudicou as suas MAP em Abrangência e Medida-F.

Então, nenhuma EG pertencente às classes Rocha Sedimentar Orgânica e Época foram classificadas, no conjunto das vinte features por esse classificador (Tabela 7.3). No caso de Rocha Sedimentar Orgânica, esse fato é esperado. Uma vez que poucos exemplos enviados a um algoritmo que classifica a partir de árvore de decisão, dificulta o seu processo de aprendizado para geração de um bom modelo. Diferentemente da classe Época, que possui 650 EG no corpus de referência e, mesmo assim o J48 Decision Tree zerou nas três métricas. Isso porque quando a execução envolve as features “Word” (features: “PrevW”, “Prev2W”, “NextW”, “Next2W” e a própria “Word”), a árvore de Decisão tem como nodo raiz a feature “Ini”. No momento em que “Ini” é igual a “max”, as palavras são classificadas como Outside e é gerado o nodo Folha. Já “Ini” igual a “min", o próximo caminho da árvore recai na feature “Tag” e a ramificação da árvore continua. Então, a estrutura “Ini”: min possui apenas 2 EG classificadas como Época. As demais 648 EG pertencentes a essa classe possuem “Ini”: max. Como o J48 encontrou uma quantidade maior de registros para a última configuração, ele consegue um maior número de sentenças, a fim de continuar a ramificação da árvore de Decisão.

No CRF e no Naive Bayes, as features importantes para a geração do modelo foram as de POS Tagger e as features Word. Ao analisá-las, constatou-se que tais classificadores calculam pesos mais altos para o seu conjunto de features a partir de:

1) um parser morfológico, que rotule corretamente o maior número de palavras que formam um corpus e,

2) a saída correta das features de janela, originando um contexto que condiz com o corpus de referência, o qual faz parte da EG que será classificada. A Seção Análise de Erros comprovará essa discussão.

Observou-se ainda que, no processo de REG, ocorreu uma grande quantidade de palavras que não recebeu classificação geológica, pois num corpus existem muito mais palavras que não são EN. No conjunto de textos utilizado nessa tese, das 163.790 palavras, 95,20% não são EG (155.932 palavras) e, por isso foram identificadas como Outside. O interressante das classificações feitas, especialmente pelo CRF, é que apesar do grande número de palavras que não são EG, o classificador consegue aprender a rotular aquelas

identificadas como EG. Dependendo do tipo de avaliação, a classe Outside ora é considerada como acerto, ora como erro. A próxima Seção exemplifica tais casos avaliados.

Outra questão que influenciou nos resultados foi a classificação de EG compostas. O CRF, assim como os outros algoritmos, classificam palavra por palavra, pois o corpus de entrada foi processado para que o conjunto de features as caracterize individualmente. Uma possível solução é a união das EG compostas com algum caracter, por exemplo: Bacia_ do_ Paraná. No corpus de entrada, verificou-se que a EG com sinal de hífen são caracterizadas como um todo no conjunto de features. Isso facilita o entendimento do classificador para identificar EG compostas dentro do contexto em que elas estão inseridas. Assim, features de janela, como por exemplo, “PrevW” e “NextW” tornam-se tão relevantes, tanto quanto elas são para as EG simples. Consequentemente, a feature “Word” vai considerar a EG composta como uma única palavra e não mais como palavras individuais. Então, no momento de classificar qualquer outro texto, as EG compostas serão identificadas mais facilmente.

7.3.1 J48 Decision Tree e Features “Words”

Avaliou-se o J48 Decision Tree retirando as features “Words” , pois suas saídas são extremamente amplas e envolvem todo o conjunto de palavras e caracteres do GeoCorpus. A Tabela 7.5 compara os resultados do J48 Decision Tree, no momento em ele foi executado sem as features “Word” e com todo o conjunto de features. São consideradas features “Word”: “PrevW”, “Prev2W”, “NextW”, “Next2W” e a “Word” propriamente dita. Então, constatou-se que ele não zerou a classe Época, quando essas features foram eliminadas. Isso porque a exclusão delas gera uma árvore de Decisão com a feature “Suf” assumindo um dos primeiros nodos da estrutura. Consequentemente, muitas EG de Época são classificadas, pois prefixos, gazetteers e features de POS Tagger participam da nova estrutura da árvore. Assim, observou-se que o J48 teve um ganho de 64,90% ao retirar tais features para a classe Época. Os resultados com as três métricas para esse experimento estão no Apêndice D.1.

Ao analisar a árvore de Decisão do J48, constatou-se que as features impactantes foram “Ini”, “Tag” e “NextW”. No caso da feature “Tag”, os valores “prop” e “prp” foram os únicos rotulados às palavras classificadas como EG. Os demais dezoito tipos diferentes de saídas para “Tag”, não foram rotulados nas EG. Como o objetivo do J48 é “Dividir para conquistar", ele gera a Árvore de Decisão por meio da escolha de um caminho com o maior número de ramificações. Para isso, esse classificador escolhe a configuração de features que possuem a quantidade maior de registros.

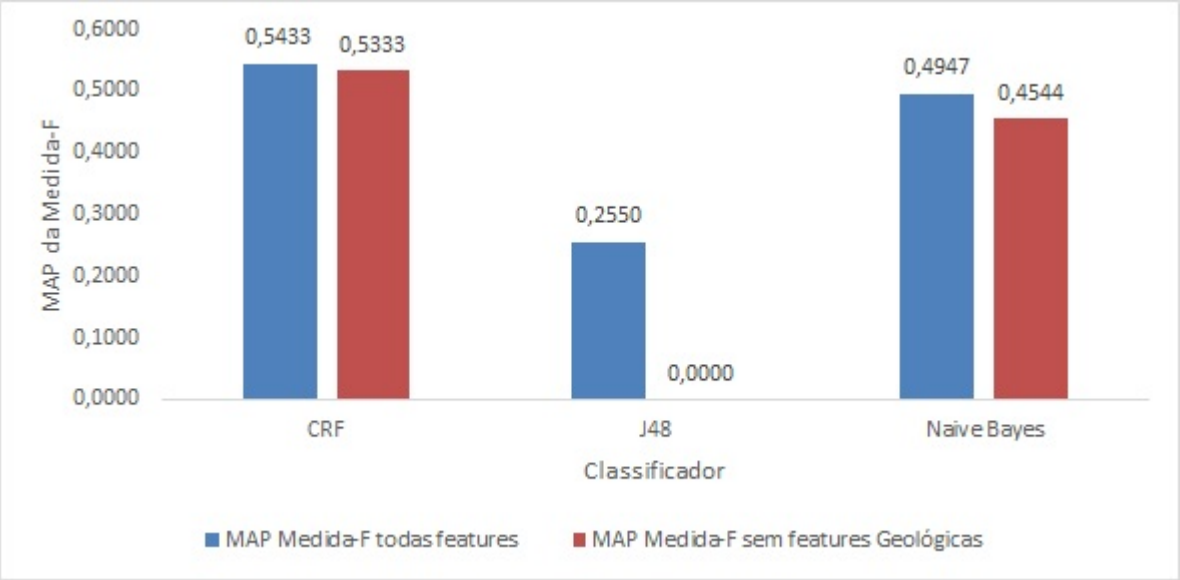
Tabela 7.5 – Resultados do J48 Decision Tree

Classes	Medida-F com todas as features	Medida F sem as features “Words"
EON	3,40%	70,20%
ERA	8,90%	62,10%
PERÍODO	2,70%	13,10%
ÉPOCA	0,00%	64,90%
IDADE	2,70%	17,20%
ROC SED SILICICLÁTICAS	72,80%	28,60%
ROC SED CARBONÁTICAS	62,40%	15,90%
ROC SED QUÍMICA	0,00%	0,00%
ROC SED ORGÂNICA	0,00%	0,00%
BACIA SEDIMENTAR	25,80%	4,60%
CONTEXTO GEOLÓGICO DE BACIA	48,10%	33,90%
UNIDADE LITOESTRATIGRÁFICA	1,00%	45,60%
OUTRO	60,90%	36,20%
Média Aritmética Ponderada	25,50%	34,89%

7.3.2 Features Geológicas

Ainda sobre a discussão das features, realizou-se um experimento, o qual retirou as features geológicas (features Prefixo, Sufixo e Gazetteer), a fim de verificar o quão representativas elas são para a subárea Bacia Sedimentar Brasileira. De acordo com a Figura 7.2, a MAP da Medida-F para cada classificador, comprova que as features Prefixo, Sufixo e Gazetteer são importantes para o REG para os classificadores probabilísticos, entretanto não são fundamentais. No CRF, a MAP permaneceu praticamente a mesma, pois, nas classes de tempo geológico, em que foram utilizadas tais features a diminuição variou de 1% à 4%. No Naive Bayes, as features geológicas mostraram-se mais importantes em relação ao CRF, porque a MAP decaiu em 0,0403 para as mesmas classes. Já para o J48 Decision Tree elas são fundamentais, porque ele classificou todas as EG como Outside. Numa classificação incluindo as features geológicas, elas desenvolvem a ramificação da árvore junto com as demais features. Logo, para a forma de classificação definida pelo J48, constata-se a imprescindibilidade de tais características para essa subárea. A próxima Seção elucidará algumas das questões explanadas na discussão dos resultados.

Figura 7.2 – Gráfico da MAP com todas as features e sem as features geológicas



7.4 Análise de Erros

A partir da apresentação dos resultados e da discussão sobre eles, alguns erros de identificação e classificação entre as EG serão apresentados e analisados. Para encontrar EG marcadas como Falsos Positivos, Falsos Negativos e erros de classificação equipararam-se o corpus de referência com a saída de cada um deles. No caso do CRF, foram geradas duas listas: uma do arquivo de referência e outra da saída desse sistema. As listas tornam a avaliação mais eficiente na busca das EG classificadas pelo CRF quando analisadas na saída xml em formato de texto. Já no J48 Decision Tree e do Naive Bayes, compararam-se dois arquivos: o all.csv (arquivo de referência) e os outputs desses classificadores. Tais informações são obtidas no segmento “Predictions on test data" dos arquivos de saída.

Durante a avaliação dos resultados, verificou-se a marcação de alguns exemplos Falsos Positivos, ou seja, EG classificadas com uma categoria, mas que não deveriam receber classificação geológica. As Tabelas 7.6, 7.7 e 7.8 ilustram exemplos de Falsos Positivos encontrados no CRF, J48 Decision Tree e Naive Bayes, respectivamente. Em cada Tabela, a primeira coluna representa a classificação feita pelos algoritmos e, a segunda coluna, a marcação feita no corpus de referência GeoCorpus. De acordo com a referência, o CRF não deveria ter classificado “Faixa Congo” e “Brasil” como EG, conforme ilustrado na Tabela 7.6. O primeiro equívoco de classificação deve-se à estrutura textual. Geralmente, as Bacias Sedimentares aparecem nos textos antecedidas por “na” (combinação da preposição “em” com o artigo “a”), caracterizadas pela feature “PrevW”. Neste exemplo, ocorre

da seguinte forma: “Subgrupo Schisto-Calcaire na Faixa Congo e com a Formação Olhos d’Água” e assim fez com que o classificador se confundisse em alguns casos. Já no exemplo seguinte, o que influenciou foi o contexto. O classificador entendeu que “Brasil” é uma Unidade geológica, ao passo que ela é Unidade Federativa.

A Tabela 7.7 indica que o J48 Decision Tree classificou “fácies" porque essa palavra recebeu a classificação de acordo com o contexto. Nesta sentença: “o espectro faciológico e as associações de fácies presentes na área de estudo", o anotador não classificou “fácies". Diferentemente de outra sentença do corpus: “sobre as fácies marinhas mais profundas" que recebeu classificação. Já a “bacia" foi classificada como Bacia Sedimentar pela alta frequência que essa palavra aparece anotada no corpus de referência.

Os Falsos Positivos apresentados pelo Naive Bayes (Tabela 7.8) ocorreram porque no primeiro exemplo, “faixa Rio Preto”, foi atribuída a feature “Tag” como nome próprio e as primeiras letras de cada palavra como maiúscula, feature “Ini”, o que faz com que o classificador interprete como EG. Em contrapartida o segundo “Schneidermann" foi classificado como Idade pelas EG pertencentes a classe Idade estarem representadas por muitos nomes estrangeiros nos textos. Como esse classificador, também aprende com os exemplos marcados no GeoCorpus, ele entendeu que essa palavra, por ser estrangeira, poderia receber a referida classe.

Tabela 7.6 – Exemplos classificados como Falsos Positivos pelo CRF

Classificação do CRF	**Classificação da Referência**
<EG Bac Sed>Faixa Congo</EG>	Faixa Congo <O>
<EG Unid Litoestr>Brasil </EG>	Brasil <O>

Tabela 7.7 – Exemplos classificados como Falsos Positivos pelo J48 Decision Tree

Classificação do J48 Decision Tree	**Classificação da Referência**
<EG Cont Geol Bacia>fácies</EG>	fácies <O>
<EG Bac Sed>bacia</EG>	bacia <O>

Tabela 7.8 – Exemplos classificados como Falsos Positivos pelo Naive Bayes

Classificação do Naive Bayes	**Classificação da Referência**
<EG Unid Litoestr>faixa Rio Preto</EG>	faixa Rio Preto <O>
<EG Idade>Schneidermann</EG>	Schneidermann<O>

Dentro da etapa de avaliação dos resultados, realizou-se uma análise dos exemplos Falsos Negativos, isto é, EG classificadas no GeoCorpus, mas que não foram classificação pelos algoritmos. O CRF deixou de classificar, por exemplo, “Bacia Pernambuco-Paraíba" e “Formação Barreiras" (Tabela 7.9. No primeiro caso, “Bacia Pernambuco-Paraíba” houve um erro de PosTagger pelo Parser OpenNLP. Na sentença: “Portanto, restam apenas para a Bacia Pernambuco-Paraíba as subbacias Miriri, Alhandra e Olinda", o parser

deveria ter etiquetado a palavra “subbacia” (feature “Next2T") como substantivo e a marcou como preposição. Também a palavra “a" (“feature NextT”) deveria ter sido rotulada como artigo e recebeu ponto. Isso influenciou no peso dessas features, fazendo com que o classificador não a considere uma EG. No segundo caso, “Formação Barreiras”, as features foram atribuídas a cada uma das palavras que compõem essa EG composta, ou invés de a caracterizarem como uma única EG. Diferentemente do exemplo anterior em que a EG “Pernambuco-Paraíba” é unida por hífen.

O J48 Decision Tree não classificou “siliciclásticas" pelos poucos exemplos encontrados para esse tipo de EG, conforme ilustrado na Tabela 7.10. Uma vez que um número mínimo de especialistas classificou essa palavra como uma Rocha Sedimentar Siliciclástica, devido a interpretação desses, influenciou na classificação. Na sequência, um outro caso de Falso-Negativo ocorreu com a EG “lamito", devido aos poucos exemplos para esse tipo de rocha.

Na Tabela 7.11, tem-se primeiramente, a EG “calcário" classificada como Outside. O Naive Bayes não a categorizou como Rocha Sedimentar Carbonática, porque de acordo com a sentença: “no interior do Calcário do subsolo", a palavra que a antecede é uma preposição e a feature “PrevT" teve como saída um substantivo. Ainda nesta sentença, a feature “NextT" recebeu o atributo numeral, enquanto deveria ser etiquetada também como preposição. O Parser se equivocou ao etiquetá-las morfologicamente. O segundo Falso-Negativo ocorreu porque a marcação de “Bacia” está separada no conjunto de features da EG “Irecê", ou seja, cada uma dessas palavras foi analisada separadamente. Então, em alguns casos em que aparece esse tipo de EG composta, o classificador não marca corretamente.

Tabela 7.9 – Exemplos classificados como Falsos Negativos pelo CRF

Classificação do CRF	**Classificação da Referência**
Bacia Pernambuco-Parnaíba <O>	<EG Bacia Sed>Bacia Pernambuco-Parnaíba</EG>
Formação <O >Barreiras <O>	<EG Unid Litoestr>Formação Barreiras</EG>

Tabela 7.10 – Exemplos classificados como Falsos Negativos pelo J48 Decision Tree

Classificação do J48 Decision Tree	**Classificação da Referência**
siliciclásticas <O>	<EG Roc Sed Siliciclástica>siliciclásticas</EG>
lamito <O>	<EG Roc Sed Siliciclástica>lamito</EG>

Tabela 7.11 – Exemplos classificados como Falsos Negativos pelo Naive Bayes

Classificação do Naive Bayes	**Classificação da Referência**
Calcário <O>	<EG Roc Sed Carbonática>Calcário</EG>
Bacia<O> de<O> <EG Bac Sed>Irecê</EG>	<EG Bac Sed>Bacia de Irecê</EG>

Outro tipo de análise realizada, envolveu os erros de classificação, os quais foram computados pelas Matrizes de Confusão, representadas pelas Figuras 7.3, 7.4 e 7.5. No gráfico da matriz de Confusão, as células da Diagonal Principal apresentam a quantidade de exemplos do corpus de teste que os algoritmos acertaram, ou seja, os Verdadeiros Positivos. As células que não fazem parte da Diagonal Principal denotam os Falsos Positivos e Falsos Negativos.

Para a criação das matrizes de Confusão foi realizada a normalização dos dados, isto é, todos os valores estão entre 0-1. Cada elemento da matriz está representado por uma cor, variando cromaticamente entre o azul claro e o azul escuro. O azul claro ilustra uma baixa correlação (igual ou próximo à 0) entre as classes e, azul escuro ilustra uma alta correlação (próxima ou igual à 1). Quanto mais escuro o tom de azul, maior coocorrências entre as classes. Um classificador perfeito, possui todos os elementos da matriz de Confusão na Diagonal Principal e as células restantes estariam em branco. Elementos fora da diagonal principal são elementos que foram erroneamente classificados. Cada letra corresponde a uma das classes geológicas:

A = Éon

B = Era

C = Período

D = Época

E = Idade

F = Rocha Sedimentar Siliciclástica

G = Rocha Sedimentar Carbonática

H = Rocha Sedimentar Química

I = Rocha Sedimentar Orgânica

J = Bacia Sedimentar

K = Contexto Geológico de Bacia

L = Unidade Litoestratigráfica

M = Outro

N = Outside

Esse tipo de gráfico é interessante, pois mostra de forma visual a distribuição das classificações, facilitando a identificação das categorias que foram mal classificadas. Ao interpretar as matrizes de Confusão, observa-se que o CRF apresentou o maior número de EG Verdadeiros Positivos em relação ao J48 Decision Tree e ao Naive Bayes, pois ele apresentou um bom equilíbrio de resultados entre Precisão e Abrangência.

Na Figura 7.3, a qual representa o CRF, pode-se identificar que a classe Contexto Geológico de Bacia (coluna K) contém uma grande quantidade de classificações erradas,

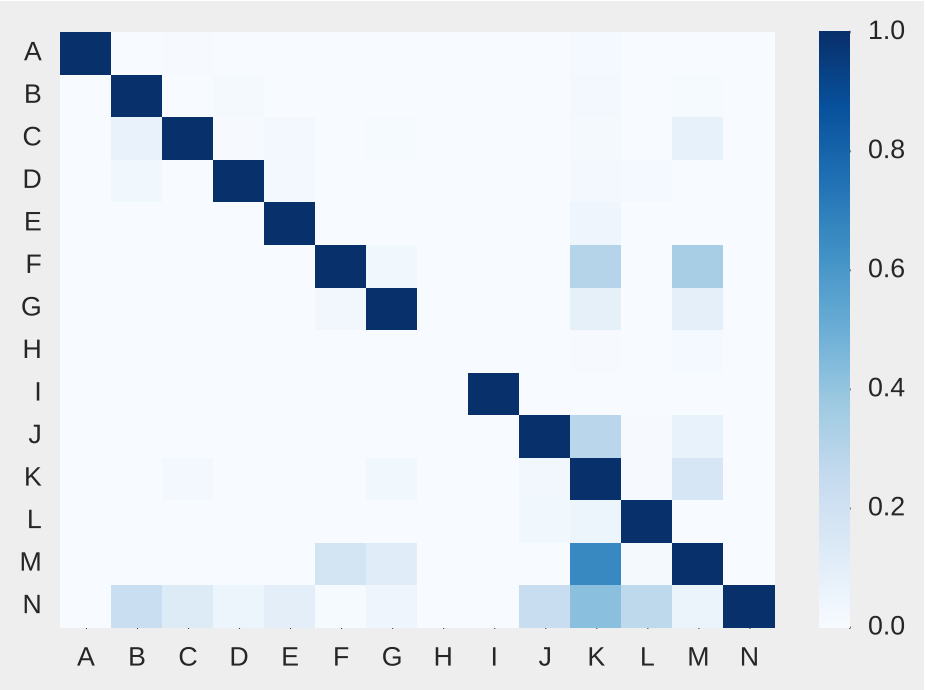


Figura 7.3 – Matriz de Confusão gerada pelo CRF

visivelmente com as classes "Rocha Sedimentar Siliciclástica", "Rocha Sedimentar Carbonática", "Bacia Sedimentar", "Outro"e "Outside". Essas classificações ocorreram devido à identificação das EG pertencentes a ela, originar mais de uma interpretação pelos especialistas durante o processo de anotação. Por exemplo: alguns anotaram “zonas de cisalhamento" como EG, outros marcaram apenas “cisalhamento" como EG. Consequentemente, isso confundiu os algoritmos na tarefa de REG.

Já na Figura 7.4 constata-se que o J48 Decision Tree não classificou nenhuma EG para as classes “D” (Época), “H” (Rocha Sedimentar Química) e “I” (Rocha Sedimentar Orgânica). Na classe Época, a árvore de Decisão mostrou que a feature “Ini” : min possui apenas 2 EG classificadas como Época. Em Rochas Sedimentares Química e Orgânica, a não classificação aconteceu pelo corpus de referência possuir poucos exemplos de EG pertencentes à essa classe. No gráfico representado pela Figura 7.5 é claramente visível que o Naive Bayes também não encontrou nenhuma EG da classe “H” (Rocha Sedimentar Química). Devido aos poucos exemplos definidos para essa classe (5 EG), o referido classsificador não conseguiu aprender a rotulá-las no corpus de teste.

A classe Outside teve grande volume de classificações, pois há muitas palavras que não são EG. Significa que, se os algoritmos encontrarem dificuldade de realizar a classificação, eles irão etiquetar as instâncias do GeoCorpus como Outside. Essa é uma decisão de classificação previsível dentro do processo de Machine Learning, pois a tendência dos algortimos é rotular as palavras de acordo com a classe que possui o maior número de exemplos. Mesmo assim, as três técnicas conseguiram realizar a classificação das EG com

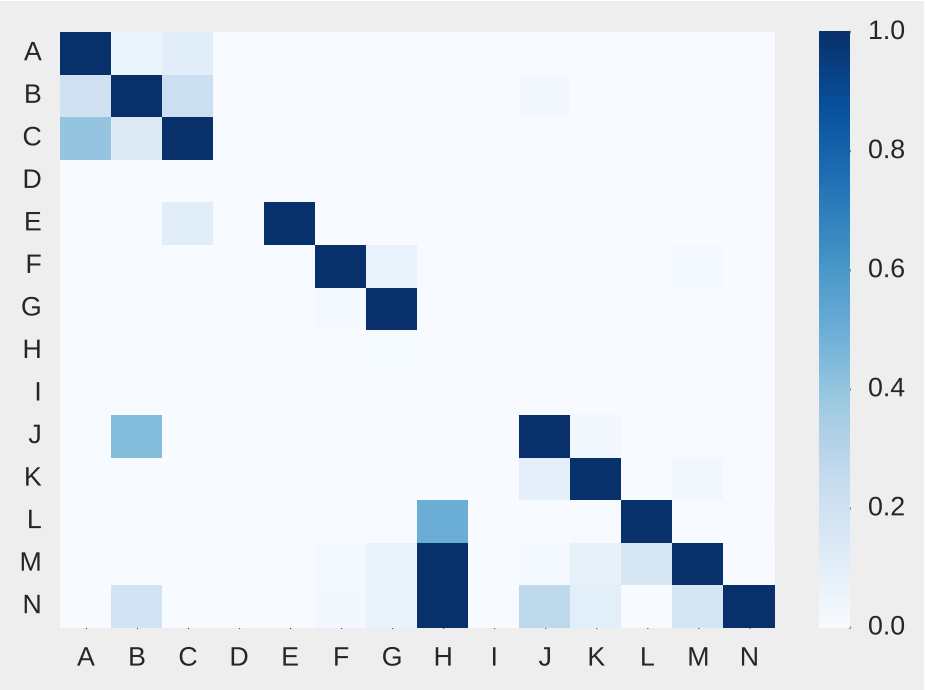


Figura 7.4 – Matriz de Confusão gerada pelo J48 Decision Tree

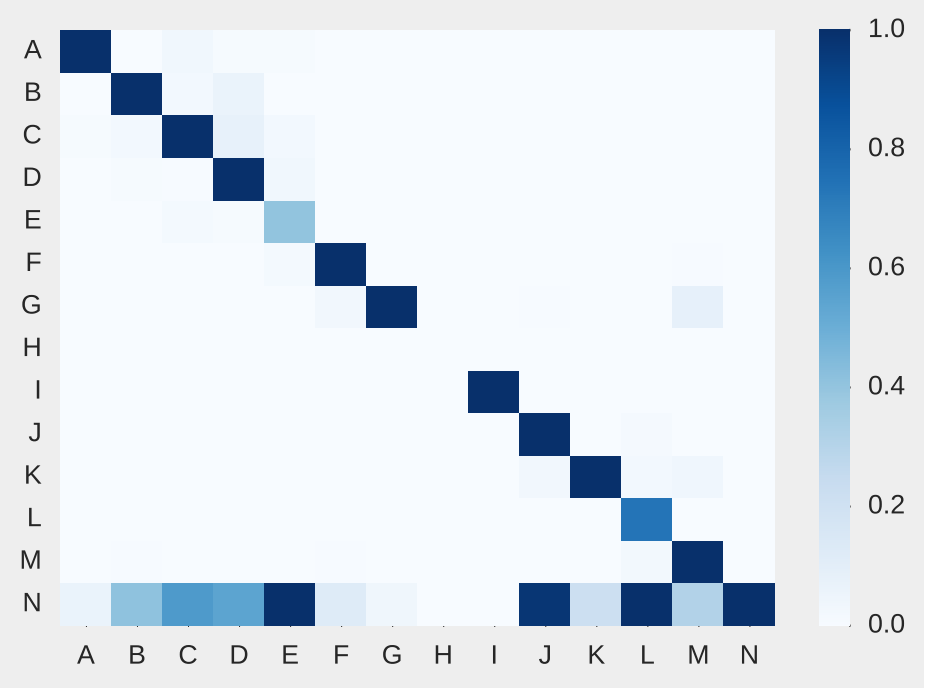


Figura 7.5 – Matriz de Confusão gerada pelo Naive Bayes

valores acima de 60% de Precisão na maioria das classes. A partir das análises nas matrizes de Confusão, alguns exemplos foram selecionados a fim de enriquecer esse estudo.

A Tabela 7.12 elucida que o CRF deveria ter classificado “quartzo" como Outro ao invés de Rocha Sedimentar Siliciclástica, na sentença: “Misturados com os grãos carbonáticos ocorrem grãos de quartzo da granulação areia fina”. Justifica-se a falha devido a feature “Next2W” ter recebido a palavra “mica”, ao passo que deveria ser a palavra “granulação”.

A feature NextT recebeu a tag de pontuação, enquanto o correto é preposição. Também houve ausência da declaração da feature “NextW”. Outro erro envolveu a sentença: “A maior parte dos depósitos glacio-influenciados do Grupo Itararé foi inicialmente interpretada como glacio-continental”. A EG “Grupo Itararé", recebeu a classe “Contexto Geológico de Bacia”, enquanto deveria ser “Unidade Litoestratigráfica”. Em relação à palavra “Grupo" observou-se que à feature “PrevT” deveria ser atribuída um verbo e não preposição. Na EG “Itararé”, a feature “Next2T” recebeu a tag verbo, sendo que o correto é advérbio, já que a palavra etiquetada é “inicialmente”.

Tabela 7.12 – Exemplos de erros de classificação pelo CRF

Classificação do CRF	**Classificação da Referência**
<EG Roc Sed Carbonatica>quartzo</EG>	<EG Outro>quartzo</EG>
<EG Cont Geol Bacia>Grupo Itararé</EG>	<EG Unid Litoestr>Grupo Itararé</EG>

Em relação ao J48 Decision Tree, a primeira EG classificada erroneamente foi “dolomito” ilustrada na Tabela 7.13. De acordo com a sentença “exibindo, da base para o topo, as formações Carrancas (diamictito, ritmito, arenito), Sete Lagoas calcário e dolomito, com intercalações pelíticas”, as features “Tag" e “Prev2T” foram etiquetadas como preposição e deveriam receber substantivo.

A EG “Áreas Cratônicas" deveria ter sido classificada como Contexto Geológico de Bacia (Tabela 7.13). Dada a sentença: “Áreas cratônicas são regiões da crosta da Terra”, o J48 não marcou “Áreas” e classificou somente “Cratônicas” como Bacia Sedimentar. Notou-se que a primeira palavra que compõem a EG (“Áreas”) recebeu a etiqueta conjunção coordenada para a feature “Next2T”, enquanto deveria ser verbo. Outra justificativa foi a saída “min" para a feature “Cap”, enquanto ela deveria receber “maxmin”. Já na EG “cratônicas” observou-se erro em duas features: “NextT” obteve como saída “ponto”, contudo deveria ser anotada por “verbo” e na “Tag” ocorreu preposição, não obstante é adjetivo.

Tabela 7.13 – Exemplos de erros de classificação pelo J48 Decision Tree

Classificação do J48 Decision Tree	**Classificação da Referência**
<EG Outro>dolomito</EG>	<EG Roc Sed Carbonática>dolomito</EG>
Áreas <O> <EG Cont Geol Bacia>Cratônicas</EG>	<EG Cont Geol Bacia>Áreas Cratônicas</EG>

Por fim, salientam-se dois casos para o Naive Bayes (Tabela 7.14). A partir da sentença: “Formação Tinguis (Cenozoico) na região e em Campo Largo, além de fraturas e falhas com indicadores cinemáticos”, observou-se que foi classificado a EG “Cenozoico” como Idade, embora seja Era. A justificativa encontra-se na feature “Next2T”, etiquetada como adjetivo, ainda que seja preposição para a palavra “na”.

No outro caso, em “depositados na Bacia Sedimentar Marginal Pernambuco-Paraíba.”, a EG “Bacia Sedimentar Marginal Pernambuco-Paraíba” teve as duas últimas palavras que a compõem classificadas como Unidade Litoestratigráfica. Nesse caso, o motivo não foi a

marcação das suas features e sim a dificuldade de classificar EG compostas e extensas. Além disso, o contexto que a envolveu confundiu o classificador, pois o texto tem como propósito apresentar uma síntese sobre os diversos aspectos estruturais e morfológicos da Formação Barreiras, classificada como Unidade Litoestratigráfica em várias partes dele.

Tabela 7.14 – Exemplos de erros de classificação pelo Naive Bayes

Classificação do Naive Bayes	**Classificação da Referência**
<EG Idade>Cenozoico</EG>	<EG Era>Cenozoico</EG>
<EG Bac Sed>Bacia Sedimentar Marginal</EG> <EG Un Litoestr>Pernambuco-Paraíba</EG>	<EG Bac Sed> Bacia Sedimentar Marginal Pernambuco-Paraíba </EG>

Em síntese, a partir da presente análise, os erros que mais chamaram a atenção, nos três classificadores, foram para classificação em EG compostas. Os equívocos cometidos pelo parser, na etiquetagem morfológica também influenciaram na saída das features de POS Tagger. O contexto, ou seja, a ligação entre as palavras que compõem o corpus e os poucos exemplos em determinadas classes prejudicaram a classificação de certas EG. No próximo capítulo, serão apresentadas as considerações finais.

8. CONSIDERAÇÕES FINAIS

A presente tese de doutorado teve como objetivo tratar a tarefa reconhecimento de EN relevantes no domínio de Geologia, especificamente, na subárea Bacia Sedimentar Brasileira, em textos do português. É relevante destacar que a língua portuguesa é carente de recursos para PLN, principalmente para EI de termos geológicos, dentro da subárea definida neste trabalho. Então, a tarefa de REN torna-se mais desafiadora.

Conforme apresentado ao longo deste trabalho, diferentes técnicas computacionais de aprendizado de máquina têm sido pesquisadas e aplicadas, assim como o uso de recursos linguísticos para a resolução desse problema. Adicionalmente, a literatura nos proporciona uma diversidade de modelos para avaliação de métodos de REN.

A partir de uma análise sobre a tarefa de reconhecimento de EN em vários domínios, buscou-se encontrar referências para o domínio da Geologia (Capítulo 3). Dessa pesquisa, muito poucos trabalhos foram realizados com o propósito de reconhecer EN para a Geologia e nenhum, especificamente, para a subárea Bacia Sedimentar Brasileira. Tal estudo originou na proposta de um método baseado em CRF para o REN geológicas, na subárea Bacia Sedimentar Brasileira, em textos do português, bem como avaliá-lo com outras técnicas.

Construiu-se um corpus de referência nomeado GeoCorpus, o qual é formado por boletins de Geociências da Petrobrás, artigos e algumas dissertações referentes à Bacia Sedimentar Brasileira (Capítulo 5). Os documentos científicos selecionados estão relacionados com a subárea definida.

Uma das finalidades de aplicar a técnica de aprendizado de máquina supervisionado CRF é analisar o comportamento do modelo mediante um corpus anotado por especialistas. O CRF gera um modelo probabilístico com aplicação eficaz em textos de linguagem natural, os quais envolvem problemas de etiquetagem de sequências estruturadas. Tal modelo pode alcançar uma classificação globalmente ideal e utilizar um conjunto de features simples, a fim de exprimir estruturas contextuais complexas. Para isso, definiram-se as features de classificação das EG. Na Seção 6.3, conjuntos de features foram definidos baseados na informação semântica, na etiquetagem morfológica POS Tagger, na estrutura das palavras identificadas como EG geológicas e nas suas classes.

Na avaliação experimental, averiguaram-se os resultados obtidos para três clasificadores.Também, realizou-se um exame crítico das features empregadas (Capítulo 7). A partir de tal análise, pode-se constatar que as features geológicas são importantes, mas as não geológicas são features genéricas que podem ser usadas para aquelas EG que não tenham prefixo, sufixo ou gazetter inseridos no seu conjunto de características. Mesmo asssim, as featues geológicas e as não geológicas complementam-se, o que melhora o resultado final em todas as classes, as quais as features geológias são utilizadas.

Ao analisar-se os três métodos de classificação, os melhores resultados alcançados pelo CRF de Medida-F foram para as classes Idade, Bacia Sedimentar e Unidade Litoestratigráfica. O CRF obteve os melhores valores de Precisão para o maior número de classes determinadas, ou seja, de 13 classes, 5 foram mais precisas e uma sexta equiparou-se com 100% de Precisão junto com o Naive Bayes, na classe Rocha Sedimentar Orgânica.

O J48 Decision Tree apresentou os mais altos valores de Medida-F em Rocha Sedimentar Siliciclástica, Rocha Sedimentar Carbonática, Contexto Geológico de Bacia e Outro. Em compensação, ele não classificou nehuma EG das classes Época, Rocha Sedimentar Química e Rocha Sedimentar Orgânica. Em ambas as classes detectou o grande número de registros pertencentes a classe Outside com valor “max” para a feature “Ini”. Essa feature é o nodo raiz da árvore de Decisão, o que gerou uma classificação errônea. Acresenta-se também a baixa representatividade das classes Rochas Sedimentares Química e Orgânica, como outro motivo para a não classificação das suas EG.

O Naive Bayes destacou-se com os melhores resultados de Medida-F nas classes Éon, Era, Período e Época. Mesmo assim, a classe Rocha Sedimentar Química não foi classificada por ele, por possuir poucos exemplos que o façam gerar um modelo de classificação para esse tipo de EG. Por ter conseguido os melhores valores de Abrangência, sua MAP foi a mais alta.

Ao considerar as características de desenvolvimento de cada uma dessas técnicas, para a geração de um modelo de classificação na subárea Bacia Sediementar Brasileira, verifica-se que: o classificador J48 Decision Tree tem como desvantagem a dificuldade de gerenciar grande quantidade de valores para uma determinada feature. Também possui dificuldade para gerenciar grande quantidade de classes e de registros para uma mesma classe. Por exemplo dos 160.633 registros, 160.226 pertecem a classe Outside. A sua vantagem é a possibilidade de interpretar a ávore de Decisão para descobrir quais são as features determinantes para atribuir às classes selecionadas. O CRF e o Naive Bayes, por gerarem modelos gráficos probabilísticas apresentam como vantagem a sua usabilidade para tarefas diferentes de aprendizado. Por exemplo, a tarefa de previsão, com o objetivo de adquirir os resultados mais prováveis para o data set de entrada e, o diagnóstico, a fim de obter as causas mais prováveis para as consequências observadas. A partir da avaliação dos três classificadores, conclui-se que o CRF foi o melhor, pois ele apresentou o valores mais altos de MAP para Precisão e Medida-F com 76,78% e 54,33%, respectivamente.

8.1 Contribuições

As principais contribuições desta tese são:

- A proposta de um método probabilístico para o reconhecimento de EG, aplicado em textos do Português, dentro da subárea Bacia Sedimentar Brasileira. O CRF gerou um modelo automático, que está disponível para classificar textos na referida subárea. Tal modelo pode analisar dados maiores, mais complexos e oferecer resultados mais rápidos e precisos, mesmo em escala muito grande.
- O desenvolvimento de um trabalho científico inédito e disponível à comunidade acadêmica. Há muitos artigos e journals que realizam REN para domínios como Medicina, Biomedicina e Jornalismo. Em contrapartida, não existem trabalhos que executem o REG para o português e, menos ainda, sobre o reconhecimento de EG neste contexo.
- Um corpus anotado de Geologia, específico para Bacia Sedimentar Brasileira está disponível para a comunidade acadêmica. Ao longo do desenvolvimento do GeoCorpus, deparamos com desafios como o processo de anotação. Primeiro, foi a grande dificuldade de encontrar anotadores com disponibilidade de classificar os textos para gerar um corpus de referência. Outro desafio foi a confiança na anotação, ou seja, conseguir especialistas que tenham conhecimento na subárea para anotar os textos. Nesse sentido, verificou-se que a necessidade de um corpus anotado é uma questão muito séria e precária. Observou-se que escolha de um domínio específico, requer especialistas na área, os quais precisarão de tempo e esforço para anotarem as EN. O artigo “Processo de construção de um corpus anotado com Entidades Geológicas visando REN", aceito no Symposium in Information and Human Language Technology (STIL), 2017 consolida a referida contribuição.
- O conjunto de features para o aprendizado dessa subárea Geológica. A partir de um estudo da literatura sobre a tarefa de REN, definiram-se as features específicas que melhor caracterizaram as EG.
- Como contribuição e objetivo final da tese, tem-se o auxílio da pesquisa de informação e do conhecimento. Visto que, a classificação de EG dentro do contexto Bacia Sedimentar Brasileira é um assunto de importância tanto para a área acadêmica, na diponibilidade de recusos especializados, quanto para a área econômica, como na construção civil e na exploração de petróleo.

8.2 Trabalhos Futuros

O aprendizado de máquina é uma ciência que não é nova, mas que está ganhando novo impulso por ser desafiadora e complexa. Especialmente, quando aplicada em domínios específicos, pouco estudados e explorados. Assim, tem-se como um dos trabalhos futuros aplicar o corpus desenvolvido em outras técnicas de aprendizado de máquina, de modo individual e combinado, como por exemplo, Redes Neurais e Suport Vector Machine (SVM).

Pretende-se expandir o corpus de referência, acrescentando mais textos anotados no GeoCorpus. Principalmente, expandir as classes que possuem o menor número de EG como Rocha Sedimentar Química e Rocha Sedimentar Orgânica.

Outro trabalho futuro compreende a melhoria da estruturação dos dados para classificação de EG compostas. Adicionalmente, melhorar o processo de POS Tagger no GeoCorpus a fim de tornar a feature “Tag” mais eficaz, aumentando o peso probabilístico dessa característica.

Ainda um outro trabalho importante seria investigar sobre as relações entre as entidades do domínio e desenvolver técnicas de extração automática dessas relações.

Por fim, acreditamos que seria interessante ampliar o domínio geológico, incluir outras subáreas, bem como fazer um trabalho mais aprofundando com as classes e seus agrupamentos.

REFERÊNCIAS BIBLIOGRÁFICAS

[AHvdH+15] Akhondi, S.; Hettne, K. M.; van der Horst, E.; van Mulligen, E. M.; Kors, J. "Recognition of chemical entities: combining dictionary-based and grammar-based approaches", *Journal of Cheminformatics*, vol. 7, Jan 2015, pp. 1–11.

[AKP15] Althobaiti, M.; Kruschwitz, U.; Poesio, M. "Combining minimally-supervised methods for arabic named entity recognition", *Transactions of the Association for Computational Linguistics*, vol. 3, Mai 2015, pp. 243–256.

[AL13] Atdag, S.; Labatut, V. "A comparison of named entity recognition tools applied to biographical texts". In: Proceedings of the Systems and Computer Science, 2013, pp. 228–233.

[AP73] Asmus, H. E.; Ponte, F. "The Brazilian Marginal Basins". Springer, 1973, cap. 3, pp. 87–133.

[AV14] Amaral, D. O. F. d.; Vieira, R. "Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields", *Linguamática*, vol. 6–1, Jul 2014, pp. 41–49.

[BdSVR08] Bruckschen, M.; de Souza, J. G. C.; Vieira, R.; Rigo, S. "Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas". Linguateca, 2008, cap. 14, pp. 247–260.

[BMSW97] Bikel, D. M.; Miller, S.; Schwartz, R.; Weischedel, R. "Nymble: A high-performance learning name-finder". In: Proceedings of the 5^{th} Conference on Applied Natural Language Processing, 1997, pp. 194–201.

[BON03] Bender, O.; Och, F. J.; Ney, H. "Maximum entropy models for named entity recognition". In: Proceedings of the 7^{th} Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2003, pp. 148–151.

[BSCB10] Batista, D.; Silva, M. J.; Couto, F.; Behera, B. "Geographic signatures for semantic retrieval". In: Proceedings of the 6^{th} Workshop on Geographic Information Retrieval, 2010, pp. 18–19.

[BSVG03] Bizzi, L. A.; Schobbenhaus, C.; Vidotti, R. M.; Gonçalves, J. H. "Geologia, Tectônica e Recursos Minerais do Brasil: texto, mapas e SIG". Brasília, BR: Companhia de Pesquisa de Recursos Minerais, 2003, 692p.

[Car12] Carvalho, W. S. “Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina”, Tese de Doutorado, Universidade de São Paulo, 2012, 80p.

[CC03] Curran, J. R.; Clark, S. “Language independent ner using a maximum entropy tagger”. In: Proceedings of the 7^{th} Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2003, pp. 164–167.

[CD12] Chatzis, S. P.; Demiris, Y. “The echo state conditional random field model for sequential data modeling”, *Expert System Applications*, vol. 39, Set 2012, pp. 10303–10309.

[CDF14] Cohen, K. B.; Demner-Fushman, D. “Biomedical natural language processing”. Amsterdam, NED: John Benjamins Publishing Company, 2014, 171p.

[CFGF13] Cohen, K.; Finney, S.; Gibbard, P.; Fan, J.-X. “The ics international chronostratigraphic chart”, *International Union of Geological Sciences*, vol. 36, Set 2013, pp. 199–204.

[Cir01] Ciravegna, F. “Adaptive information extraction from text by rule induction and generalisation”. In: Proceedings of the 17^{th} International Joint Conference on Artificial Intelligence, 2001, pp. 1251–1256.

[CN03] Chieu, H. L.; Ng, H. T. “Named entity recognition with a maximum entropy approach”. In: Proceedings of the 7^{th} Human Language Technology Conference North American Chapter of the Association for Computational Linguistics, 2003, pp. 160–163.

[Coh] Cohen, W. W. “Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data”. Capturado em: https://www.bibsonomy.org/bibtex/209e8bd7203843908df514bedaba4d377/mortimer_m8, Maio 2004.

[Coh68] Cohen, J. “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit”, *Psychological bulletin*, vol. 70–4, Out 1968, pp. 213–220.

[CPCT14] Collier, N.; Paster, F.; Campus, H.; Tran, A. M.-v. “The impact of near domain transfer on biomedical named entity recognition”. In: Proceedings of the 5^{th} International Workshop on Health Text Mining and Information Analysis, 2014, pp. 11–20.

[Cri08] Cristina Mota, D. S. “Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM”. Lisboa, PT: Linguateca, 2008, 450p.

[CSM05] Chaves, M.; Silva, M.; Martins, B. “A geographic knowledge base for semantic web applications”. In: Proceedings of the 20^{th} Brazilian Symposium on Databases, 2005, pp. 15.

[dABV15] do Amaral, D. O. F.; Buffet, M.; Vieira, R. “Comparative analysis between notations to classify named entities using conditional random fields”. In: Proceedings of Symposium in Information and Human Language Technology, 2015, pp. 27–31.

[DMP+04] Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S.; Weischedel, R. M. “The automatic content extraction (ACE) program-tasks, data, and evaluation”. In: Proceedings of the 4^{th} International Conference on Language Resources and Evaluation, 2004, pp. 837–840.

[DNF+05] Dingare, S.; Nissim, M.; Finkel, J.; Manning, C.; Grover, C. “A system for identifying named entities in biomedical text: How results from two evaluations reflect on both the system and the evaluations”, *Comparative and Functional Genomics*, vol. 6, Fev 2005, pp. 77–85.

[DPDPL97] Della Pietra, S.; Della Pietra, V.; Lafferty, J. “Inducing features of random fields”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, Abr 1997, pp. 380–393.

[DPMR14] Danger, R.; Pla, F.; Molina, A.; Rosso, P. “Towards a protein–protein interaction information extraction system: Recognizing named entities”, *Knowledge-Based Systems*, vol. 57, Fev 2014, pp. 104–118.

[FDM+05] Finkel, J.; Dingare, S.; Manning, C. D.; Nissim, M.; Alex, B.; Grover, C. “Exploring the boundaries: gene and protein identification in biomedical text”, *BioMed Central Bioinformatics*, vol. 6, Mai 2005, pp. 5.

[FMS+10] Freitas, C.; Mota, C.; Santos, D.; Oliveira, H. G.; Carvalho, P. “Second harem: Advancing the state of the art of named entity recognition in portuguese”. In: Proceedings of the 4^{th} International Conference on Language Resources and Evaluation, 2010, pp. 3630–3637.

[Fol80] Folk, R. L. “Petrology of sedimentary rocks”. Texas, EUA: Hemphill Publishing Company, 1980, 190p.

[Gab13] Gabbayl, I. “Recognising specific named entities in a new restricted domain using Conditional Random Fields”, Tese de Doutorado, University of Limerick, 2013, 162p.

[GJ13] Grotzinger, J.; Jordan, T. “Para Entender a Terra”. Porto Alegre, BR: Bookman, 2013, 768p.

[GM90] Gabaglia, G.; Milani, E. “Origem e Evolução de Bacias Sedimentares”. Rio de Janeiro, BR: E.J. Editora e Comunicação Integrada, 1990, 415p.

[HAB+97] Hobbs, J. R.; Appelt, D.; Bear, J.; Israel, D.; Kameyama, M.; Stickel, M.; Tyson, M. “Fastus: A cascaded finite-state transducer for extracting information from natural-language text”, *Finite-state language processing*, Jun 1997, pp. 383–406.

[HK99] Hallsworth, C.; Knox, R. “British geological survey rock classification scheme. Classification of sediments and sedimentary rocks”, Relatório Técnico, British Geological Survey Programmes, 1999, 44p.

[HS17] He, H.; Sun, X. “F-score driven max margin neural network for named entity recognition in chinese social media”. In: Proceedings of the 15^{th} Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 713–718.

[JAW+11] Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. “Oscar4: a flexible architecture for chemical text-mining”, *Journal of cheminformatics*, Jun 2011, pp. 1–12.

[JD17] Jochim, C.; Deleris, L. “Named entity recognition in the medical domain with constrained CRF models”. In: Proceedings of the 15^{th} Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 839–849.

[Jia12] Jiang, J. “Information extraction from text”. Springer, 2012, cap. 2, pp. 11–41.

[JM09] Jurafsky, D.; Martin, J. H. “Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition”. New Jersey, EUA: Pearson Prentice Hall, 2009, 1024p.

[JPM+11] Jaiswal, A.; Pezanowski, S.; Mitra, P.; Zhang, X.; Xu, S.; Turton, I.; Klippel, A.; MacEachren, A. M. “Geocam: A geovisual analytics workspace to contextualize and interpret statements about movement”, *Journal of Spatial Information Science*, vol. 3, Dez 2011, pp. 65–101.

[KLCS15] Keretna, S.; Lim, C. P.; Creighton, D.; Shaban, K. B. “Enhancing medical named entity recognition with an extended segment representation technique”, *Computer methods and programs in biomedicine*, vol. 119–2, Abr 2015, pp. 88–100.

[KLW15] Kim, S.; Lu, Z.; Wilbur, W. J. “Identifying named entities from pubmed® for enriching semantic categories”, *BioMed Central Bioinformatics*, vol. 16, Fev 2015, pp. 1–10.

[KOT+04] Kim, J.-D.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y.; Collier, N. “Introduction to the bio-entity recognition task at JNLPBA”. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, 2004, pp. 70–75.

[Kri12] Krippendorff, K. “Content analysis: An introduction to its methodology”. Califórnia, EUA: Sage, 2012, 456p.

[KRL+15] Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; et al.. “The chemdner corpus of chemicals and drugs and its annotation principles”, *Journal of cheminformatics*, Jan 2015, pp. 1–17.

[LMP01] Lafferty, J. D.; McCallum, A.; Pereira, F. C. N. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”. In: Proceedings of the 18^{th} International Conference on Machine Learning, 2001, pp. 282–289.

[LPN17] Liu, F.; Perez, J.; Nowson, S. “A language-independent and compositional model for personality trait recognition from short texts”. In: Proceedings of the 15^{th} Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 754–764.

[LVC+15] Liu, H.; Verspoor, K.; Comeau, D. C.; MacKinlay, A. D.; Wilbur, W. J. “Optimizing graph-based patterns to extract biomedical events from the literature”, *BioMed Central Bioinformatics*, Out 2015, pp. 1–15.

[MAM08] Mansouri, A.; Affendey, L. S.; Mamat, A. “Named entity recognition approaches”, *International Journal of Computer Science and Network Security*, vol. 8, Fev 2008, pp. 339–344.

[Mar00] Martins-Neto, M. A. “Tectonics and sedimentation in a paleo/mesoproterozoic rift-sag basin (espinhaço basin, southeastern brazil)”, *Precambrian Research*, vol. 103, Out 2000, pp. 147–173.

[Mar05] Martins-Neto, M. “A bacia do São Francisco: Arcabouços estratigráfico e estrutural com base na interpretação de dados de superfície e subsuperfície”, *Simpósio sobre o Cráton do São Francisco*, vol. 3, Mar 2005, pp. 283–286.

[Mar16] Martins-Neto, M. A. “Classificação de bacias sedimentares: uma revisão comentada”, *Revista Brasileira de Geociências*, vol. 36–1, Mar 2016, pp. 165–176.

[MBMB01] McCray, A. T.; Bodenreider, O.; Malley, J. D.; Browne, A. C. “Evaluating UMLS strings for natural language processing”. In: Proceedings of the American Medical Informatics Association Symposium, 2001, pp. 448–452.

[ME15] Majumder, A.; Ekbal, A. “Event extraction from biomedical text using CRF and genetic algorithm”. In: Proceedings of the 3^{rd} International Conference on Computer, Communication, Control and Information Technology, 2015, pp. 1–7.

[MMDR04] Maraschin, A. J.; Mizusaki, A. M. P.; De Ros, L. F. “Near-surface k-feldspar precipitation in cretaceous sandstones from the potiguar basin, northeastern brazil”, *The Journal of geology*, vol. 112, Mai 2004, pp. 317–334.

[MRB+07] Milani, E. J.; Rangel, H. D.; Bueno, G. V.; Stica, J. M.; Winter, W. R.; Caixeta, J. M.; Neto, O. P. “Bacias sedimentares brasileiras: cartas estratigráficas”, *Boletim de Geociências da Petrobrás*, vol. 15, Jan 2007, pp. 183–205.

[MUSC+13] Marrero, M.; Urbano, J.; Sánchez-Cuadrado, S.; Morato, J.; Gómez-Berbís, J. M. “Named entity recognition: fallacies, challenges and opportunities”, *Computer Standards & Interfaces*, vol. 35, Set 2013, pp. 482–489.

[NNB09] Nédellec, C.; Nazarenko, A.; Bossy, R. “Information extraction”. Springer, 2009, cap. 31, pp. 663–685.

[NS07] Nadeau, D.; Sekine, S. “A survey of named entity recognition and classification”, *Lingvisticae Investigationes*, vol. 30, Jan 2007, pp. 3–26.

[OOG16] Ogg, J. G.; Ogg, G.; Gradstein, F. M. “A Concise Geologic Time Scale: 2016”. Amsterdam, NED: Morgan Kaufmann Publishers, 2016, 240p.

[OTK02] Ohta, T.; Tateisi, Y.; Kim, J.-D. “The GENIA Corpus: An Annotated Research Abstract Corpus in Molecular Biology Domain”. In: Proceedings of the 2^{nd} International Conference on Human Language Technology Research, 2002, pp. 82–86.

[PCA+86] Petri, S.; Coimbra, A. M.; Amaral, G.; Ojeda, H. O. Y.; Fúlfaro, V. J.; Ponçano, W. L. “Código brasileiro de nomenclatura estratigráfica”, *Revista Brasileira de Geociências*, vol. 16, Dez 1986, pp. 372–376.

[Qui92] Quinlan, J. R. “C4.5: Programming for Machine Learning”. Califórnia, EUA: Morgan Kauffmann, 1992, 302p.

[San09] Santos, D. “Caminhos percorridos no mapa da portuguesificação: A linguateca em perspectiva”, *Linguamática*, Mai 2009, pp. 25–59.

[SC07] Santos, D.; Cardoso, N. “Reconhecimento de entidades mencionadas em português: Documentação e atas do HAREM, a primeira avaliação conjunta na área”. Lisboa, PT: Linguateca, 2007, 413p.

[Sch94] Schmid, H. “Probabilistic part-of-speech tagging using decision trees”. In: Proceedings of International Conference on New Methods in Language Processing, 1994, pp. 1–9.

[Set04] Settles, B. “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, 2004, pp. 104–107.

[SLZ13] Sheppard, S.; Lawson, N. D.; Zhu, L. J. “Accurate identification of polyadenylation sites from 3’ end deep sequencing using a Naïve Bayes classifier”, *BioMed Central Bioinformatics*, vol. 29, Ago 2013, pp. 2564–2571.

[SM09] Sutton, C.; McCallum, A. “Piecewise training for structured prediction”, *Journal Machine Learning*, vol. 77, Dez 2009, pp. 165–194.

[SMG10] Sobhana, N.; Mitra, P.; Ghosh, S. “Conditional random field based named entity recognition in geological text”, *International Journal of Computer Applications*, vol. 1, Fev 2010, pp. 143–147.

[Sob12] Sobhana, N. “Enhancing retrieval of geological text using named entity disambiguation”, *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, Jan 2012, pp. 2250–2459.

[Sux08] Suxiang, Z. “Based Cascaded Conditional Random Fields Model for Chinese Named Entity Recognition”. In: Proceedings of the 9^{th} International Conference on Signal Processing, 2008, pp. 1573–1577.

[SZZ+03] Shen, D.; Zhang, J.; Zhou, G.; Su, J.; Tan, C.-L. “Effective Adaptation of a Hidden Markov Model-Based Named Entity Recognizer for Biomedical Domain”. In: Proceedings of the Workshop on Natural Language Processing in Biomedicine, 2003, pp. 49–56.

[TKSDM03] Tjong Kim Sang, E. F.; De Meulder, F. “Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition”. In: Proceedings of the

7^{th} Conference on Natural Language Learning at North American Chapter of the Association for Computational Linguistics Human Language Technologies, 2003, pp. 142–147.

[TXT+05] Tanabe, L.; Xie, N.; Thom, L. H.; Matten, W.; Wilbur, W. J. "GENETAG: a tagged corpus for gene/protein named entity recognition", *BioMed Central Bioinformatics*, vol. 6, Ago 2005, pp. 1–7.

[UGS] UGS, C. I. d. E. "Tabela Cronoestratigráfica Internacional". Capturado em: http://www.stratigraphy.org/ICSchart/ChronostratChart2017-02BRPortuguese.pdf, Ago 2017.

[VG05] Viera, A. J.; Garrett, J. M. "Understanding interobserver agreement: the kappa statistic", *Society of Teachers of Family Medicine*, vol. 37–5, Maio 2005, pp. 360–363.

[VWPM+88] Van Wagoner, J.; Posamentier, H.; Mitchum, R.; Vail, P.; Sarg, J.; Loutit, T.; Hardenbol, J. "An overview of the fundamentals of sequence stratigraphy and key definitions", *Society of Economic Paleontologists and Mineralogists*, vol. 42, Jan 1988, pp. 39–45.

[WFHP16] Witten, I. H.; Frank, E.; Hall, M. A.; Pal, C. J. "Data Mining: Practical Machine Learning Tools and Techniques". Califórnia, EUA: Morgan Kaufmann, 2016, 664p.

[WKQ+08] Wu, X.; Kumar, V.; Quinlan, J. R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G. J.; Ng, A.; Liu, B.; Philip, S. Y.; et al.. "Top 10 algorithms in data mining", *Knowledge and information systems*, vol. 14, Jan 2008, pp. 1–37.

[Zac12] Zaccara, R. C. C. "Anotação e classificação automática de entidades nomeadas em notícias esportivas em português brasileiro", Dissertação de Mestrado, Universidade de São Paulo, 2012, 72p.

[ZD15] Zirikly, A.; Diab, M. T. "Named Entity Recognition for Arabic Social Media". In: Proceedings of the 1^{st} Conference of the North American Chapter of the Association for Computational Linguistics– Human Language Technologies, 2015, pp. 176–185.

APÊNDICE A – GUIDELINES SOBRE ANOTAÇÃO DO RECONHECIMENTO DE ENTIDADES NOMEADAS GEOLÓGICAS EM TEXTOS DA LÍNGUA PORTUGUESA

1. Introdução

Este documento contém instruções para a anotação das Entidades Nomeadas (EN) em textos da Língua Portuguesa na área da Geologia. De forma resumida, uma contextualização é fundamental no apoio a essas anotações, na qual cita-se: a definição do Reconhecimento de Entidades Nomeadas (REN), as classes que farão parte da tarefa de REN bem como a ferramenta que auxiliará na anotação dessas EN.

2. Reconhecimento de Entidades Nomeadas (REN)

O REN consiste na identificação e na classificação de expressões linguísticas, as EN, as quais são na sua maioria nomes próprios, que remetem a um referente específico no texto [1]. Ele pode abranger os mais variados domínios, como Jornalismo, Geologia, Biologia e Medicina. Na Geologia, a classificação pode ser feita com o uso de classes como: Bacia Sedimentar, Era, Período, Rocha Sedimentar Orgânica, onde as EN são, por exemplo, Bacia do Paraná, Cenozoico, Triássico, Carvão, etc.

Como definimos trabalhar no domínio da Geologia, a EN de nosso interesse refere-se a Entidades Geológicas (EG), que consistem em termos específicos no texto, desde que esses façam parte de uma subárea geológica. Então, para fins de delimitação do escopo deste trabalho, a subárea definida é Bacia Sedimentar Brasileira, pois a quantidade de EG como um todo, no referido domínio, é demasiado ampla e para a obtenção de resultados mais específicos na sua avaliação.

REN é uma tarefa complexa que envolve vários desafios no processo de reconhecimento das EG e divide-se em duas etapas: identificar e classificar as EG. A segunda etapa é mais complexa que a primeira devido a ambiguidade das palavras. Significa que à mesma EG pode ser atribuída mais de uma classificação, dependendo do contexto e do domínio na qual ela está inserida, como: Nome, Rio, Local e Bacia Sedimentar. No exemplo: “O rio São Francisco faz parte da bacia São Francisco”, o primeiro São Francisco é classificado como Rio e o segundo como Bacia Sedimentar.

3. Processo de Anotação

Essa tarefa constitui da anotação automática de um corpus (conjunto de textos) formado por teses, dissertações e Boletins de Geociências da Petrobras. O domínio em questão é Bacia Sedimentar Brasileira, o qual compreende as seguintes classes geológicas: Eon, Era, Período, Época, Idade, Rocha Sedimentar Siliciclástica, Rocha Sedimentar

Carbonática, Rocha Sedimentar Química, Rocha Sedimentar Orgânica, Bacia Sedimentar, Contexto Geológico de Bacia, Unidade Estratigráfica e Outro. Segue uma breve descrição das classes geológicas [2].

1. Eon: Maior subdivisão de tempo dentro da escala de tempo geológico, representadas por Hadeano, Arqueano ou Arcaico, Proterozoico e Fanerozoico.
Ex.: Litologicamente, é representado por rochas graníticas e gnáissicas, com núcleos granulíticos e charnoquíticos, arqueanos a proterozoicos.

2. Era: Corresponde a subdivisão de Eon. São eras: Cenozoico, Mesozoico, Paleozoico.
Ex.: Este complexo de rochas vulcânicas de maior densidade modificou a dinâmica deposicional dos sedimentos Cenozoicos.

3. Período: É a subdivisão de uma Era. São eles: Quaternário, Neogênico, Paleogênico, Cretácico, Jurássico, Triássico, Pérmico, Carbônico (Mississípico e Pensilvânico), Devônico, Silúrico, Ordovícico, Câmbrico.
Ex.: ao redor de 180 Ma (Jurássico): diques e derrames de composição toleítica.

4. Época: Subdivisão do período no tempo geológico. Alguns exemplos: Holocênico, Pleistocênico, Pliocênico, Miocênico, Oligocênico, Eocênico, Paleocênico, Cretácico Superior, Cretácico Inferior, Jurássico Superior, Jurássico Médio, Jurássico Inferior, entre outras.
Ex.: Durante o Oligoceno , a deformação é pequena quando comparada aos outros períodos de deformação.

5. Idade: Subdivisão de Época. Alguns exemplos: Pleistocênico Superior, Pleistocênico Médio, Calabriano, Gelasiano, entre outras.
Ex.: maior incidência entre 80 Ma 1 e 90 Ma (Santoniano/Turoniano): – predominam intrusões de composição básica a intermediária.

6. Rocha Sedimentar Siliciclástica: Originam-se de fragmentos de rochas ígneas, metamórficas ou sedimentares. Alguns exemplos: arenito, argilito, lutito, siltito, conglomerado, ritmito, tilito, folhelho, diamictito, varvito, etc.
Ex. Datações pelo método K - Ar em ilitas diagenéticas dos arenitos da Formação Juruá, indicam que o preenchimento dos reservatórios por petróleo ocorreu no Neotriássico.

7. Rocha Sedimentar Carbonática: originam-se de sedimentos resultantes do depósito de materiais dissolvidos em águas. Alguns exemplos: calcário, dolomito, coquina etc.
Ex.: Provavelmente a mineralidade sentida no vinho seja resultado de uma série de combinações do solo argiloso e calcário Kimmeridgiano.

8. Rocha Sedimentar Química: originam-se de sedimentos que contenham mais de 50% de ferro, como evaporitos.

9. Rocha Sedimentar Orgânica: Originam-se dos restos e secreções dos organismos vivos. Exemplo: carvão, etc.
Ex.: Apenas recentemente ocorreu alguma recuperação, com a elevação dos preços e o maior consumo de Carvão no complexo termoelétrico de Tubarão-SC.

10. Bacia Sedimentar: São grandes áreas de sedimentação, ou seja, deposição de sedimentos (agregados de matéria orgânica e/ou mineral), formada por rochas sedimentares. Sua formação foi a partir do Paleozóico. (São elas: Bacia do São Francisco, Bacia do Espírito Santo, Bacia de Campo, Bacia do Paraná, entre outras).
Ex.: A Guerra de 1989 estudou a influência da sobrecarga do Banco Vulcânico de Abrolhos sobre a estruturação halocinética da Bacia do Espírito Santo.

11. Contexto Geológico de Bacia: classificação relacionada aos eventos geológicos espacial e temporal, ou seja, são os estágios relacionados a Tectônica, Sedimentação e Magmatismo. Ex.: Rifte, Margem Passiva ou Drifte e Intracratônica.

12. Unidade Litoestratigráfica: compreende três componentes estratigráficos: Formação, Grupo e Membro. A Formação Estratigráfica consiste no conjunto de camadas de rochas que possui as mesmas propriedades físicas, podendo conter a mesma associação de fósseis. Algumas formações são constituídas por um único tipo de rocha, como o calcário. Exemplos: Formação Irati, Formação Abrolhos, Formação Lagoa Feia, Formação Coqueiros, Formação Pendência, etc.
Ex.: A Formação Abrolhos de idade cenozoica é caracterizada por uma associação litológica complexa que engloba rochas básicas.

O segundo componente é constituído por duas ou mais formações contíguas associadas que tenham em comum, propriedades litológicas distintas e diagnósticas. Alguns exemplos: Javari, Tapajós, Curuá, entre outros.

Já o terceiro componente, Membro Estratigráfico representa a subdivisão litológica de uma formação. Ex.: Arari, Fazendinha, Ururiá, etc.

13. Outro: São as EG que não se enquadram em nenhuma das classes anteriores, mas pertencem à subárea Bacia Sedimentar Brasileira. Ela é uma classe utilizada apenas para os casos em que o especialista achar muito relevante anotá-la, pois o foco está nas classes definidas anteriormente.

4. IdENGEO

O IdENGEO é Sistema de Marcação de Entidades Nomeadas Geológicas, que auxilia os anotadores na classificação das ENs de forma automática. Ele possui: a área para edição dos textos, a função Filtro, formada pelos botões “Desmarcar Tudo”, “Marcar Tudo” e o menu colorido correspondendo às doze classes geológicas, além dos botões: “Carregar Texto”, “Adicionar Marcação”, “Salvar” e "Remover Marcação". O Filtro serve para visualizar apenas um tipo de classificação ou mais de um, conforme a necessidade da informação que se deseja extrair.

Existem duas tarefas que devem ser observadas nos textos:

A) Verificar a classificação das EG já identificadas, corrigindo-as caso necessário e

B) classificar as EG, as quais não foram atribuídas uma classe.

Dessa forma, tais tarefas são realizadas com a ajuda do IdENGeo, que funciona de acordo com as instruções seguintes.

1. Acessar o link http://www.inf.pucrs.br/linatural/idengeo2/marcacao.html

2. Após o acesso ao link acima, aparecerá a seguinte janela (Figura A.1):

Figura A.1 – Tela inicial do sistema de antoação de EG: IdENGeo

3. Clicar na caixa de texto e selecionar o texto desejado (Figura A.2).

4. Para classificar as EG que não receberam uma classe ou expandir a classificação de uma entidade, deve-se:

4.1 Selecionar o trecho de texto que se deseja transformar em entidade. Na imagem a seguir, foi selecionada a EN Neoproterozoico como exemplo (Figura A.3).

4.2. Clicar no botão “Adicionar Marcação” (Figura A.4).

4.3. Selecionar uma classe no menu, que se localiza abaixo das classes coloridas, à direita da tela, conforme a próxima figura (Figura A.5).

5. Salienta-se que para selecionar uma classe que pertence aos grupos Tempo Geológico e Rochas Sedimentares, deve-se clicar num desses dois grupos para que seja expandido um submenu com as classes pertencentes a cada um desses grupos. Neste exemplo, a classe escolhida pertence ao grupo Tempo Geológico. Então, clica-se em “Tempo Geológico” para depois clicar na classe pertencente a esse grupo (Figura A.6).

6. No momento em que a classificação estiver concluída, clique no botão “OK” (Figura A.7).

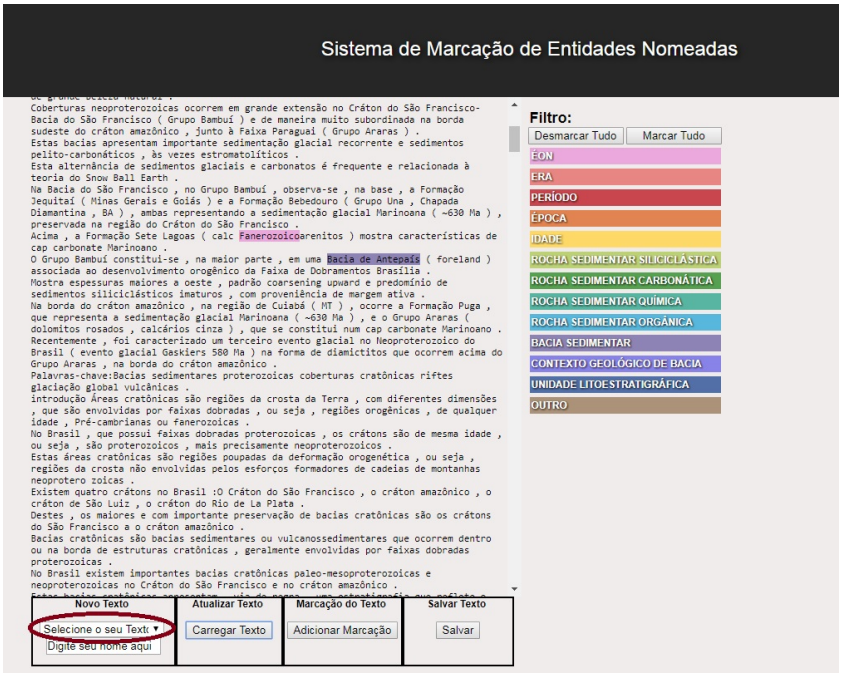


Figura A.2 – Seleção do texto que será classificado

7. Digite o seu nome na caixa de texto, a qual apresenta a mensagem “Digite seu nome aqui” (Figura A.8).

8. Concluídas as alterações no arquivo de texto, clique no botão “Salvar” para confirmar os procedimentos realizados (Figura A.9).

O mesmo arquivo de texto XML pode ser carregado posteriormente para seguir com as marcações. Caso o arquivo não seja salvo, as alterações feitas serão perdidas ao fechar o sistema. Portanto, aconselha-se salvar com frequência essas tarefas, para evitar perda do trabalho.

Observações:

1a) Caso a classificação envolva o grupo “Rochas Sedimentares”, o menu estendido apresentar-se-á de acordo com a Figura A.10.

2a) Para remover a classe de uma palavra, deve-se: a) Clicar na palavra; b) Clicar no botão “Remover Marcação”; c) Clicar no botão “OK” e d) Clicar no botão “Salvar”; Seguem as imagens da sequência das ações acima. a) Neste caso, o exemplo é remover a classificação de “Amazônico” (Figura A.11).

b) Clicar no botão “Remover Marcação” (Figura A.12).

c) Clicar no botão “Ok” (Figura A.13).

d) Clicar no botão “Salvar” (Figura A.14).

3a) Caso precise alterar a classe de uma EG deve-se:

a) Clicar na palavra que receberá a nova classe;

b) Clicar na nova classe escolhida no menu abaixo do filtro colorido e

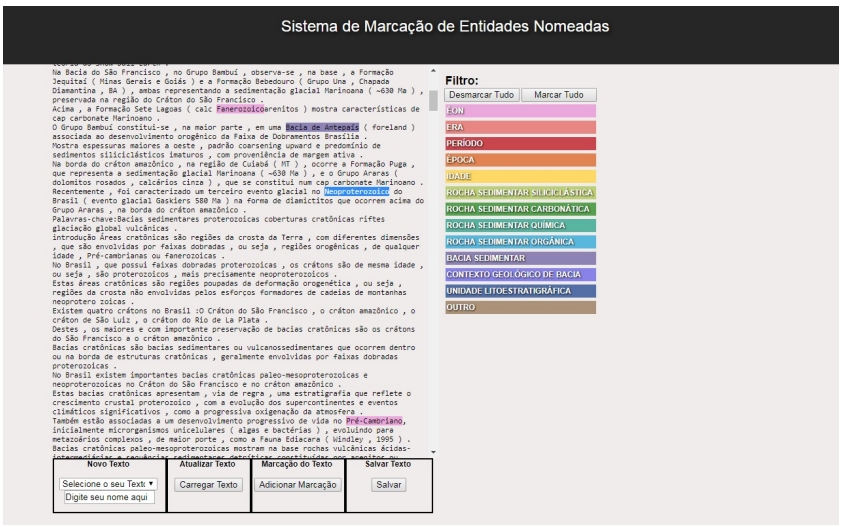


Figura A.3 – Seleção da EG a ser classificada

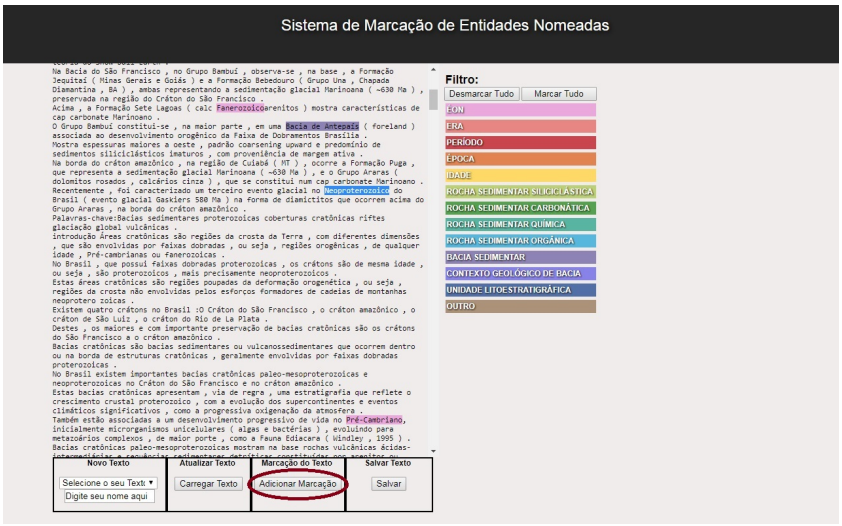


Figura A.4 – Adicionar Marcação da EG

c) Clicar no botão “OK”.

Ao término de todas as alterações feitas, clique no botão “Salvar”.

4a) Se os textos estiverem no computador local e não no servidor, clicar no borão “Carregar Texto”. (Figura A.15).

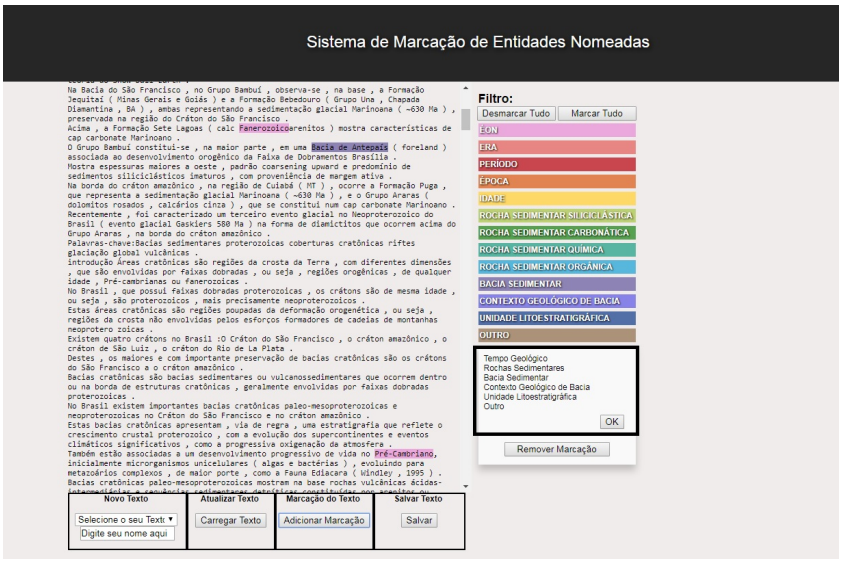


Figura A.5 – Seleção da classe geológica

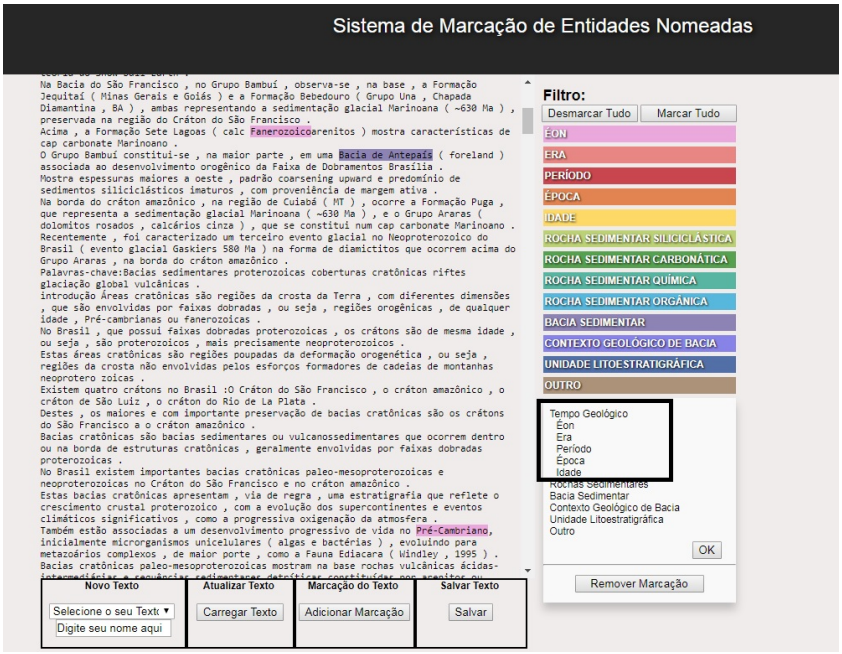


Figura A.6 – Seleção de uma classe do grupo Tempo Geológico

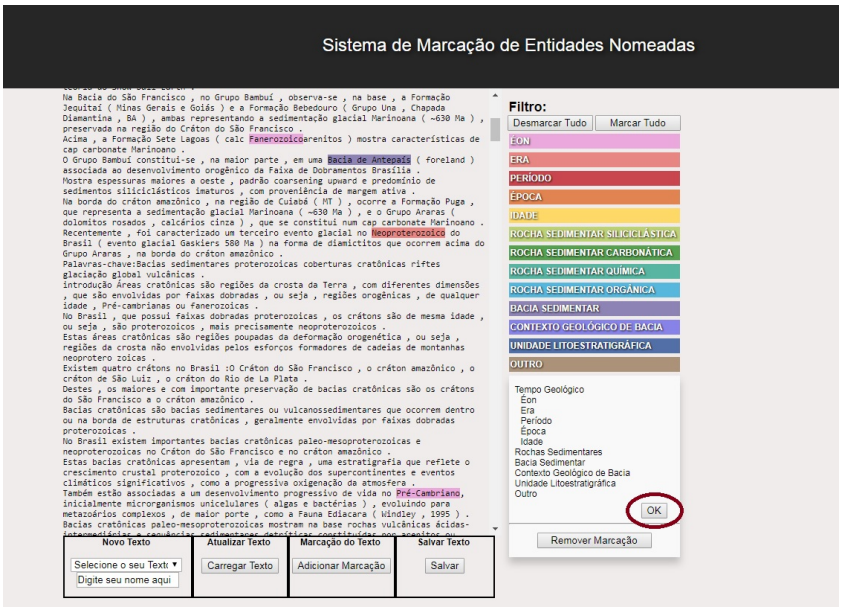


Figura A.7 – Confirmação da classe geológica através do botão “Ok”

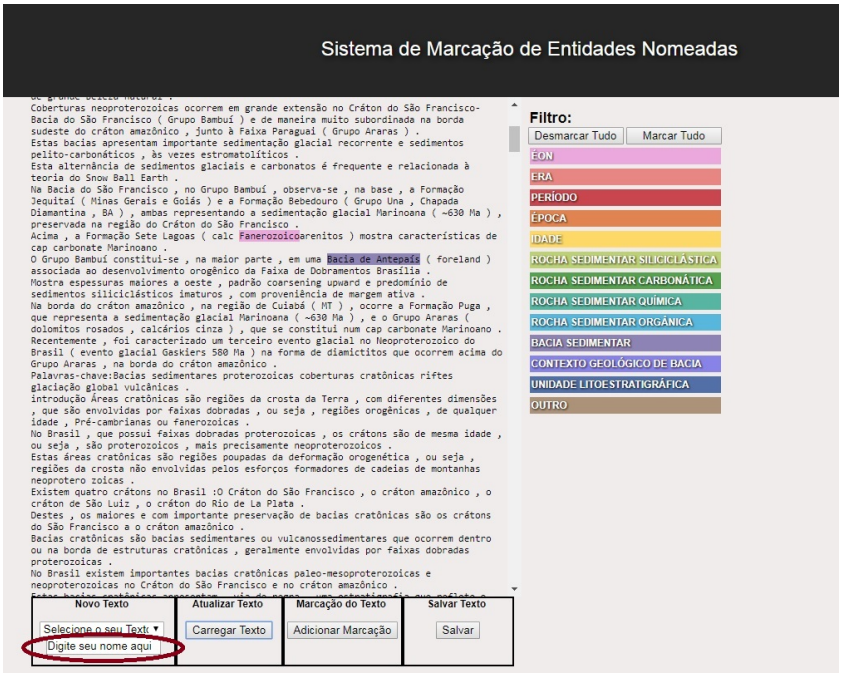


Figura A.8 – Inserção do nome do especialista da classificação

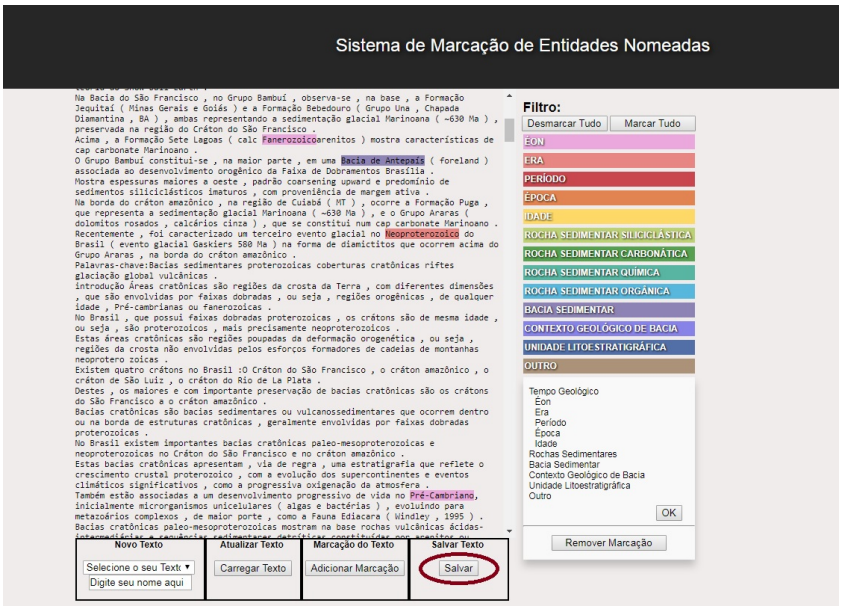


Figura A.9 – Salvar o texto classificado

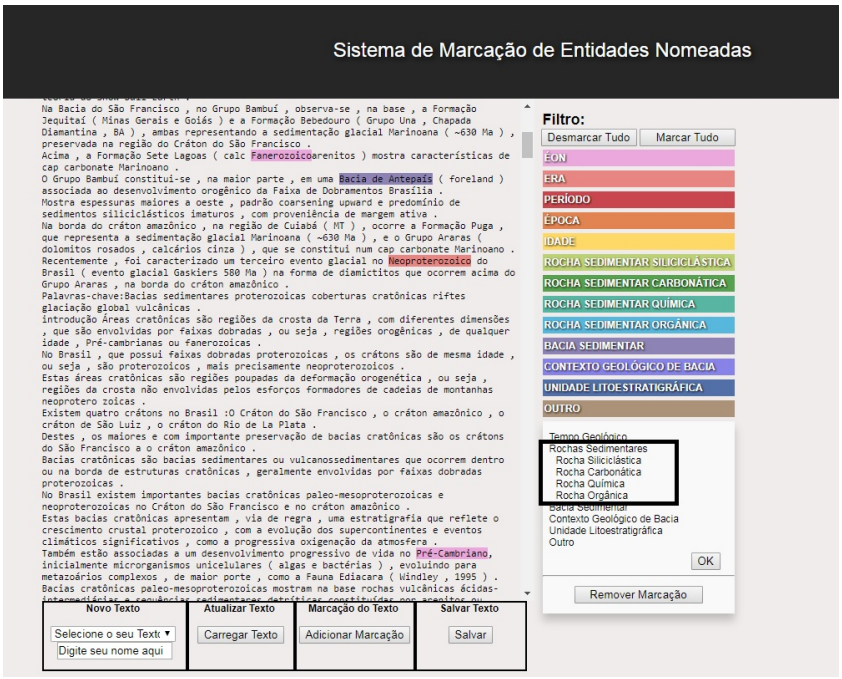


Figura A.10 – Classes do grupo Rochas Sedimentares

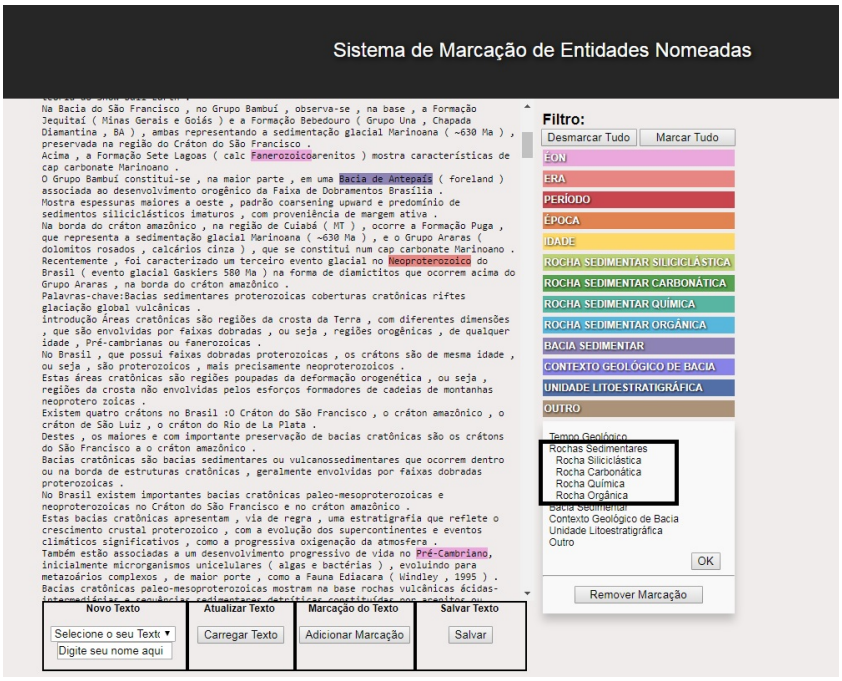


Figura A.11 – Clique na palavra para remover a sua classe

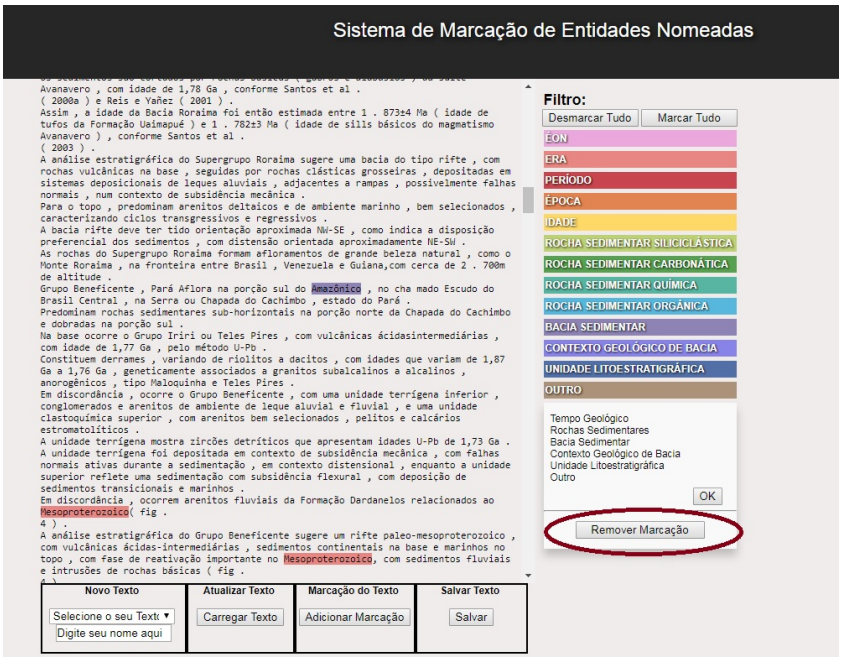


Figura A.12 – Remover classe

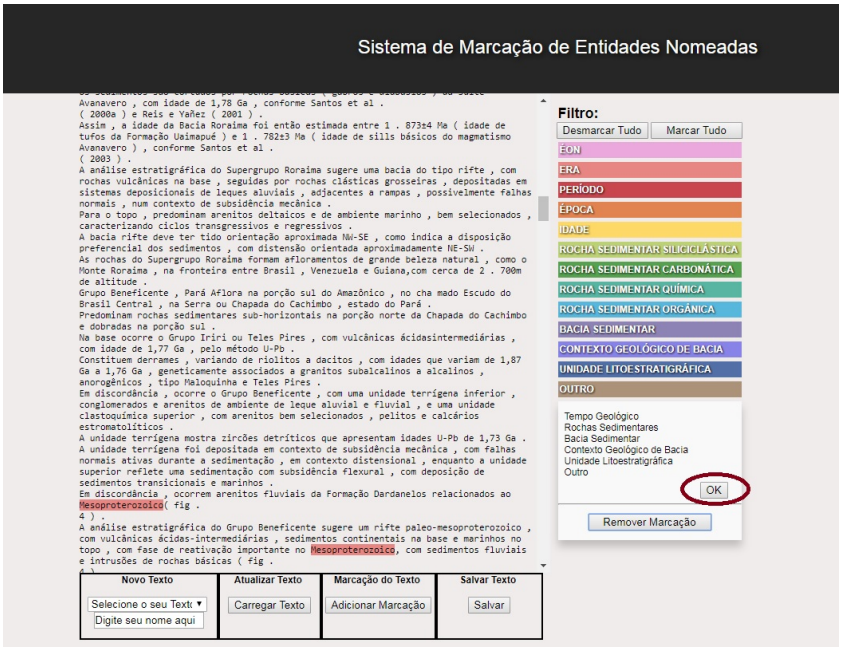


Figura A.13 – Confirma ação de remover classificação

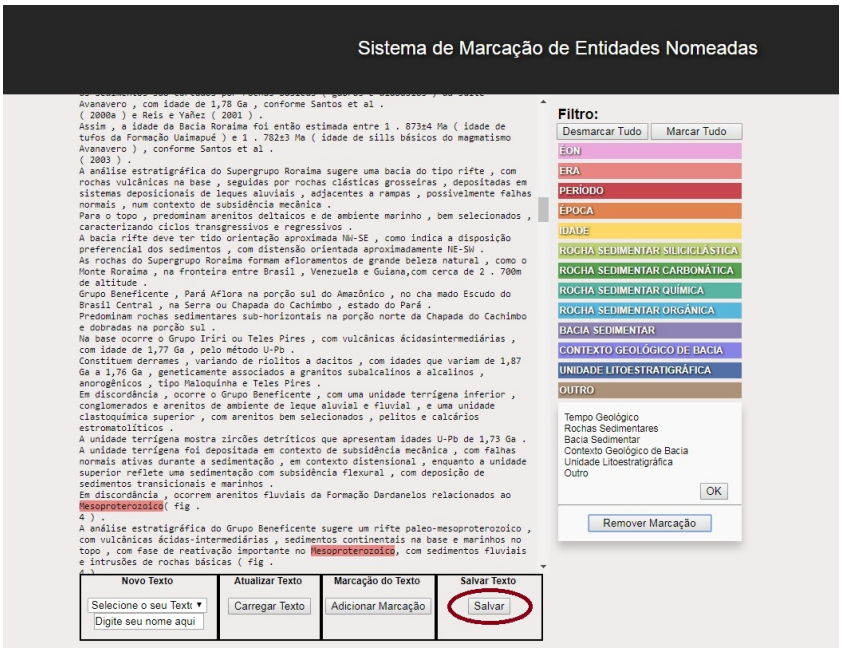


Figura A.14 – Salva ação de remover classificação

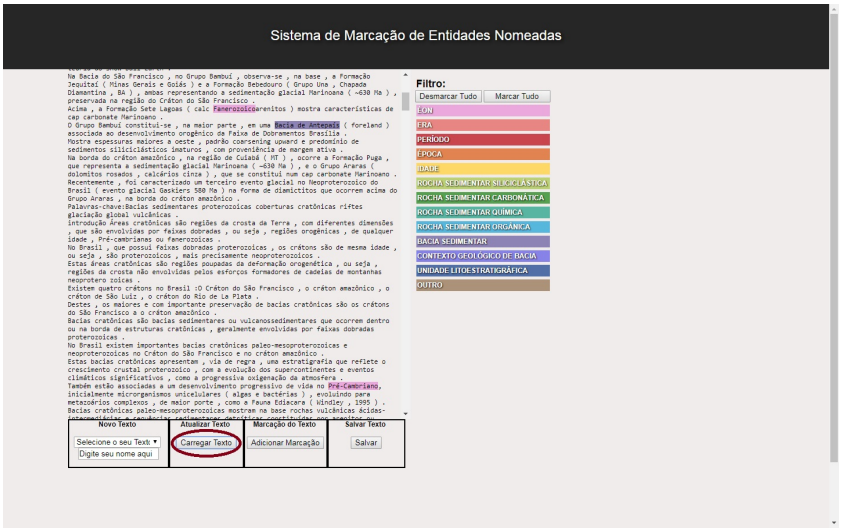


Figura A.15 – Acão de carregar o texto, caso ele esteja salvo localmente

APÊNDICE B – VETOR DE FEATURES

O

’prev2Cap’: ’null’, ’prev2T’: ’null’, ’prev2W’: ’null’, ’prevCap’: ’null’, ’prevW’: ’null’, ’prevT’: ’null’, ’nextCap’: ’maxmin’, ’cap’: ’max’, ’word’: ’O’, ’next2W’: ’constitui’, ’next2T’: ’v-fin’, ’tag’: ’art’, ’nextT’: ’prop’, ’simb’: ’alfa’, ’nextW’: ’Lopingiano’, ’next2Cap’: ’min’, ’ini’: ’max’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

Lopingiano

’prev2Cap’: ’null’, ’prev2T’: ’null’, ’prev2W’: ’null’,’nextCap’: ’min’, ’cap’: ’maxmin’, ’word’: ’Lopingiano’, ’prevCap’: ’max’, ’next2W’: ’a’, ’next2T’: ’art’, ’prevT’: ’art’, ’prevW’: ’O’, ’nextT’: ’v-fin’, ’simb’: ’alfa’, ’nextW’: ’constitui’, ’next2Cap’: ’min’, ’tag’: ’prop’, ’ini’: ’max’, ’pref’: ’false’, ’suf’: ’true’, ’gaz’: ’true’

constitui

’nextW’: ’a’, ’nextCap’: ’min’, ’cap’: ’min’, ’word’: ’constitui’, ’prevCap’: ’maxmin’, ’next2W’: ’subdivisão’, ’next2T’: ’n’, ’prevT’: ’prop’, ’prevW’: ’Lopingiano’, ’nextT’: ’art’, ’simb’: ’alfa’, ’tag’: ’v-fin’, ’next2Cap’: ’min’, ’ini’: ’min’, ’prev2Cap’: ’max’, ’prev2W’: ’O’, ’prev2T’: ’art’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

a

’prevT’: ’null’, ’nextW’: ’subdivisão’, ’nextCap’: ’min’, ’cap’: ’min’, ’word’: ’a’, ’prev-Cap’: ’min’, ’next2W’: ’posterior’, ’next2T’: ’adj’, ’prevT’: ’v-fin’, ’prevW’: ’constitui’, ’nextT’: ’n’, ’simb’: ’alfa’, ’tag’: ’art’, ’next2Cap’: ’min’, ’ini’: ’min’, ’prev2Cap’: ’maxmin’, ’prev2W’: ’Lopingiano’, ’prev2T’: ’prop’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

subdivisão

’nextW’: ’posterior’, ’nextCap’: ’min’, ’cap’: ’min’, ’word’: ’subdivisão’, ’prevCap’: ’min’, ’next2W’: ’do’, ’next2T’: ’v-pcp’, ’prevT’: ’art’, ’prevW’: ’a’, ’nextT’: ’adj’, ’simb’: ’alfa’, ’tag’: ’n’, ’next2Cap’: ’min’, ’ini’: ’min’, ’prev2Cap’: ’min’, ’prev2W’: ’constitui’, ’prev2T’: ’v-fin’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

posterior

’nextW’: ’do’, ’nextCap’: ’min’, ’cap’: ’min’, ’word’: ’posterior’, ’prevCap’: ’min’, ’next2W’: ’Permiano’, ’next2T’: ’prop’, ’prevT’: ’n’, ’prevW’: ’subdivisão’, ’nextT’: ’v-pcp’, ’simb’: ’alfa’, ’tag’: ’adj’, ’next2Cap’: ’maxmin’, ’ini’: ’min’, ’prev2Cap’: ’min’, ’prev2W’: ’a’, ’prev2T’: ’art’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

do

’nextW’: ’Permiano’, ’nextCap’: ’maxmin’, ’cap’: ’min’, ’word’: ’do’, ’prevCap’: ’min’, ’next2W’: ’null’, ’next2T’: ’null’, ’prevT’: ’adj’, ’prevW’: ’posterior’, ’nextT’: ’prop’, ’simb’: ’alfa’, ’tag’: ’vpcp’, ’ini’: ’min’, ’prev2Cap’: ’min’, ’prev2W’: ’subdivisão’, ’prev2T’: ’n’, ’pref’: ’false’, ’suf’: ’false’, ’gaz’: ’false’

Permiano

’nextW’: ’null’, ’cap’: ’maxmin’, ’word’: ’Permiano’, ’prevCap’: ’min’, ’next2W’: ’null’, ’next2T’: ’null’, ’prevT’: ’v-pcp’, ’prevW’: ’do’, ’nextT’: ’null’, ’nextCap’: ’null’, ’simb’: ’alfa’, ’tag’: ’prop’, ’prev2Cap’: ’min’, ’prev2W’: ’posterior’, ’prev2T’: ’adj’, ’pref’: ’false’, ’suf’: ’true’, ’gaz’: ’true’

APÊNDICE C – CONFIGURAÇÃO DO ARQUIVO DE INPUT PARA OS CLASSIFICADORES J48 E NAIVE BAYES

Feature Suf: ário
Feature NextCap: min
Feature Next2T: n
Feature Word: calcários
Feature gaz: calcários
Feature PrevCap: min
Feature Next 2W: subordinadamente
Feature Cap: min
Feature PrevT: prp
Feature PrevW: por
Feature NextT: prp
Feature Simb: simb
Feature NextW: e
Feature Ini: min
Feature Pref: null
Feature Prev2Cap: min
Feature Tag: prp
Feature Prev2W: representadas
Feature Next2Cap: min
Feature Prev2T: verbo
Class: Rocha Sedimentar Carbonática

APÊNDICE D – RESULTADO DO J48 DECISION TREE SEM AS FEATURES “WORDS"

Tabela D.1 – J48 Decision Tree sem as Features “Words"

Classes	Precisão	Abrangência	Medida F
EON	94,10%	55,90%	70,20%
ERA	59,40%	65,00%	62,10%
PERÍODO	66,20%	7,20%	13,10%
ÉPOCA	79,90%	54,60%	64,90%
IDADE	59,40%	10,00%	17,20%
ROC SED SILICICLÁTICAS	36,20%	23,60%	28,60%
ROC SED CARBONÁTICAS	41,40%	9,90%	15,90%
ROC SED QUÍMICA	0,00%	0,00%	0,00%
ROC SED ORGÂNICA	0,00%	0,00%	0,00%
BACIA SEDIMENTAR	30,80%	2,50%	4,60%
CONTEXTO GEOLÓGICO DE BACIA	54,90%	24,50%	33,90%
UNIDADE LITOESTRATIGRÁFICA	71,90%	33,40%	45,60%
OUTRO	44,20%	30,60%	36,20%
OUTSIDE	97,10%	99,70%	98,40%

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br