Processo de construção de um corpus anotado com Entidades Geológicas visando REN

Daniela do Amaral[1], **Sandra Collovini**[1], **Anny Figueira**[1], **Renata Vieira**[1], **Marco Gonzalez**[1]

[1]Faculdade de Informática
Pontificia Universidade Católica do Rio Grande do Sul
90619-900 – Porto Alegre – RS – Brasil

{daniela.amaral, sandra.abreu, anny.figueira}@acad.pucrs.br

renata.vieira@pucrs.br, marcoaigonzalez@gmail.com

Abstract. *This article presents the building process of GeoCorpus, developed for the Geology domain, more specifically for the Bacia Sedimentar Brasileira subarea. The annotation is focused on Geological Entities in Portuguese text, and aims at Named Entity Recognition in the proposed domain. A case study validated both the annotation process and a tool which supported the specialists in the identification and classification of Geological Entities.*

Resumo. *Este artigo apresenta o processo de construção do GeoCorpus, desenvolvido para o domínio de Geologia, mais especificamente, para a subárea Bacia Sedimentar Brasileira. A anotação restringe-se às Entidades Geológicas contidas nos textos em Português e visa o Reconhecimento de Entidades Nomeadas no domínio proposto. Um estudo de caso validou o processo de anotação desse corpus e de uma ferramenta que auxiliou os especialistas na identificação e classificação das Entidades Geológicas.*

1. Introdução

Este trabalho apresenta a construção do corpus GeoCorpus, o qual está sendo anotado com Entidades Geológicas (EG) e visa o Reconhecimento de Entidades Nomeadas (REN) no domínio de Geologia. REN consiste na identificação e na classificação de expressões linguísticas, na sua maioria nomes próprios (como pessoa, local ou organização) que remetem para um referente específico [Mota et al. 2007]. Essas expressões são chamadas de Entidades Nomeadas (EN), e podem variar conforme os diferentes domínios, por exemplo, Medicina, Biologia e Geologia. O que constitui um tipo (classe ou categoria) de EN é sua aplicação, ou seja, em Biologia as classes de interesse podem ser os genes, proteínas [Cohen and Demner-Fushman 2014] e as doenças.

Neste trabalho, o domínio em foco é o da Geologia, em que as ENs de interesse são as Entidades Geológicas (EG). As EG, consideradas neste estudo, consistem em termos específicos no texto, desde que esses façam parte das classes definidas de acordo com a subárea Bacia Sedimentar Brasileira. A escolha do domínio geológico deve-se ao fato de que o REN para Geologia é pouco encontrado na literatura. Enquanto REN para Medicina, Biomedicina e Biologia apresenta uma gama bem maior de trabalhos [Zaccara 2012] [Akhondi et al. 2015] [Collier et al. 2014] [Dánger et al. 2014] [Majumder and Ekbal 2015] [Ohta et al. 2002].

Destaca-se que a adequada identificação e classificação de ENs sob domínios específicos como o de Geologia, representa um grande desafio aos pesquisadores de PLN. Em especial, devido à carência de bases textuais nesse domínio, em Português, e de ferramentas automáticas para capturá-las. Logo, o trabalho apresentado aqui descreve o processo de anotação manual de entidades geológicas visando a construção do GeoCorpus para a tarefa de REN.

Este artigo está organizado da seguinte forma. A seção 2 apresenta o estudo do domínio de Geologia. A construção do corpus é descrita no seção 3. Na sequência, o processo de anotação do corpus geológico, bem como a ferramenta de anotação utilizada são detalhados na seção 4. Um estudo de caso é apresentado na seção 5. Por fim, as considerações finais são relatadas na seção 6.

2. Estudo do Domínio

Dentre os domínios de pesquisa estudados para a tarefa de REN, destaca-se o de Geologia devido a carência de trabalhos que envolvam EG, além da falta de ferramentas automatizadas para extrair tais informações, principalmente para textos do Português.

A partir do estudo do domínio de Geologia, verificou-se várias subáreas, como Sedimentologia, Cronoestratigrafia, Petrografia e Estratigrafia. Houve assim a necessidade de delimitar uma subárea de estudo, no caso a subárea Bacia Sedimentar Brasileira, devido a grande quantidade de EG no domínio em questão e para se obter uma avaliação mais especializada na tarefa de REN. As bacias sedimentares são definidas conforme uma concepção geográfica, isto é, uma área caracterizada pelo acúmulo espesso de sedimentos por um grande período de tempo geológico. A "Bacia do São Francisco" é um exemplo didático para essa definição [Martins-Neto 2005]. Na próxima seção as classes geológicas das Bacias Sedimentares Brasileiras consideradas neste trabalho são apresentadas.

2.1. Determinação das Entidades Geológicas e suas Classes

Com base no estudo da subárea Bacia Sedimentar Brasileira, na orientação de geólogos e professores dessa subárea foram definidas as classes geológicas deste trabalho. A seguir são apresentadas as referidas classes juntamente com alguns exemplos de EG de acordo com [Cohen et al. 2013] e [Hallsworth and Knox 1999].

- **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. Exemplo na sentença: "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. Exemplo: "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, Pérmico (Permiano), Carbônico (Carbonífero), Câmbrico, Devônico (Devoniano), Silúrico (Siluriano), Ordovícico (Ordoviciano), Mississípico e Pensilvânico, esses dois últimos, apenas para a América do Norte. Exemplo: "Em torno de 180 Milhões de anos (**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. Exemplo na sentença: "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. Exemplo na sentença: "Maior incidência entre 80 Milhões de anos (Ma) 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. Exemplo na sentença: "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. Exemplo na sentença: "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. Exemplo na sentença: "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. Exemplo na sentença: “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. Bacias Sedimentares Brasileiras: 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. Exemplo na sentença: “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. Por exemplo: Intracratônica ou Sinéclise, Rifte, Drifte e Margem Passiva. Exemplo na sentença: “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 Estratigráfica: compreende três componentes estratigráficos: Formação, Grupo e Membro [ESTRATIGRAFICA-SBG 1986]. A Formação consiste na unidade principal da litoestratigrafia. Uma formação é constituída por um corpo rochoso e pode conter um ou mais tipos de rochas, estruturas sedimentares e fósseis. Já o segundo, o Grupo, é constituído por duas ou mais formações contíguas associadas, que tenham propriedades litológicas distintas e diagnósticas em comum. O terceiro componente, Membro Estratigráfico, representa a subdivisão litológica de uma formação. Ele consiste de uma entidade que possui características litológicas próprias, as quais permitem diferenciá-las das partes adjacente da formação. Exemplos: Formação Irati, Formação Abrolhos, Javari, Tapajós, Curuá, Arari, Fazendinha, etc. Exemplo na sentença: “A bacia do Rio do Peixe tem como substrato rochas sedimentares cretáceas dos **grupos Bauru e Caiuá** e esporádicas e localizadas ocorrências de basaltos da **Formação Serra Geral**.”

13. Outro: Esta é uma classe de exceção, pois o foco está nas classes definidas anteriormente. Deve ser utilizada apenas para os casos em que o especialista achar um termo muito relevante à subárea Bacia Sedimentar Brasileira, mas que não se enquadra exatamente nas classes anteriores. Exemplos: fácies, módulo calcários, organismos fósseis.

3. Construção do Corpus

Para a construção do corpus GeoCorpus realizou-se a leitura de trabalhos científicos para a identificação de EG relacionadas à subárea Bacia Sedimentar Brasileira com

[Cohen et al. 2013] e [Hallsworth and Knox 1999]. 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 [Cohen et al. 2013], nomes de rochas sedimentares [Hallsworth and Knox 1999], nomes de bacias sedimentares brasileiras [Martins-Neto 2005] [Bizzi et al. 2003], os estágios relacionados à Tectônica, Sedimentação e Magmatismo e unidades estratigráficas. Dentre os serviços 'on-line' utilizados para a formação do corpus geológico 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. Por fim, a homogeneidade foi estabelecida, principalmente, para não misturar textos com outros elementos, como imagens, tabelas e gráficos. Como 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.

O corpus é constituído de 52 textos, em que cada documento corresponde a um arquivo de texto com tamanho entre 10 Kbytes e 53 Kbytes (de 1.460 palavras a 7.793 palavras). O processo de anotação é descrito na seção a seguir.

4. Processo de Anotação

Nesta seção é apresentado o processo de anotação das EG contidas nos textos do GeoCorpus, o qual segue as mesmas etapas de REN: identificar as EG e após, classificá-las em uma das classes geológicas descritas na seção 2.1. Destaca-se que a etapa de classificação é mais complexa devido a ambiguidade das palavras, em que uma mesma EG pode ser classificada com mais de uma classe dependendo do contexto e do domínio que está inserida. Por exemplo, na sentença: "O **rio São Francisco** faz parte da **bacia São Francisco**", a primeira EG é classificada como Rio e a segunda como Bacia Sedimentar.

Basicamente, o processo de anotação envolve os seguintes passos:

1º) Marcar os termos que referem-se a uma EG no texto, caso não tenham sido anotados e atribuir uma classe;

2º) Verificar a delimitação da EG (palavras que formam uma EG) já marcada no texto, corrigindo-a caso necessário;

3º) Verificar a classificação da EG já identificada, corrigindo-a caso necessário.

O segundo e o terceiro passos ocorreram, porque o GeoCorpus foi processado num modelo de classificação, desenvolvido para um experimento inicial [Amaral and Vieira 2014]. Esse modelo possui várias classes de Geologia e não se restringiu a uma subárea específica. A anotação dos textos será realizada com o auxílio da

ferramenta IdENGeo descrita a seguir.

4.1. IdENGeo

A IdENGeo é uma ferramenta de marcação de Entidades Nomeadas em textos do domínio de Geologia, a qual 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 à uma ou mais classes de interesse).

- **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 poder 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 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 4. 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 classes seguiu a organização de classes por grupos, conforme apresentado na seção 2.1. Caso o anotador necessite remover a classe escolhida, deve utilizar o botão "Remover Marcação".

5. Estudo de Caso

Nesta seção descrevemos um estudo de caso com o objetivo de validar o processo de anotação e a ferramenta IdENGeo. Realizou-se esse estudo com base na experiência da anotação de um texto do GeoCorpus, o qual corresponde a um arquivo de 5.984 palavras. O texto foi anotado por um geólogo e o tempo total, estimado pelo anotador, foi de noventa minutos. A anotação resultou num total de 549 EG distribuídas nas seguintes classes ilustradas na Tabela 1.

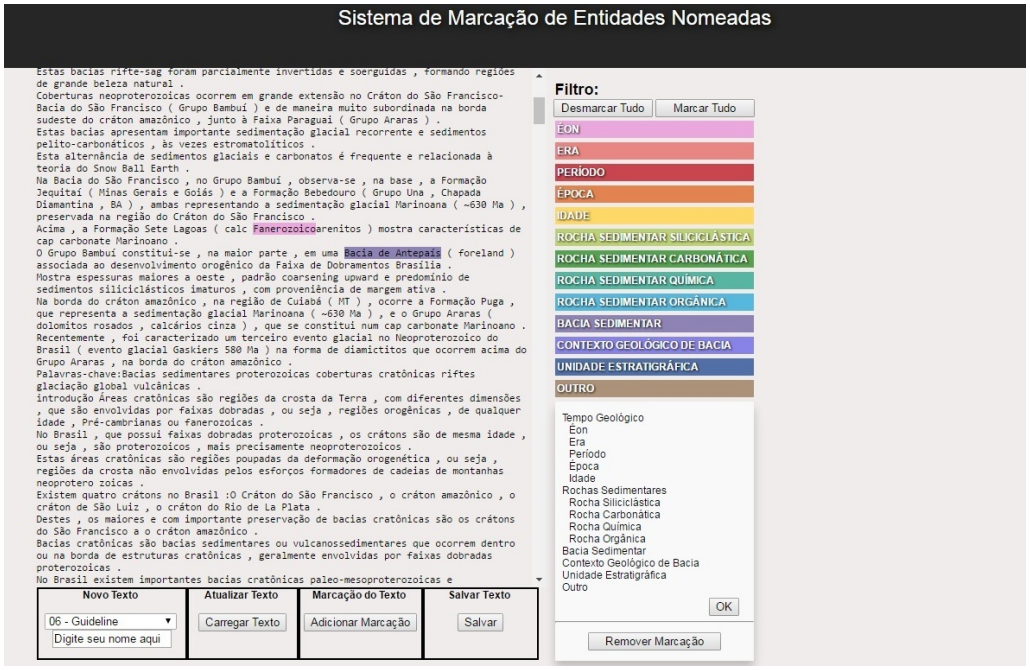

Figura 1. Interface Gráfica do IdENGeo

A partir desses resultados, constatou-se que as classes Rocha Sedimentar Siliciclástica e Unidade Estratigráfica foram as mais frequentes no texto, pois o seu assunto compreende a estratigrafia e a paleogeografia da Formação Brejo Santo, na Bacia do Araripe, com base no estudo de aspectos sedimentológicos, faciológicos e paleontológicos.

Já as classes Rocha Sedimentar Química e Rocha Sedimentar Orgânica não tiveram ocorrência pelo mesmo motivo acima, ou seja, o tema que o texto aborda. Observou-se também que a classe Outro apresentou vários casos, os quais o especialista julgou relevantes para a subárea Bacia Sedimentar Brasileira. Como por exemplo as EG “organismos fósseis”, “ostracodes” e “conchostráceos”.

O anotador notou que um dos pontos de dificuldade é a delimitação das palavras que formam uma EG. 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 Estratigráfica 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 analisada refere-se ao aspecto morfológico das palavras dispostas nos textos, ou seja, a forma em que o termo geológico está inserido na sentença. Significa que, quando uma expressão é constituída por um substantivo seguida de um adjetivo, esse 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.

Alguns anotadores testaram as funcionalidades da ferramenta e constataram que

Tabela 1. Resultado das Entidades Geológicas no estudo de caso.

Classes	Entidades Geológicas
Era	3
Período	20
Época	3
Idade	2
Rocha Sedimentar Siliciclástica	140
Rocha Sedimentar Carbonática	11
Bacias Sedimentares Brasileiras	47
Contexto Geológico de Bacia	41
Unidade Estratigráfica	121
Outro	161
Total	**5.575**

o IdENGeo apresenta uma característica importante que define o bom emprego de um sistema: a usabilidade. Segundo eles, o IdENGeo é de rápido e fácil aprendizado. Adicionalmente, é uma ferramenta que resolve com satisfação as tarefas para as quais ela foi projetada. Destacaram também a importância de selecionar as EG por classe através do filtro, uma vez que com esse recurso, é possível verificar as classes de cada texto.

Para este estudo de caso, o anotador finalizou as suas considerações ao expor que, devido ao tipo de texto ser uma tese, não foi possível realizar a anotação em único momento. Então, para solucionar essa questão, foi inserido no IdENGeo, a nova funcionalidade "Atualizar Texto" (descrita na seção 4.1), com o objetivo de facilitar o trabalho manual.

6. Considerações Finais

Este artigo descreve o processo de anotação manual de EG, a fim de construir o GeoCorpus com o propósito de que, a partir dele, seja realizada a tarefa de REN. Apresentamos a ferramenta de anotação e um experimento inicial sobre a tarefa. O domínio do GeoCorpus é Geologia e Bacia Sedimentar Brasileira é a subárea que o especializa com o objetivo de torná-lo mais eficaz na identificação das EG. A construção do corpus iniciou com a escolha do domínio, a determinação de uma subárea, a decisão das EG e de suas classes e os textos que o compõem. Em síntese, a metodologia que envolveu a sua anotação consistiu da identificação e classificação dos termos considerados como EG, além da conferência de algumas EG já classificadas no texto.

A ferramenta IdENGeo tem por objetivo auxiliar na tarefa de REN, de modo que o trabalho de anotação seja mais simples e eficiente. O processo de anotação e o uso da ferramenta foram analisados com um estudo de caso da anotação de um texto. Essa tarefa resultou num total de 549 EG distribuídas nas classes que semanticamente condisseram com o assunto que o texto abordou. Dois importantes desafios deste trabalho são: primeiro, a grande dificuldade de encontrar anotadores com disponibilidade de classificar os textos para gerar um corpus de referência; segundo, a confiança na anotação, ou seja, conseguir especialistas que tenham conhecimento na subárea definida para anotar os textos.

Como trabalhos futuros, iremos finalizar a anotação completa do corpus. Atualmente, estamos na etapa de geração da anotação que está sendo feita manualmente por geólogos, entre eles professores, doutorandos e alunos de graduação do curso de Geologia da UNISINOS do 6º semestre.

Pretende-se melhorar a ferramenta de anotação com a modificação do acesso da "função filtro" e a visualização do "menu classificação das EG". Ainda, a conclusão do GeoCorpus gerará um recurso que será utilizado em sistemas de aprendizado máquina para o REN Geológicas. O fruto desse corpus é elemento fundamental para uma tese que está em desenvolvimento. Sua disponibilidade é relevante para a pesquisa em Geologia, para tarefas de PLN, como relações entre EG e resolução de correferência. Destaca-se também a relevância econômica com a exploração do petróleo, uma vez que o GeoCorpus compreende um conjunto de textos sobre bacias sedimentares brasileiras. 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.

Referências

Akhondi, S. A., Hettne, K. M., Van Der Horst, E., Van Mulligen, E. M., and Kors, J. A. (2015). Recognition of chemical entities: combining dictionary-based and grammar-based approaches. *Journal Cheminformatics*, 7(S-1):S10.

Amaral, D. O. F. d. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

Bizzi, L. A., Schobbenhaus, C., VIDOTTI, R. M., and GONÇALVES, J. H. (2003). *Geologia, Tectônica e Recursos Minerais do Brasil: texto, mapas e SIG*. CPRM.

Cohen, K. B. and Demner-Fushman, D. (2014). *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company.

Cohen, K. M., Finney, S. C., Gibbard, P. L., and Fan, J.-X. (2013). The ics international chronostratigraphic chart. *Episodes*, 36(3):199–204.

Collier, N., Paster, F., Campus, H., and Tran, A. M.-v. (2014). The impact of near domain transfer on biomedical named entity recognition. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL 2014*, pages 11–20, Gothenburg,Sweden. Association for Computational Linguistics.

Dánger, R., Pla, F., Molina, A., and Rosso, P. (2014). Towards a protein-protein interaction information extraction system: Recognizing named entities. *Knowledge-Based Systems*, 57:104–118.

ESTRATIGRAFICA-SBG, C. E. D. N. (1986). Códìgo brasì lei ro de nomenclatura estratígráfìca. guia de nomenclatura estratigráfica. *Revista Brasileira de Geociências*, 16(4):370–415.

Hallsworth, C. and Knox, R. (1999). Bgs rock classification scheme. volume 3, classification of sediments and sedimentary rocks.

Majumder, A. and Ekbal, A. (2015). Event extraction from biomedical text using crf and genetic algorithm. In *Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, pages 1–7. IEEE.

Martins-Neto, M. (2005). A bacia do são francisco: Arcabouços estratigráfico e estrutural com base na interpretação de dados de superfície e subsuperfície. *SBG, Simp. Craton São Francisco*, 3:283–286.

Mota, C., Santos, D., and Ranchhod, E. (2007). Avaliação de reconhecimento de entidades mencionadas: princípio de arem. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, pages 161–175.

Ohta, T., Tateisi, Y., and Kim, J.-D. (2002). The genia corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 82–86, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zaccara, R. C. C. (2012). *Anotação e classificação automática de entidades nomeadas em notícias esportivas em Português Brasileiro*. PhD thesis, Universidade de São Paulo.