Pontifícia Universidade Católica do Rio Grande do Sul
Laboratório de PLN

Relatório Geocorpus

GeoCorpus

Profa. Renata Vieira
Profa. Silvia Moraes
Bernardo Consoli
Nikolas Lacerda

Porto Alegre, Brasil
09 de abril de 2020

PUCRS

Sumário

1 Introdução

Geocorpus é um corpus de avaliação para Língua Portuguesa que reúne vários trabalhos científicos na área de Geologia. Este corpus, desenvolvido originalmente por Daniela Amaral [1], contém trabalhos cujo tema são as entidades geológicas (EG) relacionadas à subárea Bacia Sedimentar Brasileira. Seus textos são essencialmente teses, dissertações, artigos e boletins de Geociências da Petrobras. Esses textos foram recuperados e selecionados a partir de termos geológicos da tabela Cronoestratigráfica [UGS], que contém nomes de rochas sedimentares [HK99], nomes de bacias sedimentares brasileiras [Mar05;BSVG03], os estágios relacionados à Tectônica, Sedimentação e Magmatismo e unidades estratigráficas.

Neste relatório, vamos apresentar e discutir as modificações que foram feitas nesse corpus visando melhorar a classificação de entidades feita pelo aprendizado de máquina.

2 Modificações

O corpus continha vários problemas que impactaram negativamente nos experimentos realizados. Seguem abaixo as modificações feitas.

2.1 Remoção de categorias vazias

Descrição: Constavam no GeoCorpus algumas categorias com o identificador vazio.

Exemplo: <EM CATEG=" ">quartzo</EM>

Solução: As categorias vazias foram removidas do corpus visando sua padronização. É importante ressaltar que essa mudança não impacta na anotação.

2.2 Remoção de categorias aninhadas

Descrição: Constavam no Geocorpus algumas categorias aninhadas.

Exemplo: <EM CATEG="EstruturaSedimentar"><EM CATEG="baciaSedimentar">Bacia do Paraná </EM></EM>

Solução: Nesses casos foi mantida apenas a categoria correta. O aninhamento que visava especialização foi removido, mantendo-se assim, a categoria mais genérica.

2.3 Correção de anotação de categoria

Descrição: Algumas entidades foram categorizadas em mais de uma classe

Exemplo: <EM CATEG="ERA">Neoarqueano</EM> ...
<EM CATEG="PERIODO">Neoarqueano</EM>

Solução: Nesses casos a anotação foi refeita. Enviamos os casos identificados para um expert e ele atribuiu a classe correta à entidade.

2.4 Remoção de linhas duplicadas

Descrição: Havia algumas linhas repetidas no GeoCorpus.

Exemplo: A linha 766 e 776 eram iguais, constando a mesma frase: "Grãos de silicato de zircônio incrustados em rochas metamórficas do grupo Warrawoona na Austrália ocidental foram datados em até 4 , 4 bilhões de anos , indicando que por essa época uma crosta estava se consolidando."

Solução: Foram removidas do GeoCorpus linhas exatamente iguais, visto que repetições tendem a atrapalhar o aprendizado de máquina. Ao todo, foram removidas 73 linhas, com uma diminuição de 51 entidades que estavam repetidas entre essas linhas.

2.5 Correção de linhas quebradas indevidamente

Descrição: Havia um padrão de quebra de linha indevida no GeoCorpus. Em algumas frases com vírgula ou abertura de parênteses, havia uma nova linha segmentando a frase em duas partes.

Solução: Como nem todas as linhas com parênteses ou vírgulas possuiam a quebra de linha indevida, as frases que apresentavam essa quebra foram corrigidas manualmente, visando a melhora do aprendizado de máquina.

2.6 Remoção da categoria outro

Descrição: Existia uma categoria chamada 'outros' no GeoCorpus, com 737 entidades.

Solução: Todas as entidades dessa categoria foram passadas para um expert, e recategorizadas em categorias mais específicas, para as entidades que não se encaixavam nas categorias já existentes, foram criadas categorias novas indicadas pelo próprio expert. Isso foi necessário pois a classe 'outros' era muito ampla e dificultava o aprendizado do classificador automático. Por isso, optamos pela substituição desta categoria por categorias mais especificas.

2.7 Padronização das categorias

Descrição: Não havia um padrão no nome das categorias do corpus, algumas estavam todas em maiúsculo, outras em minúsculo, e as com palavras compostas alternavam.

Solução: Todas as categorias estão nomeadas com o padrão Camel Case

2.8 Entidades não anotadas

Descrição: Constavam no GeoCorpus 2913 entidades que deveriam ser anotadas e não foram, consideramos essas entidades sendo entidades que foram anotadas uma ou mais vezes no corpus mas em determinadas frases não foram anotadas.

Exemplo:
Frase 1: grânulos subangulosos de <EM CATEG="sedimentaresSiliciclasticas">**quartzo**</EM>.
Frase 2: em adição a outros minerais detríticos como o **quartzo**.
Nesse exemplo, a entidade quartzo aparece categorizada na primeira frase, porém na segunda, quartzo não está categorizado.

Solução: Todas as entidades que deveriam estar categorizadas foram categorizadas através de um script. Foi possível a utilização de um script pois as palavras que não foram categorizadas não tinham problema de contexto e nem duplo significado, que fariam elas aparecerem em um momento categorizadas e outro não.

3 Análise do GeoCorpus

Após modificarmos o corpus na tentativa de obter um melhor resultado fizemos uma análise nas entidades do corpus modificado, juntamente com uma analise nas entidades do corpus não modificado, para fins de comparação. Com isso, queremos observar o nosso ganho, ou perda, no número de entidades e no número de classes do Geocorpus.

No Geocorpus antigo tinhamos uma soma de 6126 entidades anotadas, divididas em 20 classes, com as devidas modificações efetuadas, observa-se o impacto que tivemos na quantidade de categorias e de classes. No Geocorpus modificado, temos um total de 8954, um ganho de 2828 entidades em relação ao corpus anterior, além disso, o corpus possui 30 classes, um aumento de 10 classes em relação ao anterior.

Também foi analisado o número de entidades distintas do GeoCorpus novo, ou seja, o número de entidades diferentes que cada classe possui.

Com essa tabela, podemos notar que dentro das 8954 entidades anotadas, há um total de 1229 entidades distintas, um número bem inferior ao número total de entidades, o que demonstra que há muitas repetições das mesmas entidades, algo que temos que levar em consideração em momentos de aprendizado.

4 Resultados

A primeira versão do GeoCorpus foi usada para a elaboração do artigo 'Embeddings for Named Entity Recognition in Geoscience Portuguese Literature'. Naquele trabalho, o corpus gerou resultados bem satisfatórios quando usado no treinamento de uma rede neural (Table 1).

No entanto, após as modificações e correções que foram feitas no GeoCorpus, reproduzimos o experimento treinando o modelo com a nova versão para tentar atingir melhores resultados.

Nota-se que com a nova versão do Geocorpus, obtivemos um avanço considerável nas métricas do experimento.

Tabela 1 – Comparação GeoCorpus: versão Original x GeoCorpus: versão Revisada

Classe	#Instâncias(Original)	#Instâncias(Revisada)
Tempo		
idade	796	799
eon	288	256
era	326	414
epoca	650	687
periodo	637	714
Rochas		
metamorficas	197	378
magmaticas	222	582
sedimentaresSiliciclasticas	741	1102
sedimentaresCarbonaticas	240	355
sedimentaresQuimicas	5	12
sedimentaresOrganicas	22	22
Constituintes e Propriedades de Rochas		
constituinteRochaSedimentar	0	112
mineral	0	212
fosseis	0	132
estruturaSedimentar	0	86
estruturaGeologica	0	78
Sítio		
contextoGeologicoDeBacia	262	663
ambienteSedimentacao	0	146
bentonico	13	27
planctonico	44	112
campoPetrolifero	0	6
Elementos da Estratigrafia		
baciaSedimentar	243	552
unidadeEstratigrafica	578	764
unidadeGeotectonica	0	28
estratigrafia	0	247
formacao	18	0
Outros		
sistemaPetrolifero	0	93
estruturaDeBacia	40	0
geomorfologia	0	54
granulometria	67	129
elementoQuimico	0	26
procedimentoMetodologico	0	166
outro	737	0
Soma	**6.126**	**8.954**

Tabela 2 – Instâncias Distintas das Entidades do GeoCorpus - Versão Revisada

Classe	#Instâncias Distintas
Tempo	
idade	84
epoca	74
periodo	62
era	47
eon	20
Rochas	
metamorficas	59
magmaticas	58
sedimentaresSiliciclasticas	160
sedimentaresCarbonaticas	77
sedimentaresQuimicas	4
sedimentaresOrganicas	1
Constituintes e Propriedades de Rochas	
constituinteRochaSedimentar	24
mineral	6
fosseis	29
estruturaSedimentar	28
baciaSedimentar	83
estruturaGeologica	19
Sítio	
ambienteSedimentacao	32
contextoGeologicoDeBacia	121
bentonico	4
planctonico	9
campoPetrolifero	2
Elementos da Estratigrafia	
unidadeEstratigrafica	153
unidadeGeotectonica	8
estratigrafia	29
Outros	
geomorfologia	6
elementoQuimico	3
granulometria	13
procedimentoMetodologico	4
sistemaPetrolifero	10
Soma	**1.229**

Tabela 3 – Resultados dos experimentos com o GeoCorpus antigo

Embedding Model		PRE	REC	F1
Word	GeoWE	73.31%	42.38%	53.71%
Embeddings	W2V-SKPG	80.27%	64.18%	71.33%
Flair	FlairBBP	85.97%	80.41%	83.10%
Embeddings	FlairBBP$_{GeoFT}$	86.03%	82.45%	84.20%
	GeoWE+FlairBBP	86.87%	72.16%	78.84%
Stacked	W2V-SKPG+FlairBBP	86.78%	81.47%	84.04%
Embeddings	GeoWE+FlairBBP$_{GeoFT}$	86.35%	81.29%	83.74%
	W2V-SKPG+FlairBBP$_{GeoFT}$	**86.63%**	**82.71%**	**84.63%**

Tabela 4 – Resultados dos experimentos com o GeoCorpus novo

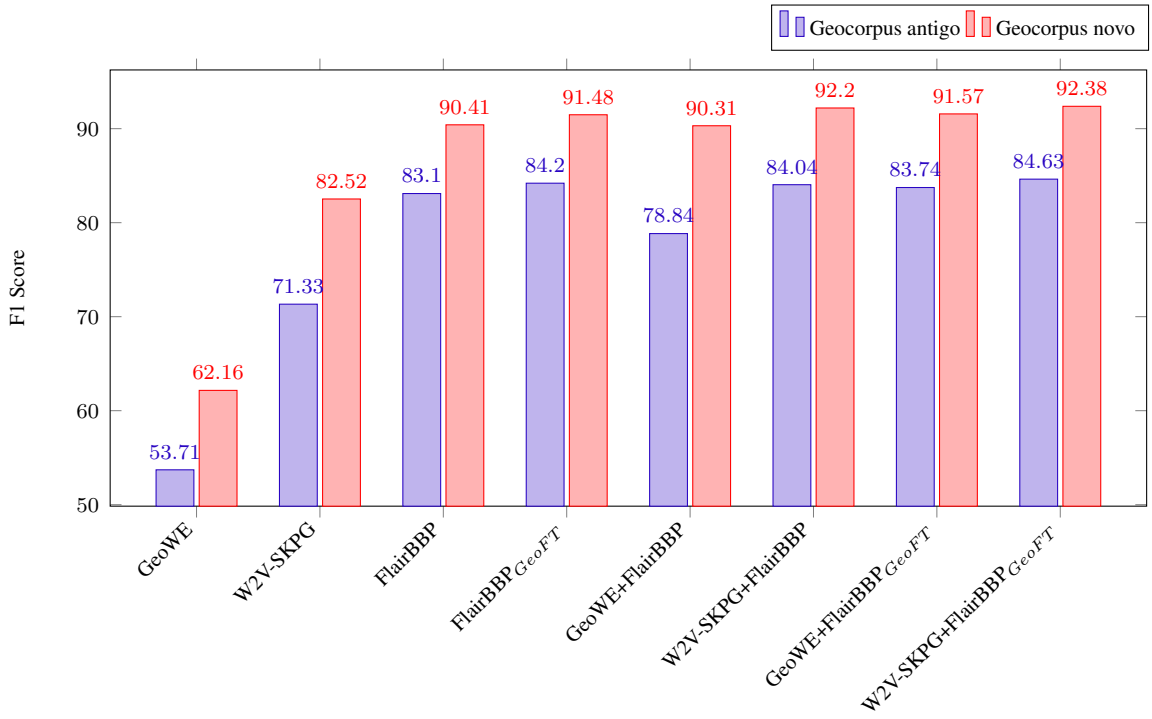
Embedding Model		PRE	REC	F1
Word	GeoWE	73.58%	53.80%	62.16%
Embeddings	W2V-SKPG	84.13%	80.98%	82.52%
Flair	FlairBBP	90.28%	90.55%	90.41%
Embeddings	FlairBBP$_{GeoFT}$	91.35%	91.62%	91.48%
	GeoWE+FlairBBP	90.31%	90.31%	90.31%
Stacked	W2V-SKPG+FlairBBP	92.31%	92.09%	92.20%
Embeddings	GeoWE+FlairBBP$_{GeoFT}$	91.71%	91.44%	91.57%
	W2V-SKPG+FlairBBP$_{GeoFT}$	**92.18%**	**92.57%**	**92.38%**

5 Conclusão

As modificações feitas no GeoCorpus visavam melhorar o aprendizado de máquina. A reprodução do experimento feito anteriormente, porém usando a versão corrigida do corpus, mostrou que essas modificações geraram o efeito esperado. Observamos uma melhora nos resultados. Na melhor configuração encontrada a medida F1 subiu de 84,63% para 92,38%.

As correções, limpezas e novas anotações produziram um aumento em média de 8.68 pontos percentuais nos testes. É importante destacar que o modelo que obteve o melhor desempenho no GeoCorpus antigo, continuou sendo o modelo de melhor desempenho no GeoCorpus novo.

No futuro, podemos identificar novas formas de melhorar o corpus, tais como unir algumas classes com poucos exemplos, e até mesmo adicionar mais informações ao corpus.



Referências

1 AMARAL, D. O. F. *Reconhecimento de entidades nomeadas na Área da geologia: bacias sedimentares brasileiras*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio Grande do Sul, 2017. Disponível em: <http://tede2.pucrs.br/tede2/handle/tede/8035>.