PUCRS

Pontifical Catholic University of Rio Grande do Sul

Artificial Intelligence Research Group

Geocorpus Report

GeoCorpus Change Report

Profa. Renata Vieira

Profa. Silvia Moraes

Bernardo Consoli

Nikolas Lacerda

Porto Alegre, Brazil

April 9, 2020

PUCRS

Contents

1 Introduction

Geocorpus is an evaluation corpus for the Portuguese language that collects several scientific works in the field of Geology. This corpus, originally developed by Daniela Amaral [1], contains works whose theme is geological entities (GE) related to the Brazilian Sedimentary Basin subdomain. The collected texts are essentially theses, dissertations, articles and bulletins from Petrobras' geosciences publications. They were recovered and selected from the geological terms of the Chronostratigraphic table, which contains names of sedimentary rocks, names of Brazilian sedimentary basins, the terms related to Tectonics, Sedimentation and Magmatism and stratigraphic units.

In this report, we will present and discuss the changes that were made to this corpus in order to improve it.

2 Modifications

The corpus contained several problems that had a negative impact on machine learning experiments carried out with it. Below are the modifications made.

2.1 Removing empty categories

Description: There were some categories in GeoCorpus with an empty identifier.

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

Solution: The empty categories were removed from the corpus in order to standardize them. It is important to note that this change does not impact the annotation.

2.2 Removal of nested categories

Description: Some annotated terms contained nested categories.

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

Solution: In these cases, only the more specialized category was removed, leaving only the more generic of the two.

2.3 Correction of category annotation

Description: Some entities have been categorized into more than one class.

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

Solution: In these cases the annotation was redone. We sent the identified cases to an expert, who assigned the correct class to the entity.

2.4 Removing duplicate lines

Description: There were some repeated lines in GeoCorpus.

Example: Lines 766 and 776 were the same, containing the same sentence: "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."

Solution: Exactly identical lines have been removed from GeoCorpus, since repetitions tend to hinder machine learning. Altogether, 73 lines were removed, containing a total of 51 entities.

2.5 Correction of improperly broken lines

Description: There was an improper line break pattern in GeoCorpus. In some sentences with a comma or opening parentheses, there was a new line segmenting the sentence into two parts.

Solution: As not all lines with parentheses or commas had the incorrect line break, the phrases that presented this break were corrected manually.

2.6 Other category removal

Description: There was a category called 'others' in GeoCorpus, with 737 entities.

Solution: All entities in this category were passed on to an expert, and recategorized

into more specific categories, for entities that did not fit into the existing categories, new categories were created indicated by the expert. This was deemed necessary because the class 'others' was very wide only hindered automatic classification.

2.7 Standardization of categories

Description: There was no pattern in the name of the corpus categories, some were all capitalized, others in lowercase, and those with compound words alternated.

Solution: All categories are named with the Camel Case standard.

2.8 Entities without annotation

Description: 2913 entities in GeoCorpus that should be annotated and were not were identified. These entities were annotated one or more times in the corpus, but in certain instances were not annotated.

Example:
Sentence 1: grânulos subangulosos de <EM CATEG="sedimentaresSiliciclasticas">**quartzo**</EM>.
Sentence 2: em adição a outros minerais detríticos como o **quartzo**.
In this example, the 'quartzo' entity appears categorized in the first sentence, but in the second, 'quartzo' is not categorized.

Solution: All entities that should be categorized were categorized using a script. It was possible to use a script because the words that were not categorized had no problem of context or double meaning, which would make them appear at one time categorized and at another time not.

3 Analysis of GeoCorpus

After modifying the corpus in an attempt to obtain a better result, we performed an analysis on the entities of the modified corpus, together with an analysis on the entities of the unmodified corpus, for comparison purposes. With that, we want to compare and contrast the two versions of GeoCorpus.

In the original Geocorpus we had a sum of 6126 registered entities, divided into 20 classes, with the necessary modifications made, the impact we had on the number of

Table 1 – Comparison GeoCorpus: Original version x GeoCorpus: Revised version

Class	#Instances(Original)	#Instances(Revised)
Time		
age	796	799
eon	288	256
era	326	414
epoch	650	687
period	637	714
Rocks		
metamorphics	197	378
magmatics	222	582
siliciclasticSedimentary	741	1102
carbonateSedimentary	240	355
chemicalSedimentary	5	12
organicSedimentary	22	22
Constituents and Properties of Rocks		
sedimentaryRockConstituent	0	112
mineral	0	212
fossils	0	132
sedimentaryStructure	0	86
geologicalStructure	0	78
Site		
basinsGeologicalContext	262	663
sedimentationEnvironment	0	146
bentonic	13	27
planktonic	44	112
oilField	0	6
Elements of Stratigraphy		
sedimentaryBasin	243	552
stratigraphicUnit	578	764
geotectonicUnit	0	28
stratigraphy	0	247
formation	18	0
Others		
petroleumSystem	0	93
basinStructure	40	0
geomorphology	0	54
granulometry	67	129
chemicalElement	0	26
methodologicalProcedure	0	166
other	737	0
Sum	**6.126**	**8.954**

categories and classes is observed. In the modified Geocorpus, we have a total of 8954 registered entities, an increase of 2828 entities, divided into 30 classes, an increase of 10 classes.

Table 2 – Unique Entities by class, organized into superclasses - Revised Version

Class	#Unique Instances
Time	
age	84
epoch	74
period	62
era	47
eon	20
Rocks	
metamorphics	59
magmatics	58
siliciclasticSedimentary	160
carbonateSedimentary	77
chemicalSedimentary	4
organicSedimentary	1
Constituents and Properties of Rocks	
sedimentaryRockConstituent	24
mineral	6
fossils	29
sedimentaryStructure	28
sedimentaryBasin	83
geologicalStructure	19
Site	
sedimentationEnvironment	32
basinsGeologicalContext	121
bentonic	4
planktonic	9
oilField	2
Elements of Stratigraphy	
stratigraphicUnit	153
geotectonicUnit	8
stratigraphy	29
Otheers	
geomorphology	6
chemicalElement	3
granulometry	13
methodologicalProcedure	4
petroleumSystem	10
Soma	**1.229**

The number of unique entities of the new GeoCorpus was also analyzed. With

this table, we can see that within the 8954 total entities noted, there are a total of 1229 distinct entities, a number well below the total number of entities, which demonstrates that there are many repetitions of the same entities, something that we have to take into account.

4 Conclusion

The changes made to GeoCorpus were aimed at improving machine learning. Corrections, cleanups and new annotations produced an average increase of 8.68 percentage points when repeating the tests performed in *Embeddings for Named Entity Recognition in Geoscience Portuguese Literature* [2], our first paper using GeoCorpus. It is important to highlight that the model that obtained the best performance in the old GeoCorpus, remained the model with the best performance in the new GeoCorpus. This confirms that the changes made using machine learning techniques on the corpus easier without changing the substance of the results.

In the future, we can identify new ways to improve the corpus, such as joining some classes with few examples, and even adding more information to the corpus.

Bibliography

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>.

2 CONSOLI, B. S. et al. Embeddings for named entity recognition in geoscience portuguese literature. In: CALZOLARI, N. et al. (Ed.). *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. European Language Resources Association, 2020. p. 4625–4630. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.568/>.