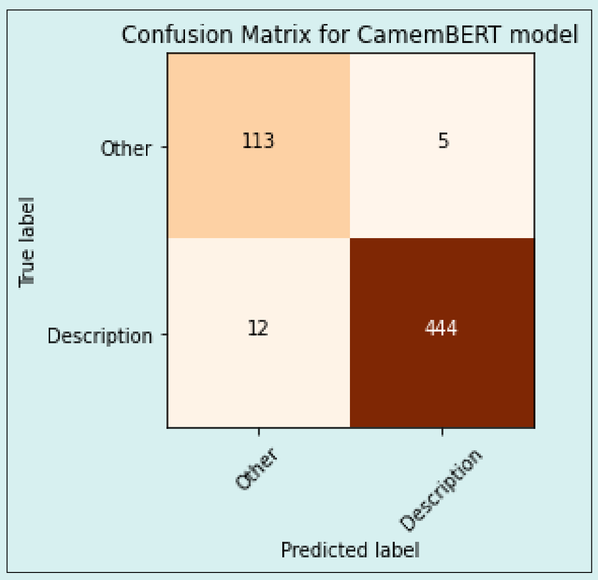
Géosciences pour une Terre durable
brgm

S0203
1605

On the application of NLP techniques to French geological descriptions

M. Khalid, C. Gracianne, M. Galant, R. Darnault,
C. Loiselet, V. Labbé



These sentences were correctly classified by a large language model:

- Alluvions recentesmelange sablegravier galet: Description
- Remblai terre noire puis colluvions basaltiques rouxvif à cailloutis emousses de basalte: Other
- Rave leger sableuse beige limon superficiel de basses terasses wurmienne: Other
- Sable argileux grisvert peu humide devenant rougeatre: Description

Introduction

Geological descriptions are written in natural language by different geologists during the acquisition process, without any constraints on the format or the content of the descriptions. As a result, the records are highly heterogeneous, ranging from brut observations to high-level interpretations and all the nuances in between. This heterogeneity imposes manual reprocessing of the data by experts to avoid mistakes occurring in downstream applications. Instead, we try to automate this process via Natural Language Processing (NLP) techniques.

Text	Label
VOLCANITES ACIDES	Description
ZONE FRACTUREE	Other
ALLUVIONS MODERNES SABLO-GRAVELEUSE	Other
CALCAIRE FRACTURE	Description
REMBLAI	Other
SABLE ARGILEUX	Description
COULEE DE BASALTE A PYROXENE ET OLIVINE	Description

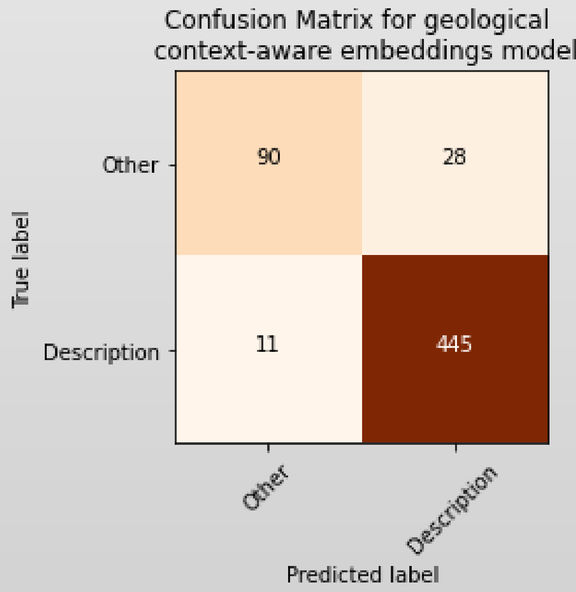
Methods

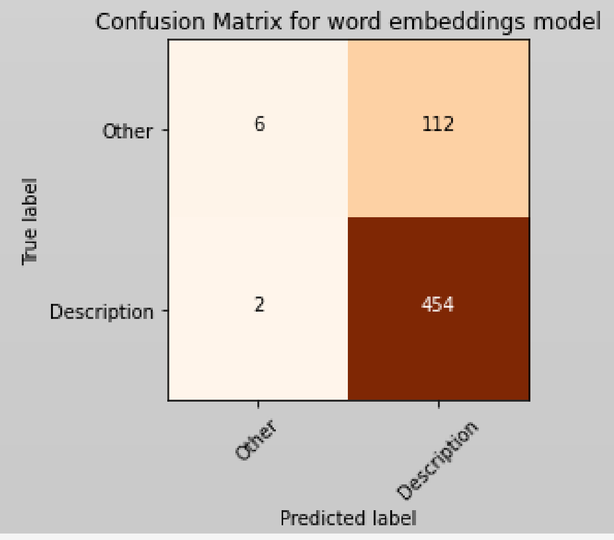
- CamemBERT (Bert based) large language model for French.
- Word embeddings with Convolutional Auto-encoders.
- Geological Context-aware word embeddings with convolutional Auto-encoders.

A very large neural net trained on a large corpus of French texts

A small neural net with a *word embeddings* layer randomly initialized

A small neural net with a *word embeddings* layer pre-trained on a corpus of geological texts in French





Take home tips:

- Use a large language model if it is available in your language.
- Geological context-aware embeddings improved the results and might have some merit.

Future works

We try to detect within a text, the part that describes the physical attributes like color and size. This will help data transfer of historical texts to modern and more structured databases.

References

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, et al.. CamemBERT: a Tasty French Language Model. 2019. ⟨hal-02445946⟩

Acknowledgments

Authors would like to thank Olivier Rouzeau for his help annotating the dataset.