Automatic Extraction of Structured Mineral Drillhole Results from Unstructured Mining Company Reports

Adam Dimeski and **Afshin Rahimi**
School of Information Technology and Electrical Engineering
The University of Queensland
a.dimeski@uqconnect.edu.au a.rahimi@uq.edu.au

Abstract

Aggregate mining exploration results can help companies and governments to optimise and police mining permits and operations, a necessity for transition to a renewable energy future, however, these results are buried in unstructured text. We present a novel dataset from 23 Australian mining company reports, framing the extraction of structured drillhole information as a sequence labelling task. Our two benchmark models based on Bi-LSTM-CRF and BERT, show their effectiveness in this task with a F_1 score of 77% and 87%, respectively. Our dataset and benchmarks are accessible online.[1]

1 Introduction

Mineral exploration involves drilling for core samples to assess their mineral composition. These assays are published in annual reports and other public announcements such as press releases. Often these results are presented in a semi-consistent non-tabular form. There is an industry demand for up-to-date mineral exploration results given that aggregate mineral composition information across a region or country can guide and optimise mineral exploration, however, current solutions involve manual collection of data directly from public company resources, which is expensive, time-intensive, and out-of-date (Riganti et al., 2015). This has become more important as the transition from fossil fuels to renewable energy has accelerated the demand for minerals such as lithium, nickel and rare earth metals. An assay report contains "drillhole sentences" which are phrases containing a unique drillhole code, depth, material, type and material percentage. See the example in Figure 1 from a mining company press release. The results are buried in long reports that contain images and natural language text with varying nomenclature, format and placement in the report across companies, geologists and mineral sectors, making their automatic extraction by regular expressions very challenging. The format of the drillhole sentences

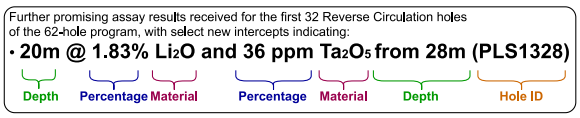

Figure 1: Excerpt from Pilbara Minerals ASX Announcement Pilbara Minerals (2021)

presents an opportunity to apply natural language processing techniques to automatically extract drillhole data. In this paper, we assess the performance of neural network models in extracting structured information about drillhole mineral exploration results. To the best of our knowledge, this work is the first to focus on extracting drillhole results from unstructured text, acknowledging prior work on extracting other geological information such as rick types Holden et al. (2019).

Our contributions are: a) developing a novel dataset for structured drillhole result extraction from 23 public Australian Stock Exchange (ASX) listed mining companies involved in mineral exploration.; b) formulating the extraction task as sequence labelling and presenting two benchmarks: a bidirectional LSTM network with a conditional random field layer (BI-LSTM-CRF) (Lample et al., 2016) and BERT (Devlin et al., 2019), showing that both perform fairly well; and c) performing error analysis and identifying major error types to guide future work.

2 Related Work

In this section, we provide insight into previous work performed on extracting geological data from reports and an overview of neural network models for sequence labelling tasks.

[1] Link to the dataset

2.1 Geological NLP

Various NLP approaches have been used to extract geological data from reports. GeoDocA is a search portal developed to search for geological terms in reports, research papers and geology results (Holden et al., 2019). GeoDocA performs part-of-speech tagging using a POS tagger from Manning et al. (2014). Although GeoDocA implements a non-neural network machine learning approach, compared to other research GeoDocA's results are most similar to the drillhole data we have extracted from public Australian mining company reports; and use a more textually similar corpus.

Consoli et al. (2020) applies a Bi-LSTM-CRF model to perform POS tagging on Portuguese geoscience literature, however, its objective was to compare different methods of word embedding. They made use of an existing corpus, GeoCorpus-2 and its predecessor from Amaral (2017) tagging rock types and numeric data such as age and period. Consoli et al. (2020) achieved a F_1 from 53.71% up to 84.63% while Amaral (2017) achieved an F_1 score of 54.33%. Challenges in developing NLP models for mining report extraction include the availability of text corpus that contains useful geological terms, and industry-level terminology as shown in Tessarollo and Rademaker (2020).

2.2 Deep Learning for Sequence Labelling

Various neural network models have been used for NLP. LSTM models and their variations have proven to be robust against newer models in natural language tasks including sequence labelling(Melis et al., 2017). Newer models are being developed to better handle more complex language structure, more recently with transformers-based models such as BERT (Devlin et al., 2018). Drillhole sentences are in a structured format, contained in unstructured text. The Bi-LSTM-CRF and BERT benchmark models used to perform sequence labelling on the drillhole sentence implement model tuning for structured data in their loss functions. BERT make use of the cross-entropy loss function to tune the model weights to sentence structure while the Bi-LSTM-CRF model uses an individual CRF layer on top of the Bi-LSTM layers that is tuned to sentence structure. One of the key performance distinctions transformers have shown is being able to better recognise sentence-level context of words, beyond just feature-based models Ghaddar et al. (2021), hence the addition of a CRF layer to a base LSTM model (Lample et al., 2016).

3 Data

The dataset consists of 50 reports from 23 ASX-listed mining exploration companies. The selection criteria for reports, extraction and segmentation of text from the PDF files and the annotation process are reported in this section. Additionally, to test the generalisation ability of the models, the dataset is split into training, dev and test sets based on a) random; b) material; and c) company, to find out if the benchmark models will be able to generalise across materials and companies.

3.1 Selecting Reports

We chose 40 publicly listed mineral companies on the Australian Stock Exchange (ASX). The selection involved sorting the mining companies according to their market capitalisation and randomly selecting 7, 7, and 6 companies from the top, middle, and bottom bins, respectively. In addition, we also included the last 20 mining companies recently listed on the ASX to include more variety in terms of formatting and materials. Annual reports dating back to 2014 were collected from the websites of these companies. Reports and companies without any drillhole results were excluded. The final corpus includes 50 reports from 23 companies which covers a variety of drillhole result formats, materials, localities, and company maturity.

3.2 Preprocessing

The 50 reports included in the corpus are in PDF format. We extracted the text using a PDF parser. Due to the inherent nature of automated PDF extraction, it introduced conversion artefacts into the extracted text resulting in the fragmentation of sentences. To split the extracted text into sentences for the sequence labelling annotation task, we used a rule-based tokeniser that splits sentences after common punctuations and new line characters. Initially, we used the Punkt sentence tokeniser Kiss and Strunk (2006), however, it yielded highly irregular sentence lengths and split drillhole sentences apart as a result of the quality of text extracted from the PDF files. The rule-based sentence tokeniser, however, worked fairly well in comparison. The corpus contained over 20,000 sentences with the majority being of a consistent length.

Set	Hole ID	Material	Percentage	Depth	Extra	Outside	Sentence Count
Train	51%	56%	56%	53%	51%	53%	17.2K
Dev	19%	16%	15%	17%	20%	15%	2.2K
Test	31%	28%	29%	30%	29%	32%	3.3K
Count	1.2K	1.4K	1.9K	2.4K	3.7K	667K	22.7K

Table 1: Tag split among each set shown as a percentage of the total count

3.3 Annotation

Annotation of the dataset was performed on the dataset text files using the IOB sequence tagging format by the author of this work. Four tags were chosen to extract the material: hole id, percentage, material, and depth. A fifth tag was included for words commonly used in drillhole sentences such as "from" and "to" when referring to the hole depth. The tagging schema is shown in Table 2.

Tag	Category	Example
H	Hole ID	PLS1328
M	Material	Li2O
P	Percentage	0.23%
D	Depth	3m
E	Extra	from
O	Outside	This

Table 2: IOB Tags

Due to the varying types of drillhole sentence formats, a set of rules was adopted to have consistent annotation of the data. The most significant rules:

- *Sentence Length:* A drillhole sentences must be separated by punctuation or words tagged as outside.
- *Non-numerical values:* Non-numeral depths and percentage were included.
- *Hole ID Format:* Drillhole sentences that use a location instead of a hole ID were not included unless directly adjacent to a hole ID.
- *Punctuation:* Punctuation that is involved with the direct indication the start of a drillhole sentence was tagged with the extra tag.
- *Filler Words:* Words that are used inside a drillhole sentence are tagged as extra when are used to refer to depth, material, percentage or Hole ID tags.

One of the challenges faced with tagging the dataset was the similarity of drillhole sentences to other geological sentences in the reports. For example, within a piece of text a part of a drillhole result might be mentioned without referring to a specific drillhole. We decided to tag these sentences even if they weren't associated with a hole ID to improve the performance of the model as these sentences also contain material, depth, and percentage attributes. For downstream applications, it will be easy to ignore such information as they are not accompanied by a drillhole ID.

3.4 Annotation Quality

In total, the entire corpus contained over 680K words with 10.8K words being tagged as part of a drillhole sentence (not tagged as outside) resulting in a highly imbalanced dataset. The corpus contained a total of 22.7K sentences. Resource constraints meant that the compiling and annotation of the dataset was a result of a single annotator. Therefore, annotation quality could not be assessed with an inter-annotator agreement. Some analysis of the annotation can be inferred from the error analysis, however the results of the Bi-LSTM-CRF and BERT models will based on a dataset with some annotation inconsistencies.

3.5 Dataset Split

The format, placement in text, and materials vary between reports. For example, a company might use the drillhole ID in parenthesis at the end of a drillhole sentence while another company might use the drillhole ID at the front of a sentence followed by a colon. Similarly, the same material might have various names depending on the type of nomenclature the geologist use. To find out if a model trained on a specific variety of data will generalise on unseen data we split the dataset into training, dev and test sets based on a) random split; b) material; and c) company. The split based on material and company was performed in a way that materials and companies in the dev and test sets were disjoint from the training set, however, Gold, 'au', was the most common material tag, accounting for 80% of all material tags, this was assumed to be in all the material datasets. Table 1 shows the proportion of each tag in each set. The percentage

	Bi-LSTM-CRF			BERT		
Dataset	**P (%)**	**R (%)**	**F_1(%)**	**P (%)**	**R (%)**	**F_1(%)**
Random	91 ± 2	67 ± 11	77 ± 7	91 ± 2	78 ± 4	84 ± 2
Material	86 ± 2	75 ± 3	81 ± 3	87 ± 1	87 ± 1	87 ± 1
Company	89 ± 4	69 ± 12	77 ± 7	87 ± 1	87 ± 1	87 ± 1

Table 3: Test set results evaluated over the three split methods averaged over runs with five different seeds for the two benchmarks, Bi-LSTM-CRF and BERT. Evaluation measures, **P** (Precision), **R** (Recall) and $\mathbf{F_1}$ scores show standard deviation values. For detailed tag-specific results see appendix.

of tags among the training, dev, and test sets were consistent across the three datasets.

4 Method

To measure the generalisation ability of a sequence labeling task trained on our dataset we used two benchmarks, both evaluated by precision, recall, and f1 using seqeval library (Nakayama, 2018).

Bi-LSTM-CRF: We use the Bi-LSTM-CRF model proposed in Lample et al. (2016) for the sequence labeling task of identifying drillhole result segments. This model uses both word and character embeddings which is suitable for our task which involves chemical formulas and numerical tokens that might only be captured through character information. The character and word embeddings are concatenated and fed to a Bi-LSTM to capture sequential and contextual information. The resulting final hidden states of the two directions are concatenated and fed into a Conditional Random Field layer that models the conditional probability of the tags.

BERT: We use BERT (Devlin et al., 2019) to find out the effect of pre-training on massive amounts of text on the performance of our task given the relatively small training set and the ability of the pre-trained transformers to transfer knowledge across tasks in low-resource settings. Due to the computational demands, we only experiment with the base (uncased) version of BERT which is lighter compared to BERT-Large in terms of the model size. Given that BERT does not take into account the characters, it is interesting to find out if it can outperform Bi-LSTM-CRF which uses character information.

4.1 Model Parameters

The Bi-LSTM-CRF model uses a default batch size of 32 sentences and embedding size of 256. Tuning of the learning rate was done by applying the "LR Range Test" Smith (2015). A learning rate value of 0.008 was set for the random and material split datasets and a learning rate of 0.005 for the company split dataset. The BERT model uses transformers library with a maximum sequence length of 512 and a default learning rate of 5e-5 for all dataset splits.

5 Results

Evaluation results are shown in Table 3. Overall, both Bi-LSTM-CRF and BERT perform well with an F1 score of 78% and 86%, respectively. Recall is considerably lower than precision for both models which is the result of the class imbalance in the training set, having a large number of outside tags. BERT outperforms Bi-LSTM-CRF substantially in terms of recall across the three dataset splits, demonstrating better adaptation to various drillhole sentence structures, contexts and nomenclature used in the mining reports. The standard deviation of Bi-LSTM-CRF across the three dataset splits and the three evaluation measures was much higher than BERT, indicating that BERT, as expected, is more robust to variation in language use. In terms of splits, while Bi-LSTM-CRF shows variation across the three datasets, BERT is able to consistently generalise to unseen examples from various companies and materials.

Upon further inspection of tag-specific results (shown in appendix), the recall of Bi-LSTM-CRF is 38% which is substantially lower than that of BERT with a recall of 66%. The identification of drillhole is an essential component of extracting the structured drillhole results from reports as it can uniquely identify a drillhole across several reports.

5.1 Error Analysis

Given the lower computational demands of the Bi-LSTM-CRF model, error analysis was performed on the model to identify the types of errors the model makes. The most frequent errors can be

categorised into five classes:

- *Context:* variability and inconsistency in context.
- *Annotation:* ambiguity in annotation.
- *UNK*: unseen words during training.
- *Split tag:* a tag split across multiple tokens.
- *Not O:* Correct tag is O, however, the model predicts otherwise.

The error type counts are shown in Table 4. Overall, the context and UNK errors are the most frequent error types that can be addressed by creating noisy data e.g. replacing unseen materials in various sentences to create noisy supervision or to increase annotation.

Split	Random	Material	Company
Error	**Count**	**Count**	**Count**
Context	1157	1053	617
Annotation	83	289	218
UNK	408	386	254
Split Tag	22	238	83
Not O	113	471	306
Total	1249	1345	883

Table 4: Error type counts for the three dataset splits for Bi-LSTM-CRF

6 Conclusion

We present our work in creating a novel dataset for extracting structured drillhole results from unstructured mining exploration reports. We formulate this task as sequence labeling and show that while both our two benchmarks Bi-LSTM-CRF and BERT perform well with an F1 score of 77% and 87%, respectively, BERT substantially outperforms Bi-LSTM-CRF and is more robust to variation in language and format. We performed error analysis on the Bi-LSTM-CRF predictions and identified context variation and unseen tokens in training data to be the most frequent error types. Our error analysis indicates improvement pathways for the Bi-LSTM-CRF model which is more efficient for use in most common computing settings.

References

Daniela Oliveira Ferreira do Amaral. 2017. *Reconhecimento de entidades nomeadas na ?rea da geologia : bacias sedimentares brasileiras*. Ph.D. thesis. Escola Polit?cnica.

Bernardo Consoli, Joaquim Santos, Diogo Gomes, Fabio Cordeiro, Renata Vieira, and Viviane Moreira. 2020. Embeddings for named entity recognition in geoscience Portuguese literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4625–4630, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

Eun-Jung Holden, Wei Liu, Tom Horrocks, Rui Wang, Daniel Wedge, Paul Duuring, and Trevor Beardsmore. 2019. Geodoca – fast analysis of geological content in mineral exploration reports: A text mining approach. *Ore Geology Reviews*, 111:102919.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *CoRR*, abs/1707.05589.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Pilbara Minerals. 2021. *Further Exceptional Drilling Results at Pilgangoora*.

Angela Riganti, Terence R Farrell, Margaret J Ellis, Felicia Irimies, Colin D Strickland, Sarah K Martin, and Darren J Wallace. 2015. 125 years of legacy data at the geological survey of western australia: Capture and delivery. *GeoResJ*, 6:175–194.

Leslie N. Smith. 2015. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186.

Alexandre Tessarollo and Alexandre Rademaker. 2020. Inclusion of lithological terms (rocks and minerals) in the open Wordnet for English. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 33–38, Marseille, France. The European Language Resources Association (ELRA).

A Detailed Results

Bi-LSTM-CRF									
Dataset	Random Split			Material Split			ASX Split		
Tag	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**	**P (%)**	**R (%)**	**F_1 (%)**
Depth	89 ± 2	70 ± 14	78 ± 9	85 ± 3	81 ± 2	83 ± 2	82 ± 1	69 ± 5	74 ± 3
Extra	91 ± 2	70 ± 10	79 ± 6	86 ± 3	76 ± 3	81 ± 1	91 ± 3	63 ± 6	74 ± 3
Hole ID	88 ± 12	39 ± 4	53 ± 4	82 ± 10	37 ± 8	50 ± 8	84 ± 7	39 ± 6	53 ± 4
Material	95 ± 1	71 ± 12	81 ± 8	90 ± 2	78 ± 3	84 ± 1	92 ± 1	54 ± 4	68 ± 3
Percentage	93 ± 4	70 ± 14	79 ± 11	86 ± 3	76 ± 3	80 ± 1	92 ± 1	60 ± 5	73 ± 3
Total	91 ± 2	67 ± 11	77 ± 7	86 ± 2	75 ± 3	81 ± 3	89 ± 4	69 ± 12	77 ± 7
BERT									
Depth	92 ± 2	83 ± 4	87 ± 2	87 ± 2	92 ± 2	89 ± 1	87 ± 2	92 ± 2	89 ± 1
Extra	91 ± 3	78 ± 5	84 ± 3	84 ± 1	89 ± 2	86 ± 1	84 ± 1	89 ± 2	86 ± 1
Hole ID	86 ± 6	62 ± 7	71 ± 4	88 ± 2	68 ± 9	76 ± 5	89 ± 2	68 ± 9	76 ± 5
Material	96 ± 2	82 ± 4	87 ± 1	92 ± 1	83 ± 2	87 ± 1	92 ± 1	83 ± 2	87 ± 1
Percentage	92 ± 3	80 ± 2	86 ± 2	88 ± 1	88 ± 2	88 ± 1	88 ± 1	88 ± 2	88 ± 1
Total	91 ± 2	78 ± 4	84 ± 2	87 ± 1	87 ± 1	87 ± 1	87 ± 1	87 ± 1	87 ± 1

Table 5: Test set results evaluated over the three split methods averaged over runs with five different seeds for the two benchmarks, Bi-LSTM-CRF and BERT. Evaluation measures, **P** (Precision), **R** (Recall) and $\mathbf{F_1}$ scores show standard deviation values.

It was shown that the variation of the F_1 score was higher between datasets for the Bi-LSTM-CRF model than the BERT model. Additionally, the variation of the F_1 score for different seeds was also higher for the Bi-LSTM-CRF model compared to the BERT model.

B Detailed Error Analysis

Bi-LSTM-CRF Error Analysis

		Dataset Split		
Code	**Description**	**Random**	**Material**	**ASX**
1	Probability of O is greatest	1138	874	577
2	Not O prediction. Followed default context when other context is required	24	177	92
3	Unknown word is a specific depth/material/percentage	408	386	254
5	General Lack/incorrect of Context	596	265	212
5.1	Spurious Tag	201	292	192
5.2	Slightly Dissimilar context	336	319	121
6	Correct, but tagged as O because depth was not a number	9	11	10
7	Error caused by split tag	280	238	83
9	Correct but not associated with drillhole	49	240	170
10	Error caused by previous error in sequence	23	76	225
C	Fully Correct	25	38	38
NO	Predicted tag was not O	113	471	306
Total		1249	1345	883

Table 6: Detailed Error Results for Bi-LSTM-CRF results

Error analysis was performed on the results for the Bi-LSTM-CRF model. Using a predefined error schema in Table 6, a rules-based error tagger was implemented to sort the errors into category types. The errors were further manually assessed for their correct error type and further categorised into error subtypes. The majority of errors were due to context errors, which were further defined into subcategories; spurious tags are tags that have been tagged outside of the drillhole sentence and the similar context subcategory was for incorrect tags that are in an uncommon or irregular form of drillhole sentence.

Errors that were as a result of incorrect annotation and the model made a correct prediction made up to 5% of all errors. In total, up to 23% of errors were due to inconsistent/incorrect annotation and the model made the correct prediction.