Automated ESG reports analysis by joint entity and relation extraction

Adrien Ehrhardt, Minh-Tuan Nguyen

SoGood 2021, 15/09/2021

Table of Contents

Problem setting

Problem setting

"Operational Research Group" at Groupe Crédit Agricole $\approx$ internal AI consultancy team.

Problem setting

"Operational Research Group" at Groupe Crédit Agricole ≈ internal AI consultancy team.

The *Corporate and Investment Banking* branch of (the) bank(s) has stakes in (very big) corporations, *e.g.* through bonds, shares.

Problem setting

"Operational Research Group" at Groupe Crédit Agricole $\approx$ internal AI consultancy team.

The *Corporate and Investment Banking* branch of (the) bank(s) has stakes in (very big) corporations, *e.g.* through bonds, shares.

Banking industry $\iff$ quantifying the **risk** / **reward** balance of an operation.

Problem setting

"Operational Research Group" at Groupe Crédit Agricole $\approx$ internal AI consultancy team.

The *Corporate and Investment Banking* branch of (the) bank(s) has stakes in (very big) corporations, *e.g.* through bonds, shares.

Banking industry $\iff$ quantifying the **risk** / **reward** balance of an operation.

Traditional default risk + many (new) types of risk, among which emerging Environmental, Societal and Governance (ESG) risks.

Problem setting

"Operational Research Group" at Groupe Crédit Agricole $\approx$ internal AI consultancy team.

The *Corporate and Investment Banking* branch of (the) bank(s) has stakes in (very big) corporations, *e.g.* through bonds, shares.

Banking industry $\iff$ quantifying the **risk** / **reward** balance of an operation.

Traditional default risk + many (new) types of risk, among which emerging Environmental, Societal and Governance (ESG) risks.

- Default risk: higher carbon costs, stranded assets, ...
- Risks to reputation: funding brown and / or shady businesses

Problem setting: ESG reports analysis

Corporations disclose "extra-financial" (ESG and/or CSR) reports.

Example of such a report:
`https://www.cotecorp.com/Groupe_PSA_2019_CSR_Report.pdf`.

Problem setting: ESG reports analysis

Corporations disclose "extra-financial" (ESG and/or CSR) reports.

Example of such a report:
`https://www.cotecorp.com/Groupe_PSA_2019_CSR_Report.pdf`.

Dedicated ESG teams at major CIB banks to:

- Analyze these reports;
- Detect potential issues;
- Detect commitments.
- ...

Problem setting: ESG reports analysis

Corporations disclose "extra-financial" (ESG and/or CSR) reports.

Example of such a report:
`https://www.cotecorp.com/Groupe_PSA_2019_CSR_Report.pdf`.

Dedicated ESG teams at major CIB banks to:

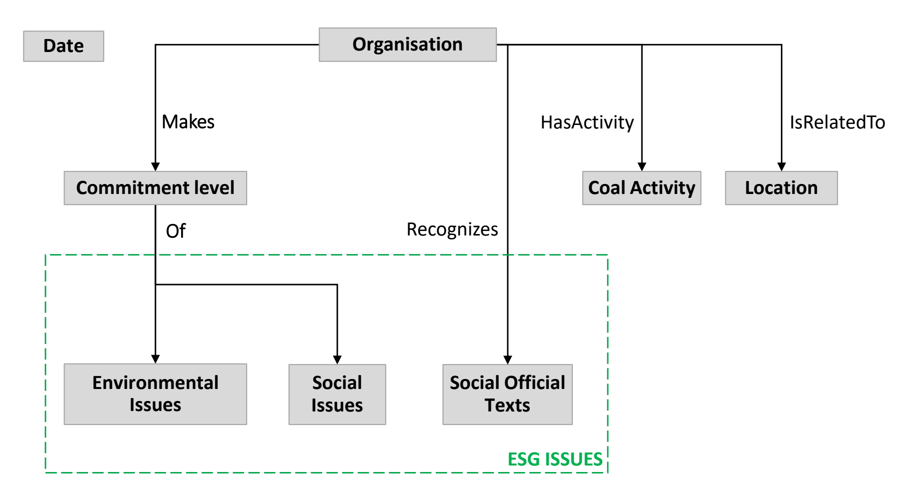
- Analyze these reports;
- Detect potential issues;
- Detect commitments.
- ...

Discrepancy between:

- Analysts' available time (up to 4000 reports to analyze per year);
- Size of the reports (308 pages in this example);
- Proportion of useful information.

Problem setting: ClimLL dataset

Data model



Problem setting: ClimLL dataset

Descriptive statistics

- 7,500 sentences from 372 paragraphs (280 into training - 92 into test set);
- All paragraphs from a given report belong to the same split;
- In total, 28,751 entities and 5,864 relations from 31 ESG reports.

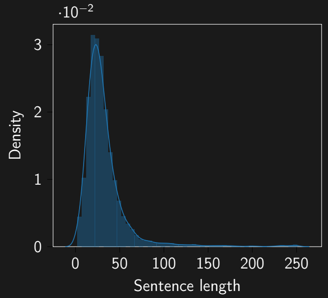
Problem setting: ClimLL dataset

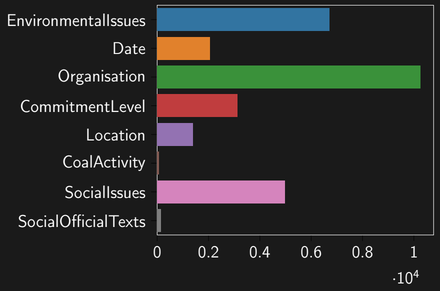
Descriptive statistics

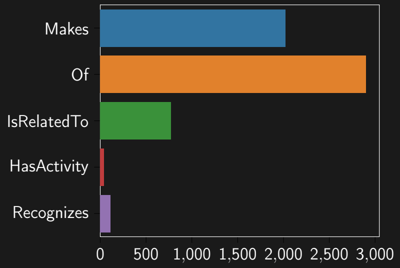
- 7,500 sentences from 372 paragraphs (280 into training - 92 into test set);
- All paragraphs from a given report belong to the same split;
- In total, 28,751 entities and 5,864 relations from 31 ESG reports.

Descriptive statistics

- 7,500 sentences from 372 paragraphs (280 into training - 92 into test set);
- All paragraphs from a given report belong to the same split;
- In total, 28,751 entities and 5,864 relations from 31 ESG reports.







Related work

Related work: Natural Language Processing

x	In addition to $\underbrace{\text{monitoring}}_{\text{Commitment}}$ $\underbrace{\text{rainwater}}_{\text{Environmental Issue}}$...
y	Commitment Environmental Issue

$$\mathrm{x} \qquad \text{In addition to } \underbrace{\text{monitoring}}\ \underbrace{\text{rainwater}}\ldots$$

$$\mathrm{y} \qquad \text{Commitment Environmental Issue}$$

Vocabulary:
(after tokenization)

$$\left\{\begin{array}{ll} 0:\ \text{Addition} & 4:\ \text{On} \\ 1:\ \text{In} & 5:\ \text{Rain\#\#} \\ 2:\ \text{\#\#ing} & 6:\ \text{To} \\ 3:\ \text{Monitor\#\#} & 7:\ \text{\#\#water} \end{array}\right\}$$

$\hat{f} = \min_{f \in \mathcal{H}} \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})$.

Related work: Natural Language Processing

$$\mathrm{x} \qquad \text{In addition to } \underbrace{\text{monitoring}}\ \underbrace{\text{rainwater}}\ldots$$

$$\mathrm{y} \qquad \text{Commitment Environmental Issue}$$

Vocabulary:
(after tokenization)

$$\left\{\begin{array}{llll} 0: & \text{Addition} & 4: & \text{On} \\ 1: & \text{In} & 5: & \text{Rain\#\#} \\ 2: & \text{\#\#ing} & 6: & \text{To} \\ 3: & \text{Monitor\#\#} & 7: & \text{\#\#water} \end{array}\right\}$$

$\hat{f} = \min_{f \in \mathcal{H}} \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})$.

$$\hat{f}\begin{pmatrix} \text{In} \\ \text{Addition} \\ \text{To} \\ \text{Monitoring} \\ \\ \text{Rainwater} \end{pmatrix} = \hat{f}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \approx \begin{pmatrix} \\ \\ \\ \text{Commitment} \\ \\ \text{Environmental Issue} \end{pmatrix}$$

Related work: Natural Language Processing

$$\mathrm{x} \qquad \text{In addition to } \underbrace{\text{monitoring}}\ \underbrace{\text{rainwater}}\ldots$$

$$\mathrm{y} \qquad \text{Commitment Environmental Issue}$$

Vocabulary:
(after tokenization)

$$\left\{ \begin{array}{ll} 0:\ \text{Addition} & 4:\ \text{On} \\ 1:\ \text{In} & 5:\ \text{Rain\#\#} \\ 2:\ \text{\#\#ing} & 6:\ \text{To} \\ 3:\ \text{Monitor\#\#} & 7:\ \text{\#\#water} \end{array} \right\}$$

$\hat{f} = \min_{f \in \mathcal{H}} \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})$.

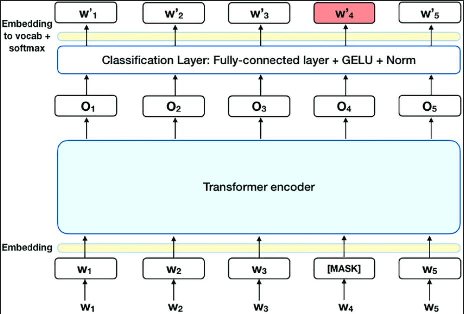
$$\hat{f}\begin{pmatrix} \text{In} \\ \text{Addition} \\ \text{To} \\ \text{Monitoring} \\ \\ \text{Rainwater} \end{pmatrix} = \hat{f}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \cancel{0} & \cancel{0} & \cancel{1} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{0} & \cancel{1} \end{pmatrix} \approx \begin{pmatrix} \\ \\ \\ \text{Commitment} \\ \\ \text{Environmental Issue} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \\ 2 \end{pmatrix}$$

Related work: Transformers

- No proximity between words (everything equidistant);
- The representation can, even after tokenization, be high-dimensional;
- The representation is not context-dependent.

Related work: Transformers

- No proximity between words (everything equidistant);
- The representation can, even after tokenization, be high-dimensional;
- The representation is not context-dependent.
- Solution: make use of the Transformer [3] neural network architecture, associated with "pre-training" on a large vocabulary, *e.g.* BERT [1].

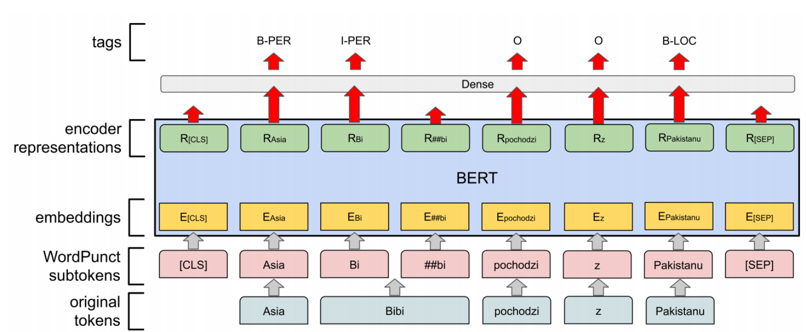


Two applications

NER & RE Pipeline: Named Entity Recognition

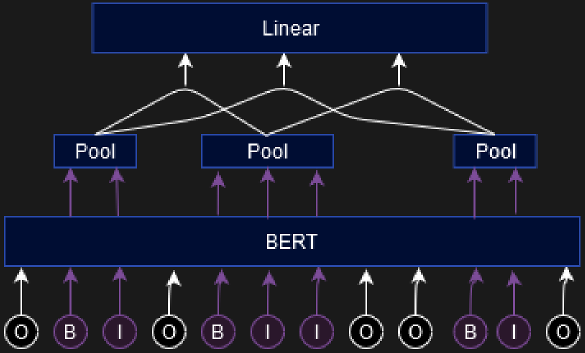
A two-stage procedure:

Stage 1: perform Named Entity Recognition.
In practice, a simple classifier *after* BERT representation.



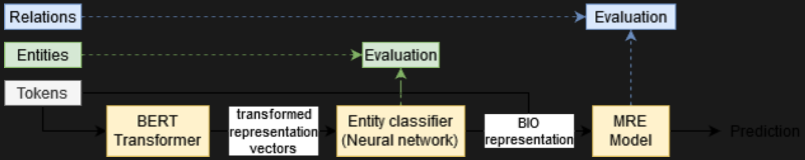
Stage 2: perform Relation Extraction on pairs of **true** entities.

In practice, a simple classifier *after* concatenating the average BERT representations of each pair of entities.



Wrapping up & predicting:

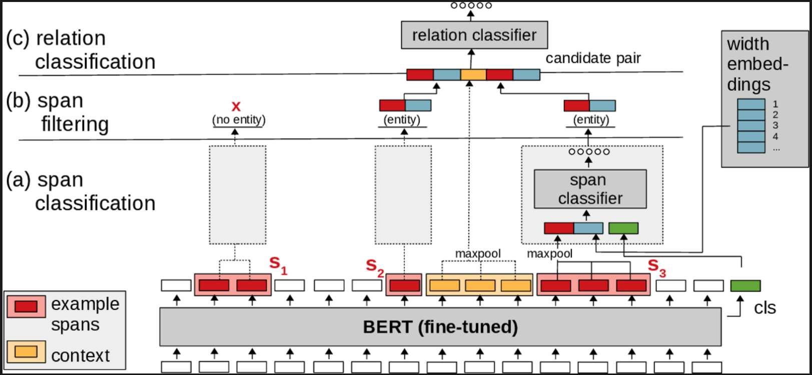
1. Predict entities
2. Predict relations based on **predicted** entities



- Subject to compound error

Joint NER & RE

SpERT [2] “blends” entity recognition & relation extraction



Numerical experiments

Numerical experiments: NER & RE Pipeline

Table: NER results on ClimLL (test set, micro-average).

Classifier	(1) Sentence-by-sentence Precision	Recall	F1	(2) 128-by-128 Precision	Recall	F1
k-nearest neighbor	0.79	**0.80**	**0.79**	0.79	**0.80**	**0.79**
Decision tree	0.42	0.46	0.44	0.43	0.47	0.45
Random forest	**0.94**	0.39	0.52	**0.93**	0.40	0.53
Neural network (512)	0.80	0.75	0.77	0.84	0.74	0.78
Neural network (1024)	0.81	0.76	**0.79**	0.83	0.75	0.78
IBM	0.87	0.70	0.78			

Numerical experiments: NER & RE Pipeline

Table: MRE results on test set.

Dataset	Average	Precision	Recall	F1
ClimLL	Micro	0.61	0.54	0.57
	Macro	0.55	0.54	0.54
CoNLL04	Micro	0.65	0.58	0.61
	Macro	0.66	0.61	0.63

Numerical experiments: Joint NER & RE

Table: Joint entity and relation extraction results on test set.

			NER			Joint NER & RE		
Dataset	Average	Model	Precision	Recall	F1	Precision	Recall	F1
CoNLL04	Micro	NER-RE	0.75	0.80	0.77	0.36	0.47	0.41
		SpERT	**0.86**	**0.91**	**0.89**	**0.71**	**0.70**	**0.70**
	Macro	NER-RE	0.71	0.74	0.73	0.41	0.51	0.45
		SpERT	**0.84**	**0.88**	**0.86**	**0.72**	**0.71**	**0.71**
SciERC	Micro	SpERT	0.64	0.72	0.68	0.31	0.45	0.37
	Macro	SpERT	0.65	0.71	0.68	0.34	0.41	0.35
ClimLL	Micro	NER-RE	0.67	0.68	0.68	0.23	0.18	0.20
		SpERT	**0.75**	**0.79**	**0.77**	**0.36**	**0.44**	**0.40**
	Macro	NER-RE	0.63	0.67	0.64	0.21	0.22	0.21
		SpERT	**0.75**	**0.78**	**0.77**	**0.46**	**0.58**	**0.50**

Conclusion

Conclusion

- ESG and CSR reports annotated so as to fit the needs of financial institutions;
- Two published works adapted to this new dataset;
- Strong results, direct application at the bank;
- Open-source code, model and API.

Future work

Future work

- Pre-training BERT on a specialized vocabulary set;
  $\rightarrow$ requires a lot of ressources and annotated documents.
- Incorporating the context of neighbouring sentences.
  $\rightarrow$ sentences in a given paragraph "share" meaning.

Demo!

https://github.com/adimajo/renard_joint

References I

[1] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[2] Markus Eberts and Adrian Ulges. "Span-based joint entity and relation extraction with transformer pre-training". In: *24th European Conference on Artificial Intelligence* (2020).

[3] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

References II

[4] Patrick Verga, Emma Strubell, and Andrew McCallum. "Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 872–884. DOI: 10.18653/v1/N18-1080. URL: https://www.aclweb.org/anthology/N18-1080.

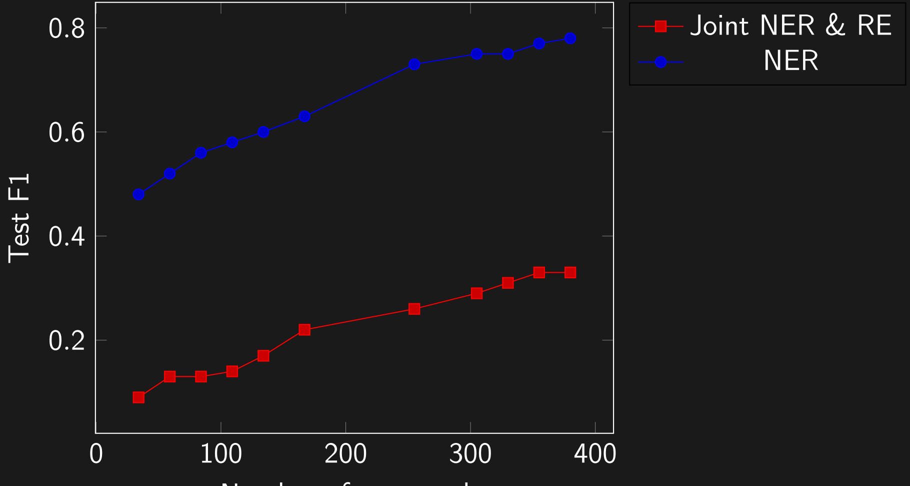
[5] Jue Wang and Wei Lu. "Two are Better than One: Joint Entity and Relation Extraction with Table-Sequence Encoders". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1706–1721. DOI: 10.18653/v1/2020.emnlp-main.133. URL: https://www.aclweb.org/anthology/2020.emnlp-main.133.

Appendix

Evolution of test F1 for the IBM model

Evolution of test F1 for the IBM model

372 paragraphs were sufficient as the F1 scores for NER and Joint NER & RE stopped improving using the IBM proprietary model.



Named Entity Recognition representation

Named Entity Recognition representation I

A popular output representation of NER is BIO (Begin, In, Out) embedding, where each word is marked as the beginning, inside, or outside of an entity (see *e.g.* [4, 5]); however, this representation does not allow overlapping entities.
Span-based methods [2], which classify spans of words, can extract the spans of these overlapping entities.

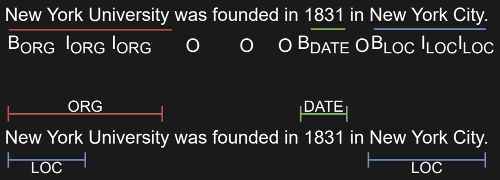


Figure: Examples of BIO (above) and span-based (below) representations.

Named Entity Recognition representation II

In the ClimLL dataset, even though the entities are presented in the span-based format in the dataset, there is no overlapping entity.

Thus, it is also possible to convert to BIO format.

Multiple relations can exist in the same sentence but relations cannot span across sentences. This facilitates splitting the paragraphs by sentence.

Evolution of loss functions

Evolution of loss functions

The entity and relation losses as well as the F1 score on the validation set throughout the training process (30 epochs) of SpERT on ClimLL.
Both entity and relation losses reached their minimum after only a few epochs while the validation F1 score kept improving.

Figure: Entity loss (left), relation loss (center) and F1-score on ClimLL w.r.t. training epochs (right).