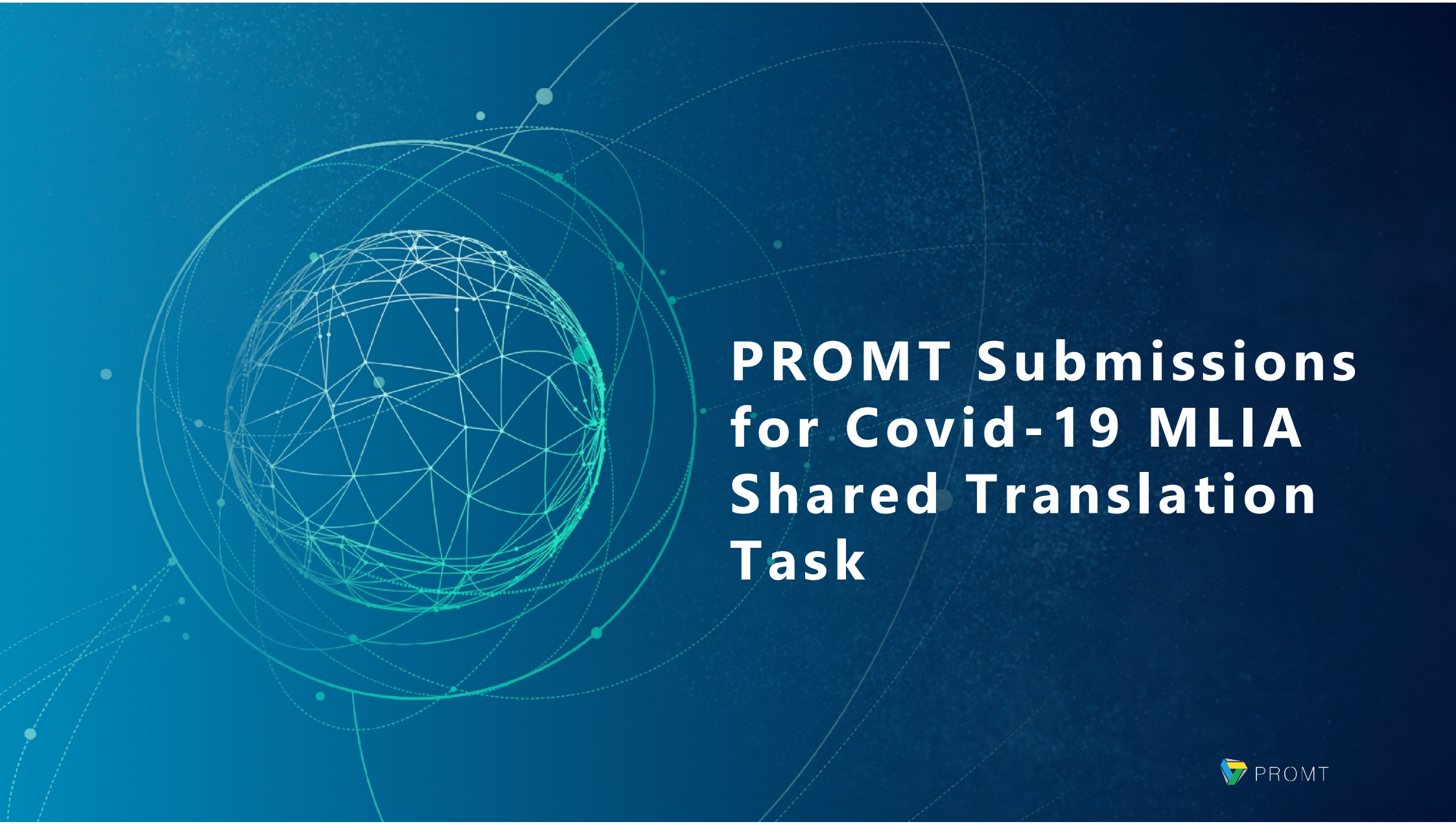
PROMT Submissions for Covid-19 MLIA Shared Translation Task
PROMT

DATA

Constrained

- all provided train data
- no filtering
- 8k SentencePiece model
- language tags on the English side

Unconstrained

- OPUS, statmt.org, private data
- no fine-tuning (no Covid-19 MLIA Data used)
- case-insensitive BPE, 8 to 16k for different models

PROMT

MODELS

Constrained

- baseline transformer, Marian
- a single multilingual model, all language pairs (EN to ES, DE, IT, EL, FR, SV)
- shared vocabulary
- vocabulary filtering (removing infrequent tokens).

Unconstrained

- baseline transformers, Marian
- all language pairs, multilingual models for IT (EN to IT, PT) and SV (EN to SV, DA, NO)
- shared vocabularies except for EN to EL

PROMT

RESULTS and FUTURE WORK

Rank top in all directions except for English-German
?

- SentencePiece
- small SentencePiece model
- multilinguality, however
    - full deduped EN corpus 1.2M sentences
    - EN source sides of the parallel corpora:
      800k (EN-SV) to 1.03M (EN-ES)

Constrained

- synthetic data

Unconstrained

- fine-tuning

PROMT