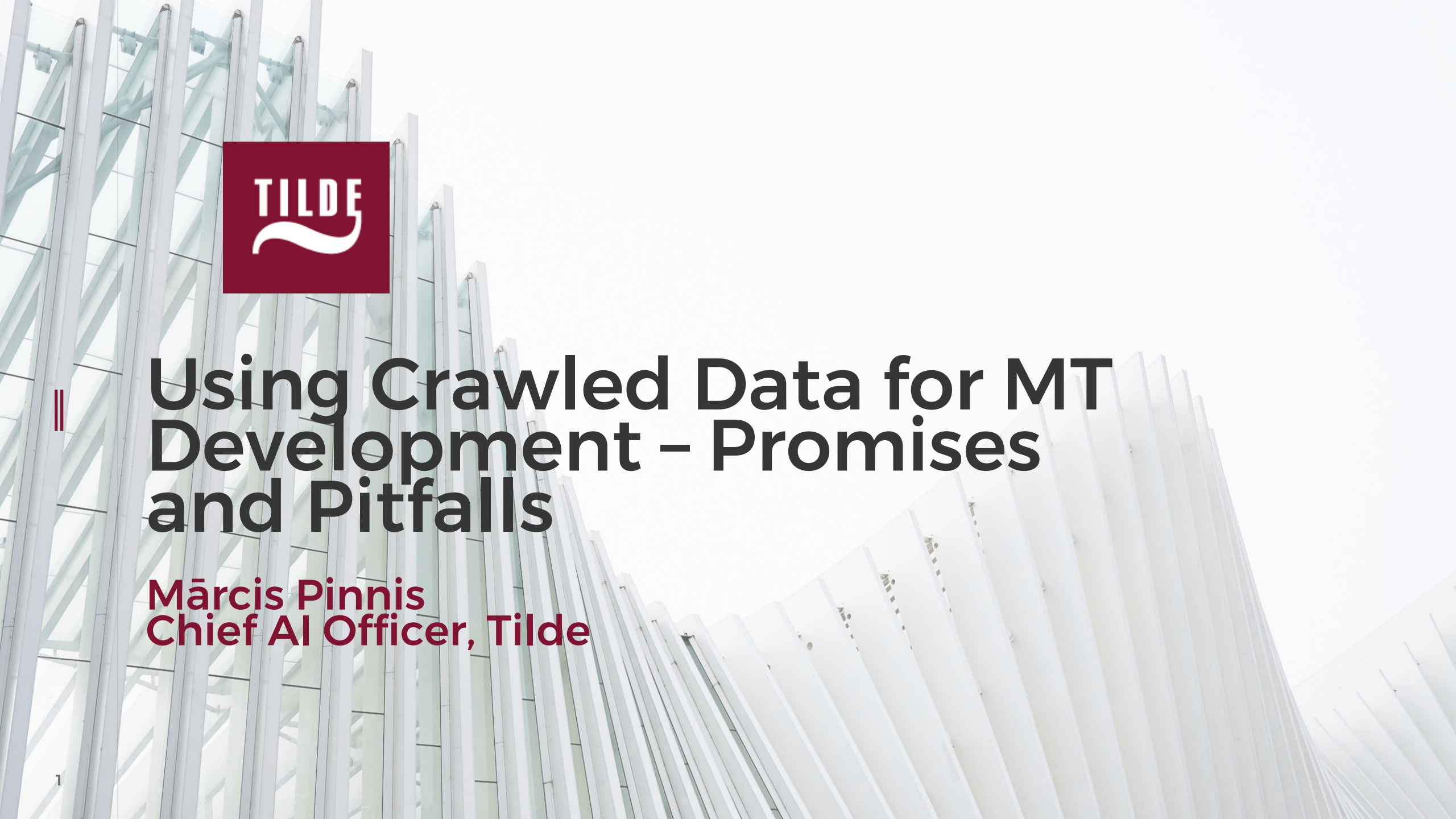


Using Crawled Data for MT Development – Promises and Pitfalls

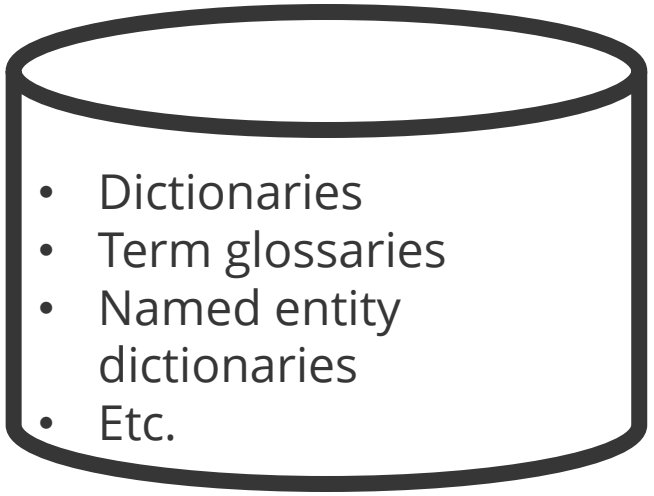
Mārcis Pinnis
Chief AI Officer, Tilde

Data for MT System Development

The main language resources for MT system development are:

Parallel corpora

Monolingual corpora



Various lexical resources

Data for MT System Development

Crawling typically focuses on:

- Mining parallel data
- Mining monolingual data

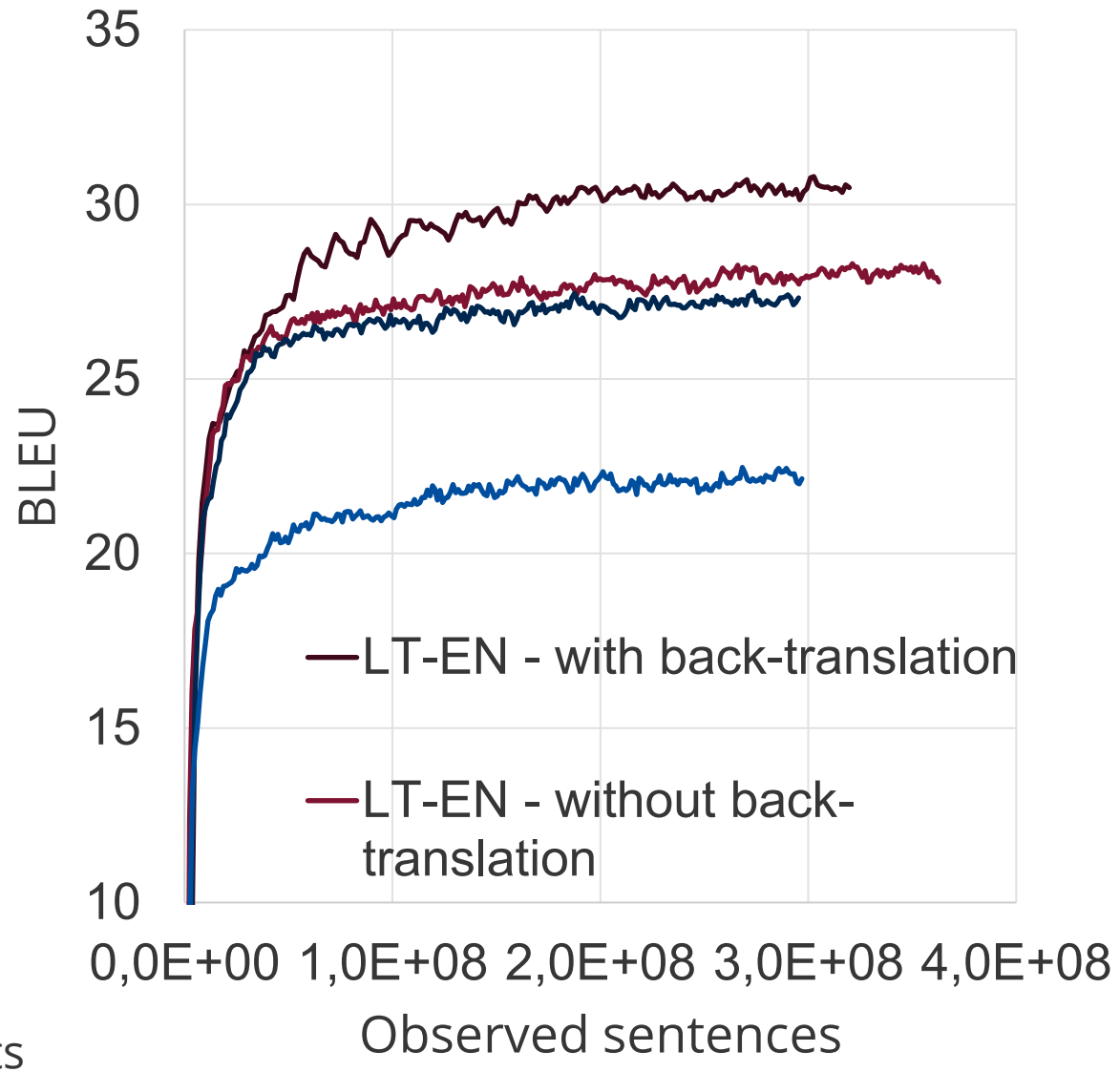
- Monolingual data in NMT are used for back-translation
- Back-translation is a method that allows to acquire synthetic parallel training data using a reverse MT system and target-side monolingual data

Monolingual Data for Back-Translation

- Back-translation helps:
  - Adjusting to the required style
  - Disambiguating domain-specific terminology
  - Improving noise-robustness of models
- **Back-translation is not magic** - it does not improve out-of-vocabulary word/phrase translation quality

Back-Translation can Boost MT Quality

Domain of the monolingual data has to match the domain of the text that will be translated.

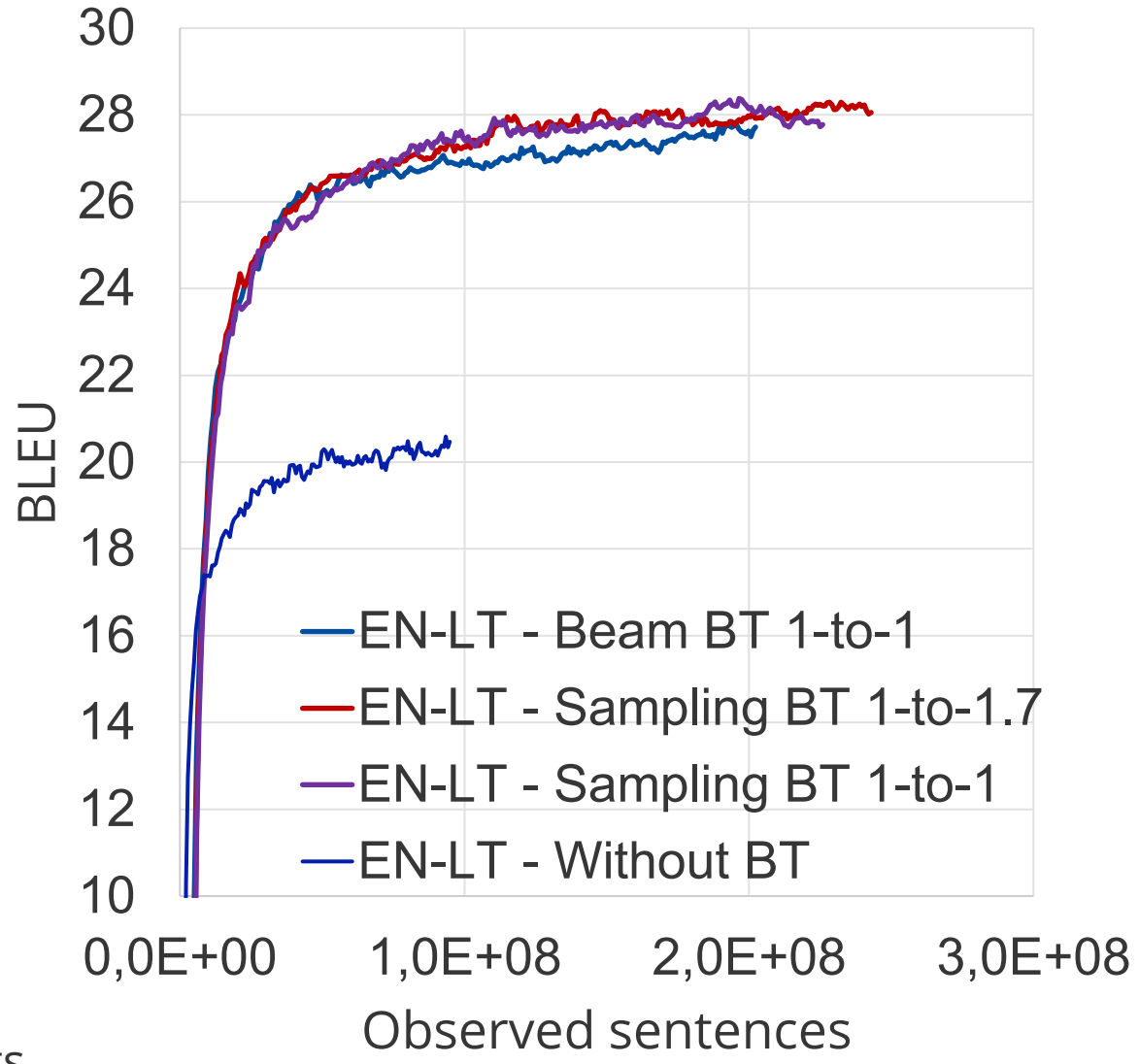


* Source of the graph: Tilde's WMT 2019 experiments

Back-Translation can Boost MT Quality

The data have to be of sufficient quantities to achieve highest quality increase

- For beam search - approximately the same amount as parallel data
- For sampling – the same or more



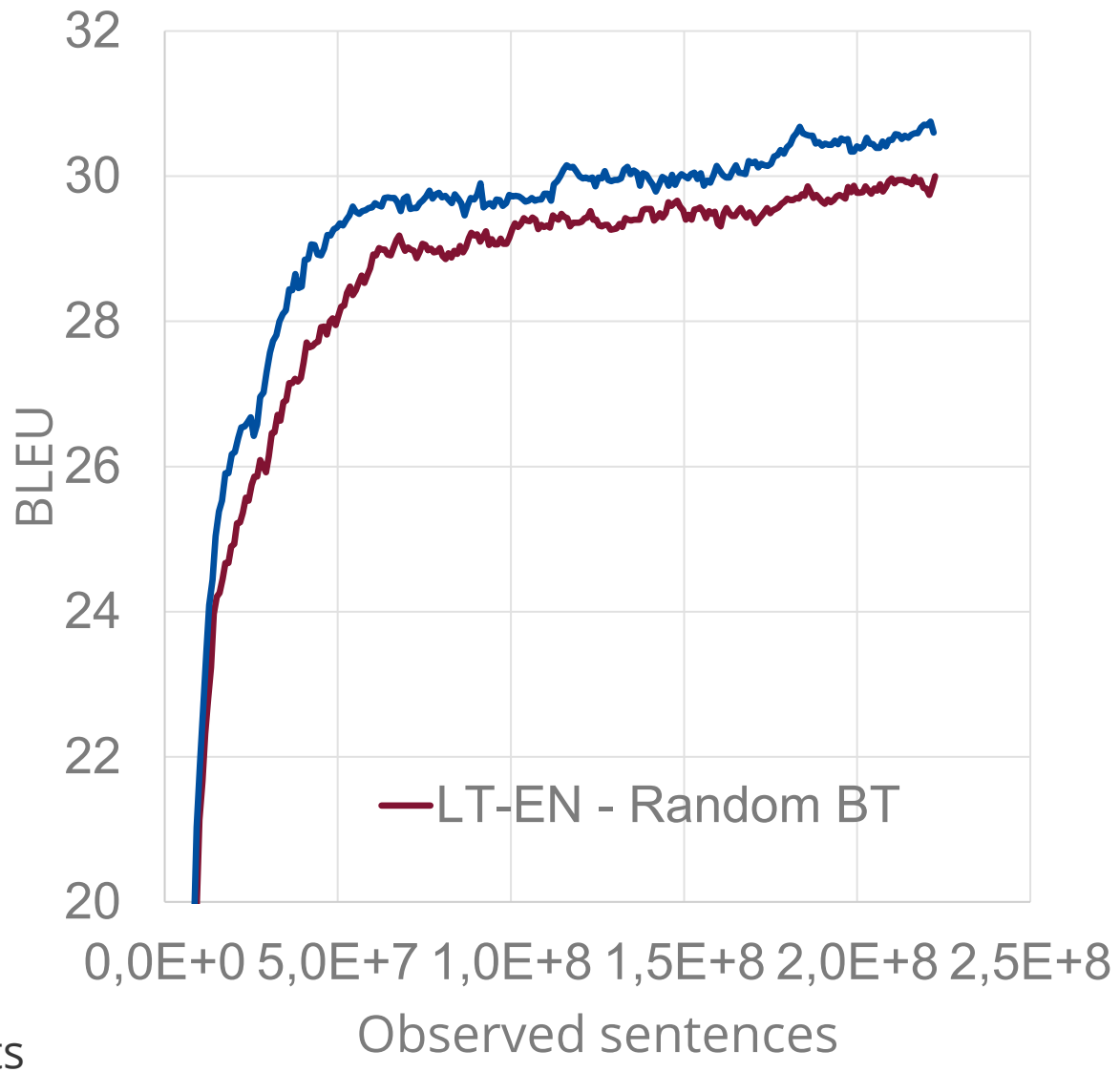
* Source of the graph: Tilde's WMT 2019 experiments

Back-Translation can Boost MT Quality

It is important that the (target language monolingual) data are source-domain adherent

- I.e., represent what users may want to write in the source language

However, if data are scarce, some data may be better than no data



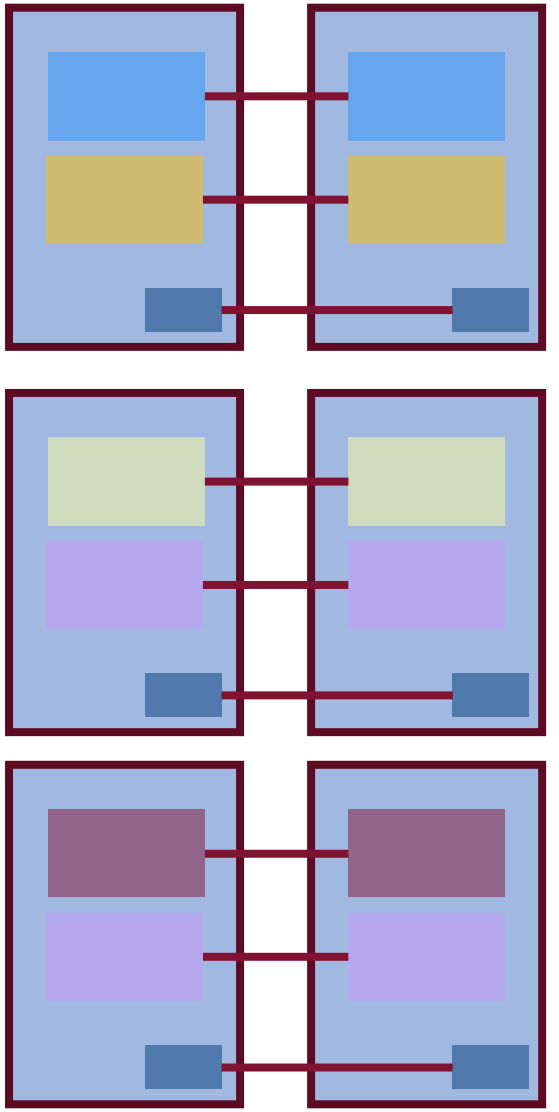
* Source of the graph: Tilde's WMT 2019 experiments

Parallel Data from the Web

Two main directions of parallel data mining are typically explored – focused or broad mining

Focused parallel data mining

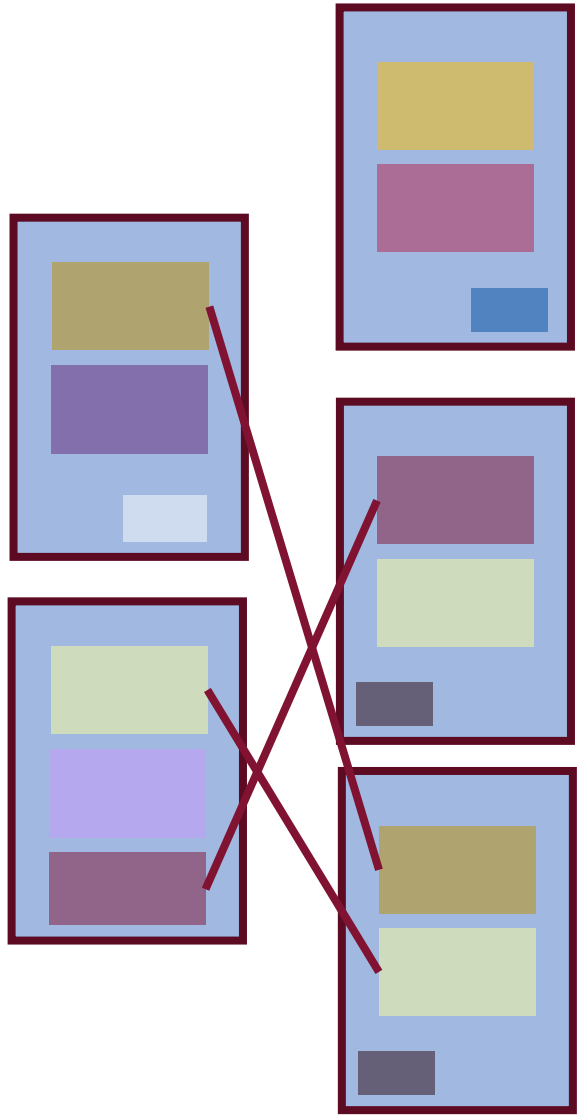
- Mining is performed for known web sites to contain parallel data
- Assumptions can be used to improve mining, e.g.:
  - Are corresponding documents identifiable?
  - Where is the main content to be mined located in documents?
- Higher (almost guaranteed) quality with smaller efforts



Parallel Data from the Web

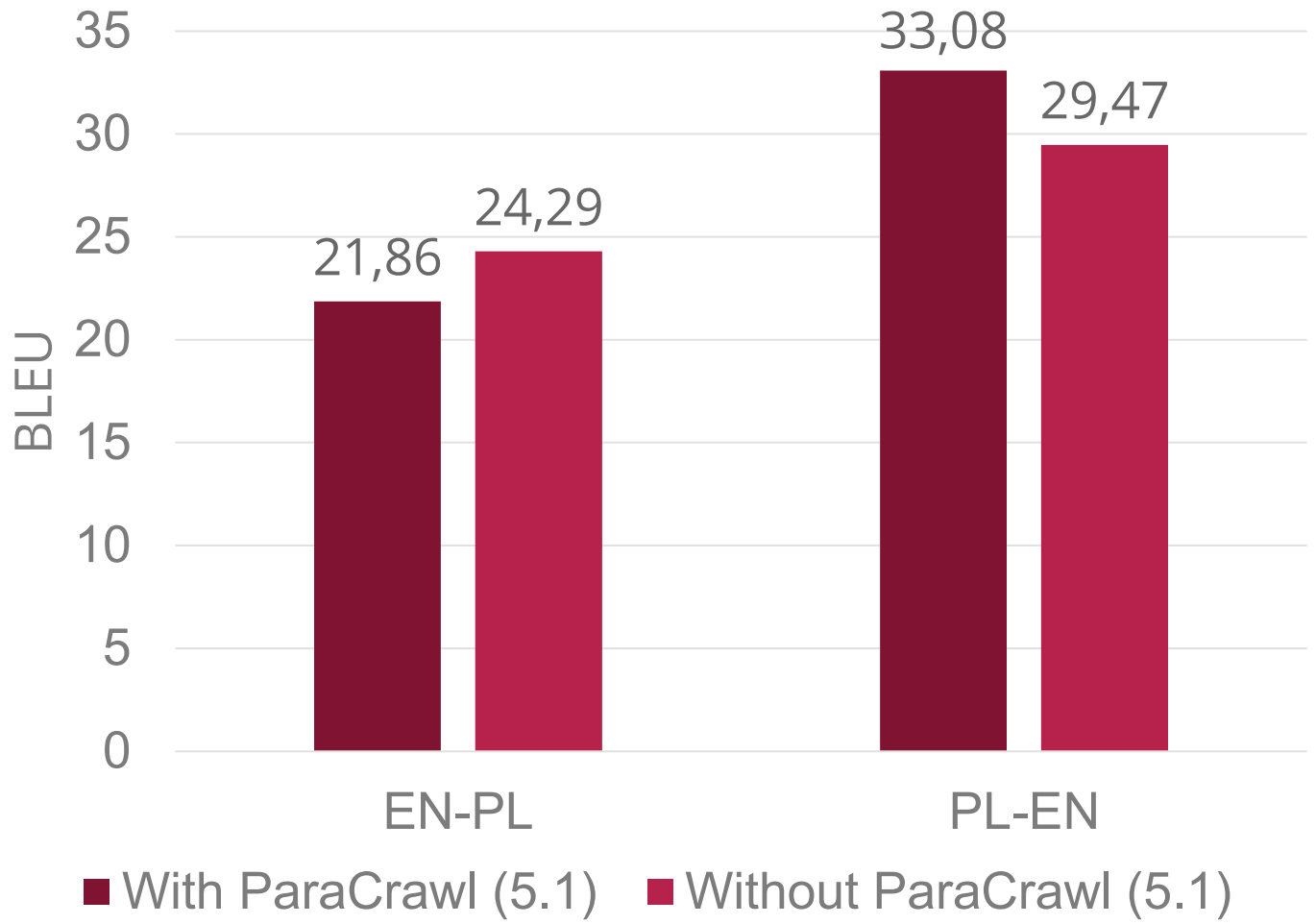
Broad parallel data mining

- The whole web is a potential source of parallel data
- Requires document alignment and more sophisticated sentence alignment methods
- Minimal to no assumptions on parallel data presence in documents can be made
- Quality depends heavily on the tools applied in the process



Parallel Data from the Web

- Parallel data from the broad parallel data mining processes may contain noise
- Noise can hinder training high quality MT systems
- It is important to evaluate quality of the crawled data and know when it is safe to use the data



* Source of the chart: Tilde's WMT 2020 experiments

Thank you!

Time for questions