PROMT Systems for WMT 2019 Shared Translation Task

Alexander Molchanov, Statistical and Neural MT team lead, PROMT LLC

Data

Human
- WMT
- OPUS
- Private Data

NMT

Back-translations
- News Crawl
- Wikipedia dumps

ER 43.6M sentence pairs 20M + 20M sentence pairs
EG / GE 51.8M sentence pairs 25M + 25M sentence pairs

Data Filtering

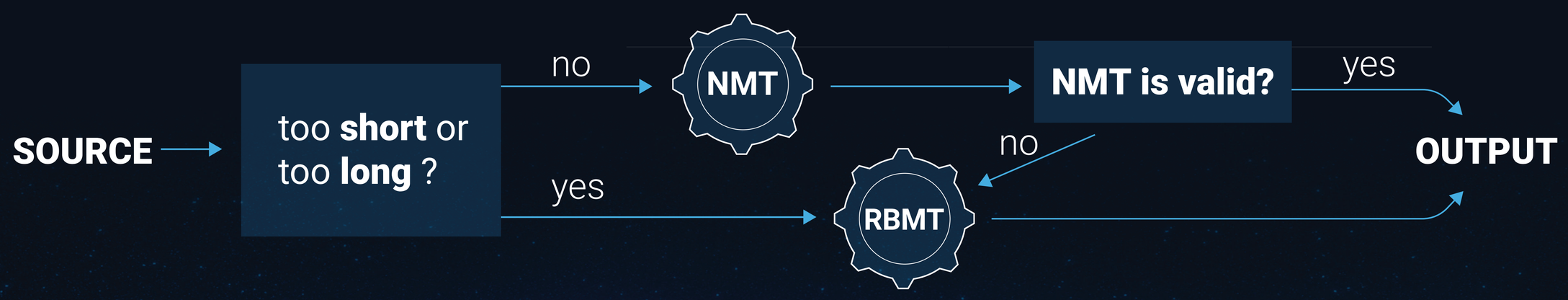
Basic filtering → Deduplication → Language detection → Parallel segments filtering (in-house classifier) → Filtering with language models

Hybrid System Architecture

NMT
- **Transformer, MarianNMT**
- **BPE** • 30K (English) + 40K (Russian) for ER • 40K for EG/GE
- **Case Feature (as a separate token)**
  World Championships 2017: Neil Black praises Scottish members of Team GB
  world <C> championships <C> 2017 : neil <C> black <C> pra@@ ises scottish <C> members of team <C> gb <U>
- **Fine-tuning on selected Data + News Crawl**

RBMT
- **Short / long sentences**
- **NMT validation** • Source-output length ratio • Output language verification • Named entities (numbers, emails etc) verification

backoff



Results

ER English-Russian
Second cluster

EG English-German
Second cluster

GE German-English
Second cluster

PROMT.COM