

Context in Machine Translation Challenge and data set

Jaakko Väyrynen
Timo Honkela
Marcus Dobrinkat

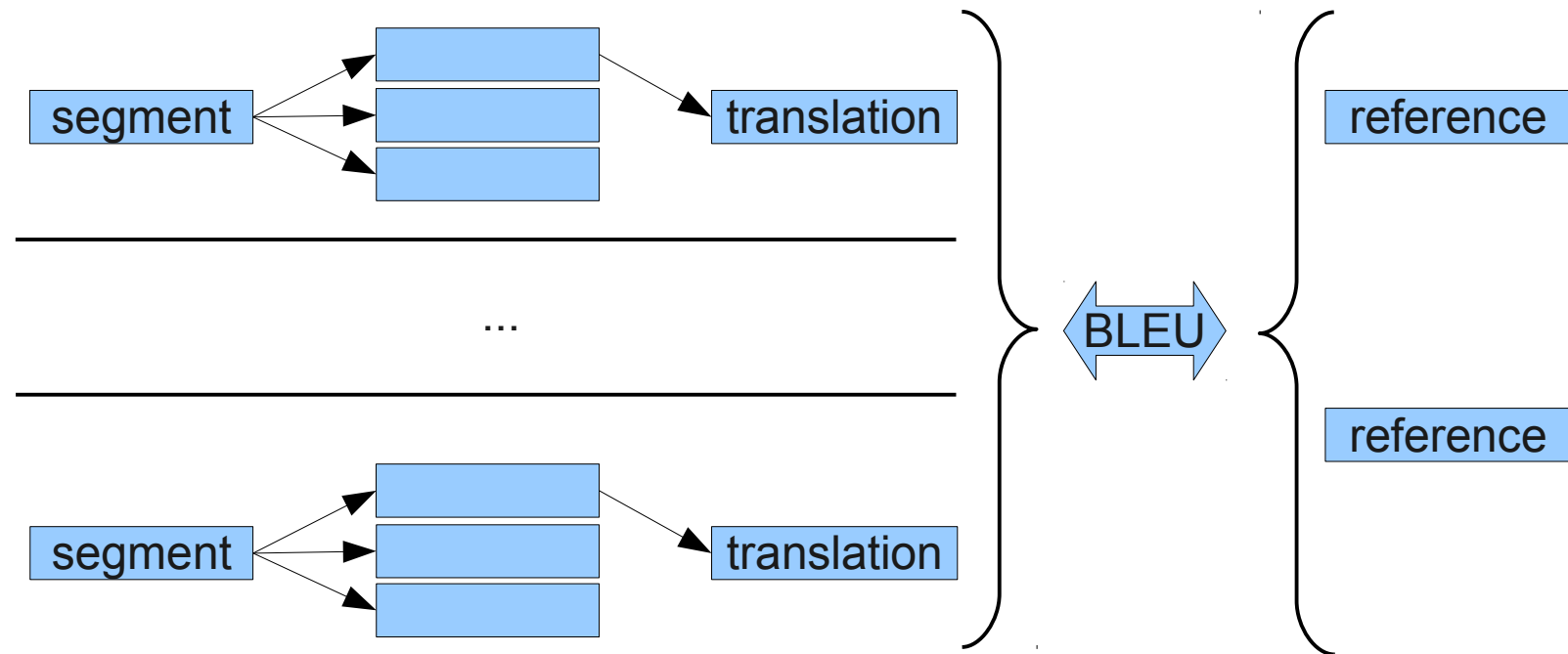
META-NET Workshop: Context in MT
Jun 14 2011

Objectives

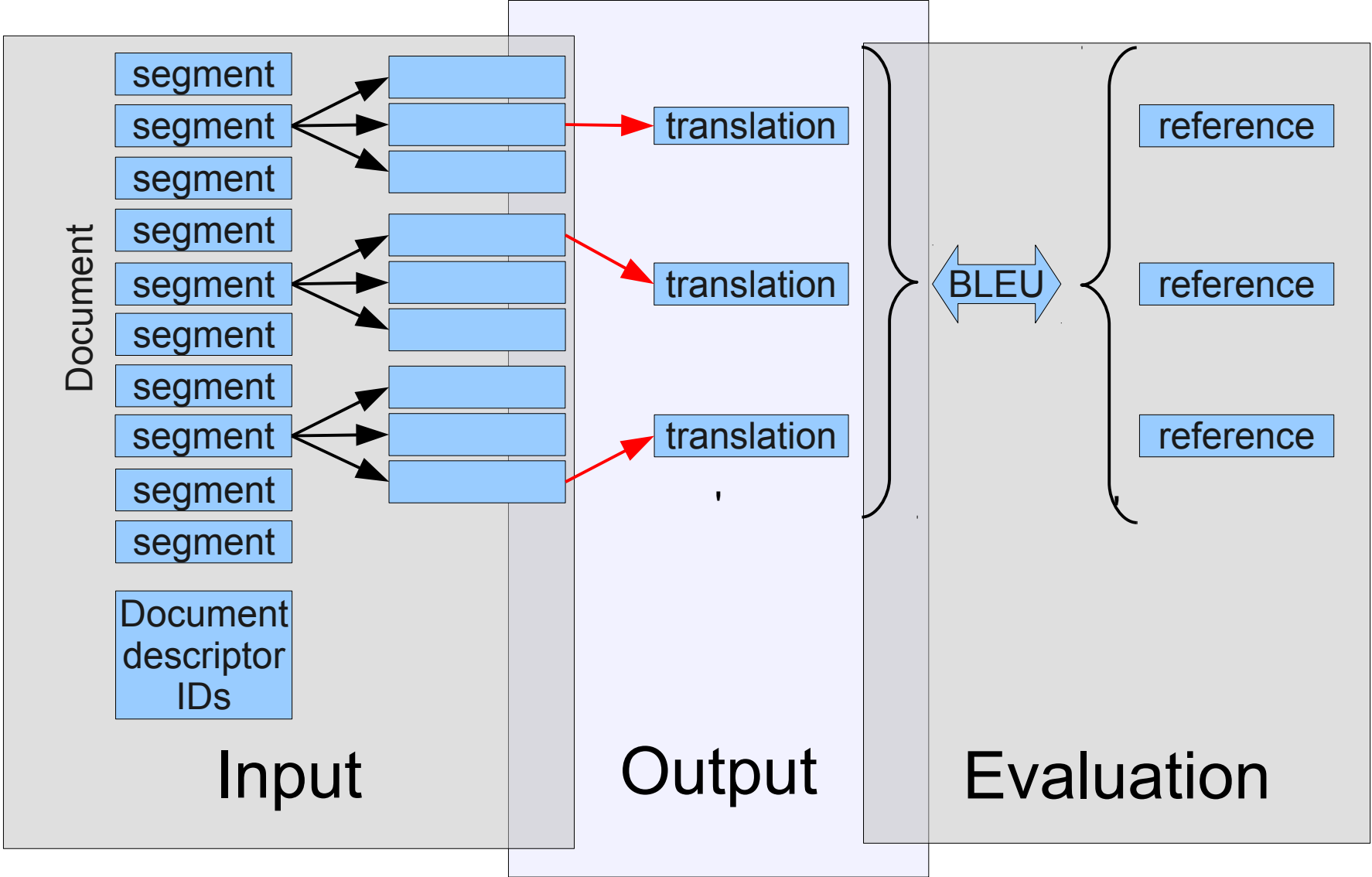
- To study the use of textual context in translation
- Can the context improve translations?
- Foster exchange between MT and ML fields



Translation without context



Challenge: take context into account



Challenge data set generation

- JRC Acquis as the original parallel corpus
 - English-Finnish and Greek-French
- MT system training ~800k segments, ~15M words
 - 4 MT systems: DFKI, FBK, LIMSI, RWTH
- ~130k segments translated by each MT system

Challenge data set statistics

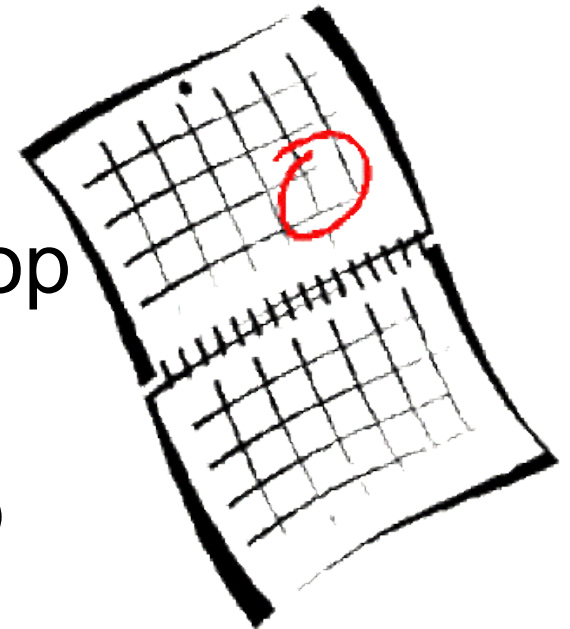
- Training set (released now)
 - ~5,000 documents,
 - ~100k segments with 100-best lists for 4 SMT systems
 - ~80k additional contextual segments
- Test set (released later) composed from
 - ~500 documents
 - ~10k segments with 100-best lists for 4 SMT systems
 - ~10k additional contextual segments

Challenge task

- Input
 - Document context
 - Document descriptors
 - N-best list of translations
- Output
 - **Choose best translation from N-best list**
 - (Document-ID, Sentence-ID, Translation-ID) integer triplets
- Evaluation
 - BLEU between selected translations and reference translations

Challenge timetable

- ICANN 2011, Espoo, Finland
 - Workshop and challenge data set released
- Three months before final workshop
 - Test set released
- Two months before final workshop
 - Submissions due
- ICANN 2012, Lausanne, Switzerland
 - Workshop and challenge results



Summary

- Interesting data set
- Task is not to translate, but to choose best translation from a list
- All exploitations of context are welcome!



Collaborators

- Special thanks to META-NET partners

