NPFL087 Statistical MT

Course Structure and Requirements

Ondřej Bojar

February 23, 2024

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Course Outline

1. Metrics of MT Quality.
2. Approaches to MT. SMT, PBMT, NMT, NP-hardness.
3. NMT (Seq2seq, Attention. Transformer). Neural Monkey.
4. Parallel texts. Sentence and word alignment. hunalign, GIZA++.
5. PBMT: Phrase Extraction, Decoding, MERT. Moses.
6. Morphology in MT. Factors or segmenting, data or linguistics.
7. Syntax in SMT (constituency, dependency, deep).
8. Syntax in NMT (soft constraints/multitask, network structure).
9. Towards Understanding: Word and Sentence Representations.
10. Advanced: Multi-Lingual MT. Multi-Task Training. Chef's Tricks.
11. **Project presentations.**

Related Classes

Informal prerequisites:

- NPFL125 Introduction to Language Technologies
- NPFL070 Language Data Resources

Recommended:

- NPFL114 Deep Learning
- *NPFL140 Large Language Models*

Course Materials

Slides:

`https://ufal.mff.cuni.cz/courses/npfl087`

Videolectures & Wiki of SMT:

`http://mttalks.ufal.ms.mff.cuni.cz/`

Books and others:

- Ondřej Bojar: Čeština a strojový překlad. ÚFAL, 2012.
- Philipp Koehn: Statistical Machine Translation. Cambridge University Press, 2009. Slides: `http://statmt.org/book/`
  Neural Machine Translation:
  `https://arxiv.org/abs/1709.07809`
  `http://mt-class.org/jhu/assets/nmt-book.pdf`

Other Good Sources

- http://mt-class.org/ (UEDIN is updated to NMT.)
- CMU (Graham Neubig) class:
  http://phontron.com/class/mtandseq2seq2017/
- http://www.deeplearningbook.org/
  by Goodfellow, Bengio, and Courville.

Grading

Key requirements:

- Work on a project (alone or in a group of two to three).
- Present project results (~30-minute talk).
- Write a report (~4-page scientific paper).

Contributions to the grade:

- 10% participation and homework (possibly CodEx exercises),
- 30% written exam,
- 50% project report,
- 10% project presentation.

Final Grade: $\geq$50% good, $\geq$70% very good, $\geq$90% excellent

ELITR Demo Videos

- https://elitr.eu/elitr-complementing-interpreters-and-starting-to-take-notes-for-yo
  ELITR final blogpost.
- http://ufallab.ms.mff.cuni.cz/~bojar/elitr/720p.mp4
  Martin Popel talking Czech; short snippets
- https://elitr.eu/a-tireless-interpreter/
  Very short demo for Open Doors Day.

Project Suggestions

http://tinyurl.com/npfl087-2024-project-ideas