



NLPL ACTIVITY B

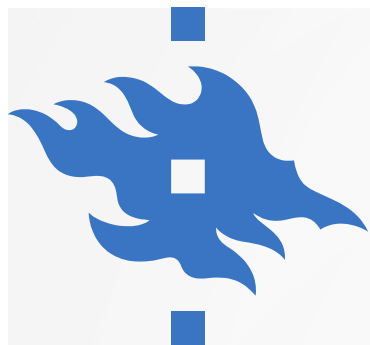
MACHINE TRANSLATION

Yves Scherrer



GOALS

1. Provide modules for **statistical machine translation** tools
 - Word alignment tools
 - Moses SMT pipeline
2. Provide modules for **neural machine translation** tools
 - HNMT, based on Theano
 - Others, depending on evolution of the field
3. Provide common training and testing **datasets**
 - WMT 2017 news translation task
 - IWSLT 2017 spoken language translation task
4. Provide **pre-trained models**
 - Helsinki WMT 2017 submissions
 - Pretrained SMT models from University of Edinburgh



CURRENT STATUS

1. Provide modules for **statistical machine translation** tools
 - Word alignment tools
 - Moses SMT pipeline
2. Provide modules for **neural machine translation** tools
 - HNMT, based on Theano
 - Others, depending on evolution of the field
3. Provide common training and testing **datasets**
 - WMT 2017 news translation task
 - IWSLT 2017 spoken language translation task
4. Provide **pre-trained models**
 - Helsinki WMT 2017 submissions
 - Pretrained SMT models from University of Edinburgh





STATISTICAL MACHINE TRANSLATION

- SMT pipeline installed on Taito and Abel:

Activity	Module Name/Version	Description	System	Install Date
B	moses/mmt-mvp-v0.12.1-2739-gdc42bcb	Moses SMT system, including GIZA++, MGIZA, fast_align	Taito	7/2017
B	moses/4.0-65c75ff	Moses SMT system Release 4.0, including GIZA++, MGIZA, fast_align, SALM	Taito, Abel	11/2017
B	efmaral/0.1_2017_07_20	efmaral and eflomal word alignment tools	Taito	7/2017
B	efmaral/0.1_2017_11_24	efmaral and eflomal word alignment tools	Taito, Abel	11/2017

- Moses Version 4.0
 - Including GIZA++, MGIZA, fast_align, SALM
- Efmaraal
 - Robert Östling's efficient word alignment tools efmaraal and eflomal



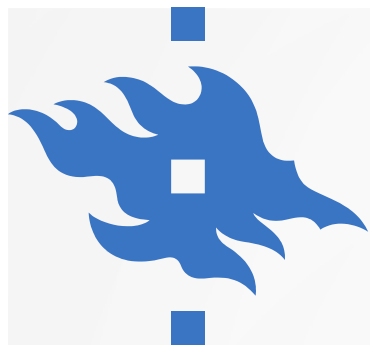
STATISTICAL MACHINE TRANSLATION

- More information on how to load and use these modules:
 - <http://wiki.npl.eu/index.php/Translation/home>
- More information about where the module files are installed (e.g., in case you need absolute paths):
 - module help mooses
- Challenge: Find the combination of versions that succeed in compiling Moses...
 - Taito: gcc/4.9.3 intelmpi/5.1.1 boost/1.58 python-env/3.4.5
 - Abel: gcc/6.1.0 boost/1.60.0 cmake/3.7.1 python3/3.5.0



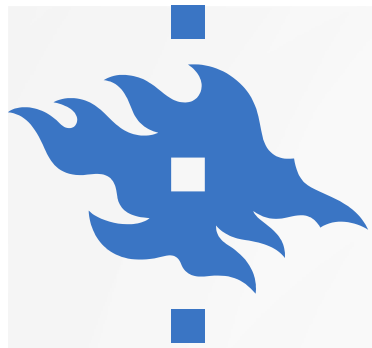
MT DATASETS

- **/proj[ects]/data/translation/iwslt17**
 - <https://sites.google.com/site/iwsltevaluation2017/data-provided>
- **/proj[ects]/data/translation/wmt17news**
 - <http://www.statmt.org/wmt17/translation-task.html> (parallel only)
- **/proj[ects]/data/translation/wmt17news_helsinki**
 - Preprocessed data and additional backtranslations used for the Helsinki WMT17 submissions (mainly EN-FI, EN-LV)
- Available on Taito + Abel, no documentation yet on Wiki
- To be updated with 2018 datasets as they become available



NEURAL MACHINE TRANSLATION

- **HNMT**
 - Requires Theano, but Theano is discontinued
 - Taito has a python-env/3.4.5-ml module with Theano included
 - No equivalent on Abel, system-wide Theano not going to happen...
- **Other NMT toolkits?**
 - OpenNMT-py (requires PyTorch)
 - Nematus (requires Theano, TensorFlow)
 - Tensor2Tensor (requires TensorFlow)
 - Marian



PRE-TRAINED MODELS

- Helsinki WMT 2017 submissions
 - EN-FI, EN-LV, possibly EN-ZH/ZH-EN
- Pretrained SMT models from University of Edinburgh