The University of Helsinki submissions to the WMT18 news task

Alessandro Raganato, Yves Scherrer, Tommi Nieminen, Arvi Hurskainen and Jörg Tiedemann
University of Helsinki

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

ACADEMY OF FINLAND

Main submissions

Language pairs:
- English<->Finnish
- English<->Estonian

Transformer architecture (Vaswani et al., 2017):
- base version, 6 layers, 8 attention heads, 512 dim. word embedding, etc.
- OpenNMT-py framework (Klein et al., 2017)

Domain labeling (Tars and Fishel, 2018):
- we added a domain label to each input sentence, according to the data source. For example, each sentence from the Europarl corpus was prepended with the <EUROPARL> label.

Subword units (Sennrich et al., 2016):
- joint BPE vocabulary of 37 000 units for each language pair

Back-translation (Sennrich et al., 2016):
- back-translated data from a SMT system (Tiedemann et al., 2016)
- back-translated data from a RNN-based system (Östling et al., 2017)

For English–>Finnish, our experiments also include:
- Rule-based system
- SMT system
- RNN-based system
- NMT system making use of a morphological analyzer and generator

English -> Finnish

Training:
- Europarl
- ParaCrawl
- Rapid
- WMT 2015 test and dev sets
- 5.5M backtranslated sentences from Finnish news2014 and news2016 corpora with a SMT system
- 5.5M backtranslated sentences from Finnish news2014 and news2017 corpora with a NMT system

Validation:
- WMT 2016 and 2017 test sets

NMT with morphological analysis and generation:
- Transformer-type NMT model trained with the Marian NMT framework (Junczys-Dowmunt et al., 2018).
- **Two-Step approach:**
  1. The Finnish corpus is analyzed with a morphological analyzer (FINTWOL by Lingsoft Inc.) and segmented into a sequence of interleaved lemmas and morphological tags.
  2. The output of the model is converted into surface forms in a separate, deterministic post-processing step.

Rule-based MT:
- Updated version of the Hurskainen and Tiedemann (2017) system, improving mainly the translation of:
- Locative expressions
- Proper names and acronyms
- Subject and object
- Comparative and superlative forms

Experiments

Results in terms of **BLEU-cased** score.

	Et→En	En→Et
HY-NMT Baseline	21.6	16.7
+Label	20.3	17.6
+Back	**26.5**	–
+Label +Back	25.4	**21.8***
+Back +Synth	**26.5***	–
+Label +Back +Synth	25.0	21.0
HY-NMT Multilingual	–	–
+Label	26.4	20.8

	Fi→En	En→Fi
Transformer +Label	19.8	15.3
Transformer +Back +Label	**23.3***	**17.8***
Multilingual +Back +Label	20.6	14.9
TwoStep +Back	–	14.5*
Seq2Seq +Back +Label	–	12.1
HY-SMT +Back	–	10.5*
HY-AH (rule based)	–	6.4*

Multilingual sub-track

We trained a multilingual model with data coming from three languages, English, Finnish and Estonian and then fine-tuned on a single language pair:
- Language labels to indicate the target language coupled with the domain labels (Johnson et al., 2016)
- 50 000 joint BPE units, to cover the three languages

We also generated **synthetic English–Estonian data by pivoting through Finnish**:
1. Train a character-level seq2seq system for Finnish-to-Estonian
2. Translate the Finnish side of the parallel English–Finnish corpus to Estonian
3. Combine the Estonian and English parts of the corpus and use also this dataset to train the final system

Finnish -> English

Training:
- Europarl
- ParaCrawl
- Rapid
- WMT 2015 test and dev sets
- 2M backtranslated sentences from Finnish news2015 corpora with a SMT system
- 6.7M backtranslated sentences from Finnish news2007-news2017 corpora with a NMT system

Validation:
- WMT 2016 and 2017 test sets

English <-> Estonian

English -> Estonian
Training:
- Europarl
- ParaCrawl
- Rapid
- 6.3M backtranslated sentences from BigEst corpus with a NMT system

Validation:
- WMT 2018 dev sets

Estonian -> English
Training:
- Europarl
- ParaCrawl
- Rapid
- 5.2M backtranslated sentences from English news2007-news2017 corpora with a NMT system

Validation:
- WMT 2018 dev sets

Statistics

Number of training sentences, with and without back-translation (Back) and synthetic data (Synth).

	Parallel	+Back	+Back +Synth
Et → En	2,178,025	7,356,697	8,942,157
En → Et	2,178,025	8,435,413	10,020,873
Fi → En	3,136,265	11,918,402	–
En → Fi	3,136,265	14,198,188	–

Shared first place in the manual evaluation for
English -> Finnish
Finnish -> English