Group Projects: Machine Translation

Matthias Huck, Alexander Fraser

LMU Munich

23 May 2017

Automatic Post-editing (APE)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

TOPIC: **Automatic Post-editing for Machine Translation**

- Post-editing is the correction of errors in machine-translated content, and typically done by humans with the purpose of bringing error-prone MT output to publishable quality.
- Can automatic post-editing (APE) fix errors in MT output?
- Build an APE system for the WMT16 APE shared task. Use any method you like, and any publicly available software and tools. Run your APE system and analyze the results.

Sources:

- `http://www.statmt.org/wmt16/ape-task.html`
- Section 7 of: Bojar et al., Findings of the 2016 Conference on Machine Translation. WMT 2016.
- Three Automatic Post-editing Shared Task papers (system descriptions) from WMT 2016: see

  `http://www.statmt.org/wmt16/papers.html`

Romanian→German Machine Translation

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

TOPIC: **Phrase-based MT from Romanian into German**

- Some language pairs are overrepresented in machine translation research, e.g. many combinations with English as either the source or target language. “Exotic” combinations such as Romanian→German are rarely explored.
- Build a phrase-based machine translation system that translates from Romanian into German. Start using the WIT3 corpus and the Moses toolkit. Which problems do you face?
- Can you utilize additional corpora? Would *pivot translation* or *lightly-supervised training* be useful?

Sources:

- *Moses* SMT toolkit: `http://www.statmt.org/moses/`
- WIT3: `https://wit3.fbk.eu/mt.php?release=2012-02-plain`
- *tokro* (de)tokenizer for Romanian:
  `https://perso.limsi.fr/aufrant/software/tokro`

Generating German Compounds

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

TOPIC: **An Exploration of Target-side Compound Splitting**

- When translating from German into English, splitting German compounds on the source side is common practice.
- Investigate whether compound splitting is useful on the target side in phrase-based translation from English into German.
- Using data from the IWSLT MT track, build an English→German baseline w/o compound splitting. Then build a system with split compounds on the target side. Can you implement a feature in Moses that would avoid stray compound parts in decoding?

Sources:

- Paper: Koehn & Knight, Empirical Methods for Compound Splitting. EACL 2003.
- `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/compound-splitter.perl`
- `https://sites.google.com/site/iwsltevaluation2015/mt-track`

Questions?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thank you for your attention

Matthias Huck

mhuck@cis.lmu.de