Text Normalization based on Statistical Machine Translation and Internet User Support
Tim Schlippe, Chenfei Zhu, Jan Gebhardt, Tanja Schultz
tim.schlippe@kit.edu

1. Overview

Introduction
- Text normalization system generation can be time-consuming
- Construction with the support of internet users (crowdsourcing):
  1. Based on text normalized by users and original text, statistical machine translation (SMT) models are created
  2. These SMT models are applied to "translate" original into normalized text
- Everybody who can speak and write the target language can build the text normalization system due to the simple self-explanatory user interface and the automatic generation of the SMT models
- Annotation of training data can be performed in parallel by many users

Goals of this paper
- Compare:
  - How does the performance of SMT evolve over amount of training data?
  - How can we modify SMT to reduce time and effort?

2. Experimental Setup

Pre-Normalization
- LI-rule by our Rapid Language Adaptation Toolkit (RLAT)

Language-specific normalization by Internet users
- User is provided with a simple readme file that explains how to normalize the sentences
- Web-based user interface for text normalization
- Keep the effort for the users low:
  - No use of sentences with more than 30 tokens to avoid horizontal scrolling
  - Sentences to normalize are displayed twice: The upper line shows the non-normalized sentence, the lower line is editable

Evaluation
- Compare quality (BLEU, edit dist.) of 1k output sentences derived from SMT, LI-rule and LS-rule to quality of text normalized by native speakers
- Create 3-gram LMs from hypotheses (1k sentences) and compare their perplexities (PPLs) on 500 manually normalized test sentences
  (Note: The 500 manually normalized test sentences have a PPL of 240.95 on a LM created with 928M tokens but a PPL of 444.05 on the LM trained with only 1k sentences normalized by native speakers.)

3. Experiments and Results

Performance for crawled French text over training data
- BLEU, Levenshtein edit dist., PPL

Duration of text normalization by native speaker
- French speaker took almost 11h for 1k sentences spread over 3 days
- Saturation of performance starts after the first 450 sentences

System improvements
- Rule-based number normalization
- Language-spec. rule-based with statistical phrase-based post-editing (hybrid):
  - Non-normalized text
  - Rule-based LI-rule
  - SMT output text
  - Human output text
  - Language-specific rule-based (LS-rule)
  - Rule LS norm.

4. Conclusion and Future Work

Conclusion
- A crowdsourcing approach for SMT-based language-specific text normalization:
  - Native speakers deliver resources to build norm. systems by editing text in our web interface
  - Results of SMT close to LS-rule, hybrid better, close to human
  - Close to optimal performance achieved after about 5 hours manual annotation (450 sentences)
  - Parallelization of annotation work to many users is supported by web interface

Future Work
- Investigating other languages
- Enhancements to further reduce time and effort