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 (Schlippe et al., 2010)
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

1. Analyze efficiency for different languages
2. Embed English annotation process for training data in MTurk
3. Reduce user effort by iterative text normalization generation and application

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:
  - Sentences to normalize are displayed twice:
    - The upper line shows the non-normalized sentence, the lower line is editable

Evaluation

- Compare quality (Edit distance) of output sentences (1k for DE and EN, 500 for BG) derived from SMT, LI-rule, LS-rule and hybrid to quality of text normalized by native speakers
- Create 3-gram LMs from hypotheses (1k for DE and EN, 500 for BG) and compare their perplexities (PPLs) on manually normalized test sentences (500 for DE and EN, 100 for BG)

Experimental Setup

- **Hybrid** SMT / Edit interface and the automatic generation of the text

Results of statistical machine translation based text normalization with crowdsourcing (Schlippe et al., 2010)

- Spelling correction
- Removal of spam
- Removal of HTML, Java script and non-protocol characters
- Separation of punctuation marks which are not in context

Future Work

- A crowdsourcing approach for SMT-based language-specific text normalization:
  - Native speakers deliver resources to build normalization systems by editing text in our web interface
  - Results of SMT which were close to LS-rule for French, even outperformed LS-rule for Bulgarian, English and German, hybrid better, close to human
  - Annotation process for English training data could be realized fast and at low cost with MTurk, however need for methods to detect and reject Turkers’ spam
  - Reduction of editing effort in the annotation process for training data with iterative-SMT and iterative-hybrid