Statistical Machine Translation based Text Normalization with Crowdsourcing

Tim Schlippe, Chenfei Zhu, Daniel Lemcke, Tanja Schultz
tim.schlippe@kit.edu

Cognitive Systems Lab

Karlsruhe Institute of Technology

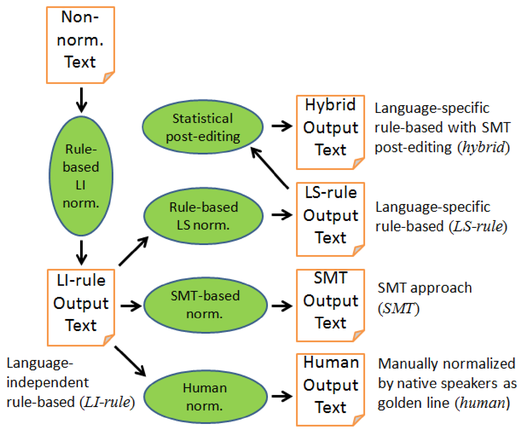
1. Overview

Introduction

- Text normalization system generation can be time-comsuming
- 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

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



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:
  - 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)

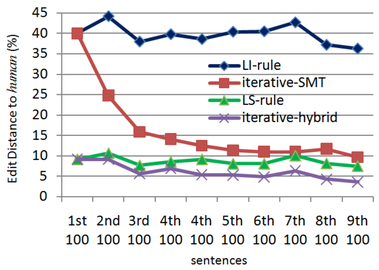
Web-based user interface for text normalization

Language-independent Text Normalization *(LI-rule)*
1. Removal of HTML, Java script and non-text parts.
2. Removal of sentences containing more than 30% numbers.
3. Removal of empty lines.
4. Removal of sentences longer than 30 tokens.
5. Separation of punctuation marks which are not in context with numbers and short strings (might be abbreviations).
6. Case normalization based on statistics.
Language-specific Text Normalization *(LS-rule)*
1. Removal of characters not occuring in the target language.
2. Replacement of abbreviations with their long forms.
3. Number normalization (dates, times, ordinal and cardinal numbers, etc.).
4. Case norm. by revising statistically normalized forms.
5. Removal of remaining punctuation marks.

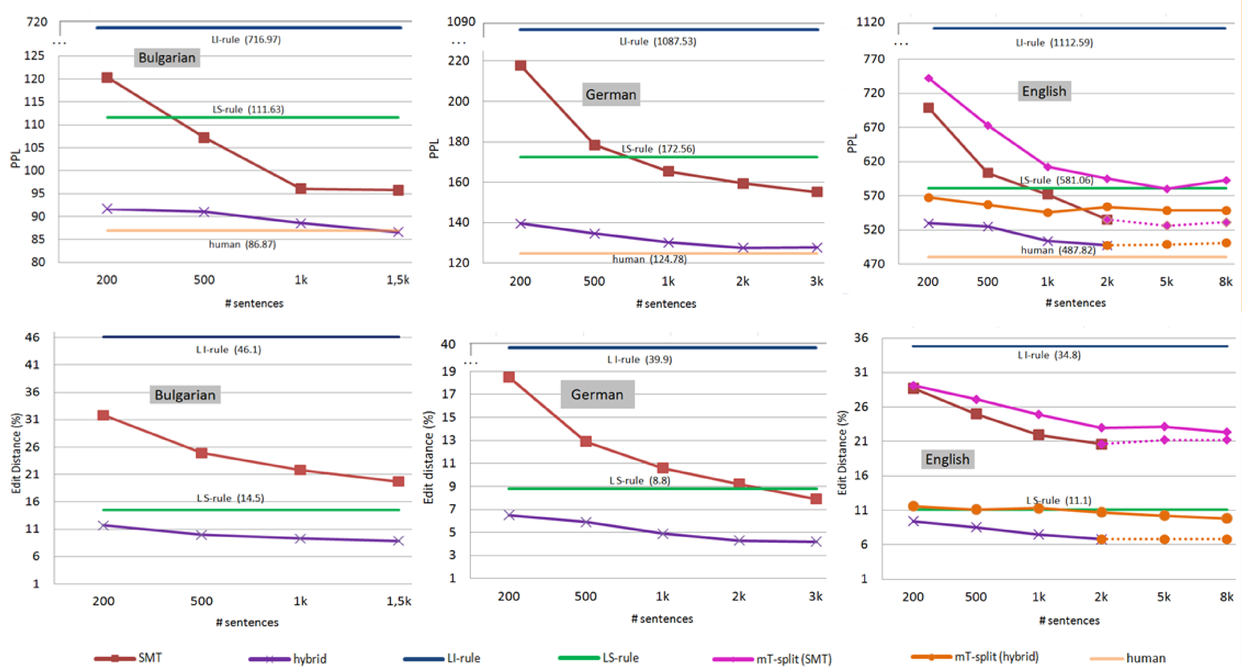
Language-independent and -specific text normalization

3. Experiments and Results

Edit Distance reduction with iterative SMT / hybrid



Performance over amount of training data



4. Conclusion and 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*