Adaptation for SMT

Prioritize translation candidates that are most relevant to current task

The problem with provenance

- Not an intrinsic text property, requires document's meta-information
- If not available: manual labeling is labor-intensive and can be arbitrary

Translation Model Adaptation Using Genre-Revealing Text Features

Marlies van der Wees Arianna Bisazza Christof Monz
Informatics Institute, University of Amsterdam

Translation model adaptation with a vector space modeling (VSM) approach

Two test sets with two genres:
- Gen&Topic*: two test sets with two genres:
- NIST 2008+2009:
  - User-generated comments (UG)
  - Newswire (NW)
- Genre adaptation by
  - Distinguishing genre from provenance and topic
  - Using genre-revealing text features for translation model adaptation
  - Eliminating the need for manual sub-corpus labels

Provenance features:

- Automatically indicators of genre can replace dependency on manual domain labels with automatic measures of genre

Automatic features:

- LDA-inferred topics
- Count of exclamation marks, question marks, repeating punctuation, emoticons, numbers, first & second person pronouns

Increased translation consistency

- Phrases with identical translations for each occurrence in a single document
- Higher consistency for genre-adapted system

Effect of source and target-side features

- Source-side genre features
- Target-side genre features

Increased translation consistency

- +4.2
- +2.7
- +0.1
- +2.6

Manual versus automatic features

- Manual provenance labels
- Automatic features (genre+LDA)

Results

Improved translation performance

- Automatic indicators of genre can replace manual sub-corpus labels
- Best system with automatic features: genre+LDA

Projection across languages

- Similar performance for feature values computed on Arabic or English side of the bitext

Conclusions

We address genre adaptation by

- Distinguishing genre from provenance and topic
- Using genre-revealing text features for translation model adaptation
- Eliminating the need for manual sub-corpus labels

The proposed method

- Improves translation quality over a competitive baseline
- Exploits features that can be projected across languages
- Increases document-level translation consistency

** Following Chen et al., Vector Space Model for Adaptation in Statistical Machine Translation, 2013