Modernizing historical Dutch: the UU system
Marijn Schraagen, Feike Dietz, Marjo van Koppen, Kalliopi Zervanou
Utrecht University, The Netherlands

Summary

• Goal: modernize 17th century Dutch text to allow use of modern NLP resources and tools
• Method: combine expert rules, translation pairs from aligned parallel text, existing SMT frameworks
• Data: parallel translation of the Bible, 1637/1888
• Results: the proposed vocabulary-based method shows promising results on an in-domain test set, performance is impaired for unrelated domains
• Future work: refinement of current method, shift to character-based methods

Introduction

• Modernization of spelling and grammar allows use of tools for modern Dutch on historical text
• Note: some features (e.g., negative concord and case marking) are lost after modernization
• Quantitative methods can be trained using parallel text, e.g., diachronic translations of the Bible

Method

The Bible text is split into a training set (32235 sentences) and a test set (5000 sentences). The following steps are incrementally applied, with associated BLEU scores [1] on the test set ($n = 4$):

- (BLEU: 0.134) No translation.
- (0.507) Baseline: construct 1-to-1 translation lexicon on training data, using sentences of equal length.
- (0.530) Perform alignment to handle sentences of unequal length, extract additional translation pairs.
- (0.581) Compile a set of manual modernization rules.
- (0.600) Construct many-to-1 translation lexicon using aligned sentences.
- (0.619) Use POS-information for already modernized words to choose the right alternative for historical words.
- (0.627) Compile rules to address punctuation differences between Bible translations.


- (0.597) Moses with basic training settings.
- (0.616) Apply MERT tuning.
- (0.639) Post-processing of incorrect output of trained Moses capitalization model.
- (0.644) Manual modernization rules on Moses output.
- (0.647) Moses with manual rules, multi-alignment, and POS patterns.
- (0.653) As above, with punctuation rules.

Discussion and future work

• Vocabulary-based method not highly suitable for unrelated texts
• Diachronic differences: e.g., en translated as negation, but used in later texts only as conjunction
• Overtranslation, i.e., arguably correct results not present in the reference translation
  - ofte-of, der-van de, hare-hun, 't-het, zo als-zoals, hebbe-heb, ...
• The current method can be refined for in-domain texts
• Character-based methods may offer wider applicability

References


Acknowledgements

This work is financed by the Netherlands Organisation for Scientific Research (NWO), grant 360-78-020.