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

1637:	Ende	het	gout	deses	lants is goet
1888:	En	het	goud	van dit	land is goed

And the gold of that land is good

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 ($n = 4$):

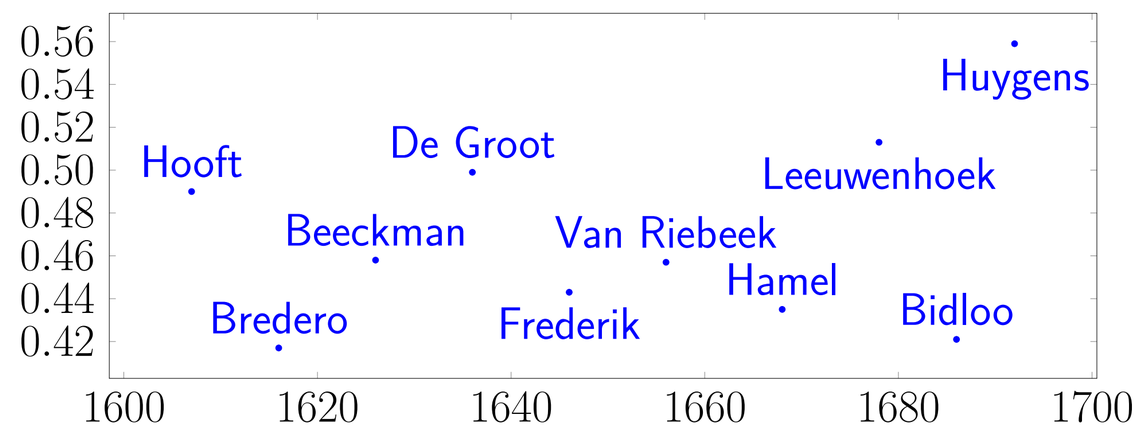
- (BLEU: 0.134) No translation.
- (0.507) Baseline: construct 1-to-1 translation lexicon on training data, using sentences of equal length.
- (0.530) Perform alignment to handle sentences of unequal length, extract additional translation pairs.
  - ◇ custom alignment algorithm using fixed anchor tokens
- (0.581) Compile a set of manual modernization rules.
  - ◇ e.g., strip case markers
- (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.
  - ◇ *haer* + V → *hen*
  - ◇ *haer* + N → *hun*
- Selection for many-to-1 and POS rules: hill-climbing optimization on BLEU score on training data.
- (0.627) Compile rules to address punctuation differences between Bible translations.

Additional approach: train the Moses SMT toolkit [2] on word level, using 2000 development sentences for minimum error rate training. Afterwards, apply steps as above.

- (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.

CLIN Shared Task test set results

- Additional phonetic rewriting rules to address OOV issues



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

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[2] Philipp Köhn, Hieu Hoang, Alexandra Birch, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

Acknowledgements

NWO Netherlands Organisation for Scientific Research

Universiteit Utrecht

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