Converting Seventeenth-Century Dutch to Modern Dutch

Erik Tjong Kim Sang
Meertens Institute Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

1 Introduction

Nederlab[1] is a five-year research project (2013-2017) which aims at making available digital versions of many historical Dutch texts, with linguistic annotations. Because of the size of the corpus, we rely on automatic tools for creating the linguistic annotations. However, the tools were developed to process twentieth century Dutch newspaper texts while our corpus contains texts from different domains written in historical variants of Dutch. Because of this difference, the linguistic annotation tools perform poorly on our corpus.

Rather than retraining our set of linguistic annotation tools, we aim at *translating* our historical corpus texts to modern Dutch, to which we then can successfully apply the tools. In this way, we only have to do the domain adaptation step once, and not adapt every tool that we want to use.

In this paper, we present experiments with translating seventeenth-century Dutch to modern Dutch. We test a state-of-the-art part-of-speech tagger for modern Dutch on the texts: Frog (Bosch et al. 2007). We describe two sets of experiments: one in which a state-of-the-art machine translation system translates the texts and one in which we use lexicon-based techniques for the translation. After a section on related work and a data section, the next two sections deal with these experiments. In the last section, we conclude.

2 Related work

Archer et al. (2015) present VARD (VARiant Detector), a tool for normalizing historical corpora. They apply it for translating text in Early Modern English to modern English so that it can be processed by standard natural language processing tools. G. Schneider, Lehmann and P. Schneider (2015) also use VARD for normalizing Early and Late Modern English so that it can be processed by a modern syntactic parser. For the language Dutch, the most relevant related work is the work of Reynaert, who has studied automatic spelling correction (Reynaert 2005) and the correction of OCR-ed text (Reynaert 2014). Translation of text to standard variants is also important for modern texts, like used on social media (Kaufmann and Kalita 2010).

[1] nederlab.nl

3 Data

For test purposes, we use two small seventeenth-century texts with gold standard base part-of-speech tags set by Hupkes (2014) and a third text without part-of-speech tags. The first text is a section of the Dutch Statenvertaling bible from the year 1637 (1370 tokens). The second is a section of the ship journal of Willem Ysbrants Bontekoe published in 1646 (1565 tokens). The third text is a part of the book *Schat der Gesontheyt* by Johan van Beverwijck from 1663 (Koomen 2007) (1612 tokens).

4 Translation with a machine translation system

In the first translation experiment, we used the state-of-the-art machine learning system Moses (Koehn 2015) for translating seventeenth-century Dutch to modern Dutch. Moses needed example texts for performing the translation task. We trained it with two versions of the Statenvertaling, one from the year 1637 (Sijs 2008) and one from 1888 (Theologencommissie 1999). Moses needed additional source text for tuning. or this purpose, we used the first chapter of the *Schat der Gesontheyt* book (Koomen 2007). We tested Moses with a part of the second chapter of this book.

We used BLEU (Papineni et al. 2002) for evaluating Moses. The translation made by Moses obtained a BLEU score of 0.283. Since we did not know what this number meant, we also computed BLUE scores for the input text and a second manual translation. These obtained 0.124 and 0.345 respectively. So the text of Moses is closer to modern Dutch than to the original seventeenth-century Dutch. More importantly, the BLEU score of Moses was better than that of any of the other translation approaches tested for this paper.

However, in order to obtain a good translation, Moses deleted words and inserted words. This makes it difficult to link word annotations in the modern text back to the original words in the historical text. For this reason, we concluded that Moses is not useful for our purposes. What we need, is a word-by-word translation.

5 Translation approaches based on lexicons

After the evaluation of a full machine translation system, we tested two alternative approaches based of lexicons. First, we tested a system which replaced historical words with their modern lemma taken from the Integrated Language Bank (GTB) offered by the Dutch Institute for Lexicography (INL 2007). In that dictionary, we looked up all words that occurred five times or more in the Statenvertaling bible translation from 1637 (8563 words). The performance of the Frog tagger improved when historical words from this lexicon were replaced by their modern lemma: from 68.2% for the original text to 82.0% on the Bontekoe text and from 63.7% to 90.4% on the bible text. However, by replacing words by lemmas, information which is needed for some linguistic analysis tasks, is lost.

Next, we derived a translation lexicon from the Statenvertaling bible, using K-vec (Fung and Church 1994). In order to improve the precision of the method, we only used sentence pairs from the bibles that

had the same number of words and we excluded unique word pairs from the output. After translating the Bontekoe text with the resulting 6150-word lexicon[2], obtained a part-of-speech accuracy of 82.1%, which is similar to that of the INL lexicon but without the disadvantage of having to work with lemmas.

6 Concluding remarks

We have evaluated three methods for improving accuracies of a modern part-of-speech tagger applied to seventeenth-century texts. The methods relied on translating the historical texts to a modern version which could then be processed by the tagger. Full machine translation proved to be unpractical, because it changed the number of words in the text, making it difficult to link the part-of-speech tags back to the original words. Translation with a manual lexicon had as a disadvantage that it produced lemmas, which are insufficient for some linguistic analysis tools. In our project Nederlab, we use the third method: translation with a learned lexicon, because it performs well and it does not have the disadvantage of working with lemmas.

References

Archer, D. et al. (2015) 'Guidelines for normising Early Modern English corpora: Decisions and justifications', *ICAME Journal* 39, DOI: 10.1515/icame-2015-0001.

Bosch, A. van den et al. (2007) 'An efficient memory-based morphosyntactic tagger and parser for Dutch', *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, pp. 99–114.

Fung, P. and Church, K. (1994) 'K-vec: A New Approach for Aligning Parallel Texts', *Proceedings of COLING94*, Kyoto, Japan, pp. 1096–1102.

Hupkes, D. (2014) *Semi-supervised training of part of speech taggers for historical Dutch using modern Dutch syntactic priors*, unpublished manuscript.

INL (2007) *Geintegreerde Taal-Bank (GTB)*, http://gtb.inl.nl/ Retrieved 27 October 2015.

Kaufmann, M. and Kalita, J. (2010) 'Syntactic normalization of twitter messages', *International conference on natural language processing (ICON)*, Kharagpur, India.

Koehn, P. (2015) *MOSES - Statistical Machine Translation System - User Manual and Code Guide*, 21 August, University of Edinburgh.

Koomen, N. (2007) *Van Beverwijck, Schat der Gesontheyt, 1663*, Volkoomen.nl.

Papineni, K. et al. (2002) 'BLEU: a method for Automatic Evaluation of Machine Translation', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318.

Reynaert, M. (2005) *Text-Induced Spelling Correction*, PhD Thesis, Tilburg University.

— (2014) 'On OCR ground truths and OCR post-correction gold standards, tools and formats', *Proceedings of DATeCH'14*, ACM, New York.

Schneider, G., Lehmann, H. M. and Schneider, P. (2015) 'Parsing Early and Late Modern English corpora', *Digital Scholarship in the Humanities* 30.3.

Sijs, N. van der, ed. (2008) *Biblia, dat is De gantsche H. Schrifture (Statenvertaling 1637)*, DBNL: Digitale Bibliotheek voor de Nederlandse Letteren.

Theologencommissie, ed. (1999) *Statenvertaling Jongbloededitie 1888*, Statenvertaling.net.

[2]Testing the learned lexicon on the bible test set would give an inflated test score since the bible was used as training material.