Converting seventeenth century Dutch to modern Dutch

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

16 November 2015

The problem

In the project Nederlab, we want to offer searching historical texts by requiring specific word classes

For example: give me all sentences with verbs preceded by the Dutch word *voor*

Since we have hundreds of thousands of historical texts, the word classes of the words in the text will be indentified automatically by software tools (part-of-speech taggers)

However...

However...

Modern part-of-speech taggers are unable to process historical texts well:

> *In ’t Jaer ons Heeren 1618, den 28. December, ben ick, Willem Ysbrantsz. Bontekoe van Hoorn, Tessel uytghevaren voor schipper, met het schip ghenaemt: Nieu-Hoorn*

Historical texts contain many words that are spelled differently, which increase the error rate of modern natural language processing tools, which have been trained on modern texts

Solution

Retrain the natural language processing tools

But retraining requires annotated historical texts which are difficult to create

Alternative: **translate the historical texts to modern Dutch**

For training this requires translated texts, which are already available and easier to create

Related research

Dieuwke Hupkes improved Dutch part-of-speech tagging of seventeenth century texts from 60% to 80-90% by translating to modern Dutch

Tessa Wijckmans works on normalizing the orthography of seventeenth century Dutch in the context of authorship attribution

Martin Reynaert is normalizing historical texts with TiCCL in order to achieve better natural language tool postprocessing

At the other end of the time spectrum, people are working on normalizing social media messages for similar reasons (Kaufmann and Kalita, 2010)

Methods for processing historical texts

1. do nothing
2. use crowd-sourcing for translating texts
3. use a part-of-speech tagger for historical Dutch
4. translate texts with machine translation
5. use a historical lexicon for translating texts
6. translate texts with a lexicon learned from examples
7. translate texts with rules learned from a lexicon

1: Do nothing: process text with modern tagger

We process two samples from texts with the MBTtagger which is part of the Frog package: the Bontekoe ship journal (1646) and the Statenvertaling bible (1637), both annotated by Dieuwke Hupkes

We register the percentage correct base tags:

Text	Size	POS
Bontekoe 1646	50 sentences; 1565 tokens	68.2%
Statenvertaling 1637	50 sentences; 1370 tokens	63.7%

These two scores will be used as baseline scores

2: Use crowd-sourcing for translating texts

The best method for translating the historical texts to modern Dutch is to have humans perform the translation task

	POS	
Text	**Baseline**	**Human**
Bontekoe 1646	68.2%	88.8%
Statenvertaling 1637	63.7%	91.2%

This approach works well but it is time-consuming

We will use these scores as ceiling scores

3: Use a part-of-speech tagger for historical Dutch

Hans van Halteren developed Adelheid, a part-of-speech tagger for Middle Dutch (14th century). The tagger is available for testing online: adelheid.ruhosting.nl

We processed the two test texts with Adelheid:

		POS	
Text	**Baseline**	**Adelheid**	**Ceiling**
Bontekoe 1646	68.2%	71.4%	88.8%
Statenvertaling 1637	63.7%	82.9%	91.2%

4: Translating texts with machine translation

Moses is a state-of-the-art free machine translation system. We trained it with two versions of the Dutch Statenvertaling bible: 1637 and 1888

Problem: in the translation process Moses may insert, delete and reorder words

If we perform part-of-speech tagging on text translated by Moses, it is difficult to link the tags to the original historical words

We need a word-by-word translation

Evaluating Moses output

We evaluated the quality of the translations by Moses with the BLEU score, a statistical method for comparing word sequences in related sentences:

		BLEU	
Text	**Baseline**	**Moses**	**Ceiling**
Beverwijck 1663	0.12373	0.28254	0.34536

Moses performed very well, better than any other automatic translation method

However, it is difficult to install, slow to run and it is hard to link tags from translated text back to the original words

5: Use a historical lexicon for translating texts

The Dutch Institute of Lexicography (INL) has developed several lexicons of historical variants of Dutch

There is an online interface for historical Dutch words: sk.taalbanknederlands.inl.nl/LexiconService

The service returns modern lemmas for historical words

Furthermore, a query for one historical word may return several candidate lemmas

Evaluating translations with the INL lexicon

We built a translation lexicon for the words that appeared five times or more in the 1637 Statenvertaling bible

When there were several candidate lemmas, we selected the one that was equal to the historical word, and otherwise the one that was most frequent in the 1888 Statenvertaling bible

Choosing the best lemma

Historical variant	Candidate lemmas
ende	**en** (57642) einde (318) eend (0)
de	**de** (41141) doen (1658)
van	**van** (22251) vinden (160)
het	**het** (21009) hebben (3018)
den	**de** (41141) den (11521)
in	en (57642) **in** (14785)
hy	**hij** (9498) hei (0)
die	**die** (11676)
dat	die (11676) **dat** (9629)
tot	**tot** (10878) totten (0)

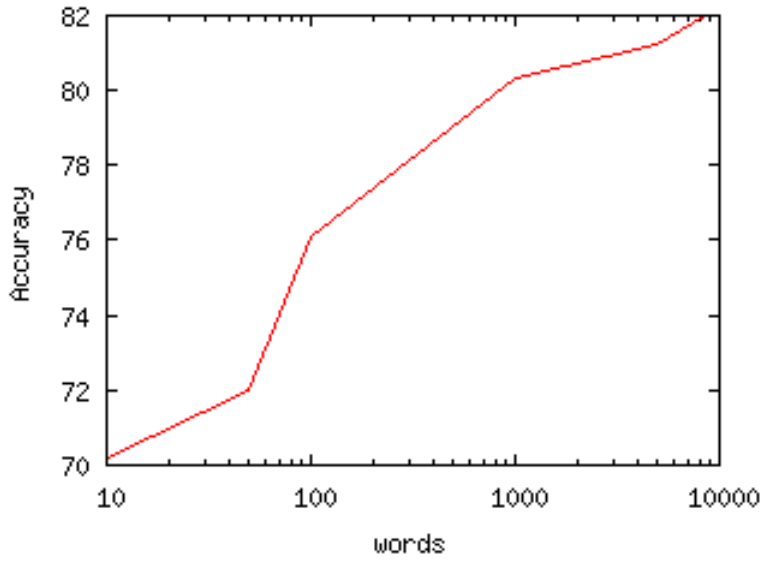
Evaluating translations with the INL lexicon

We built a translation lexicon for the words that appeared five times or more in the 1637 Statenvertaling bible

When there were several candidate lemmas, we selected the one that was equal to the historical word, and otherwise the one that was most frequent in the 1888 Statenvertaling bible

		POS	
Text	**Baseline**	**INL**	**Ceiling**
Bontekoe 1646	68.2%	82.0%	88.8%
Statenvertaling 1637	63.7%	90.4%	91.2%

Remarks on using the INL lexicon



The POS accuracies obtained by using this lexicon are excellent and the graph for the Bontekoe scores suggest further improvement is possible with a larger lexicon size

However, since the tags are estimated on lemmas rather than words, it will be difficult to obtain correct morphological information

6: Using a lexicon learned from parallel texts

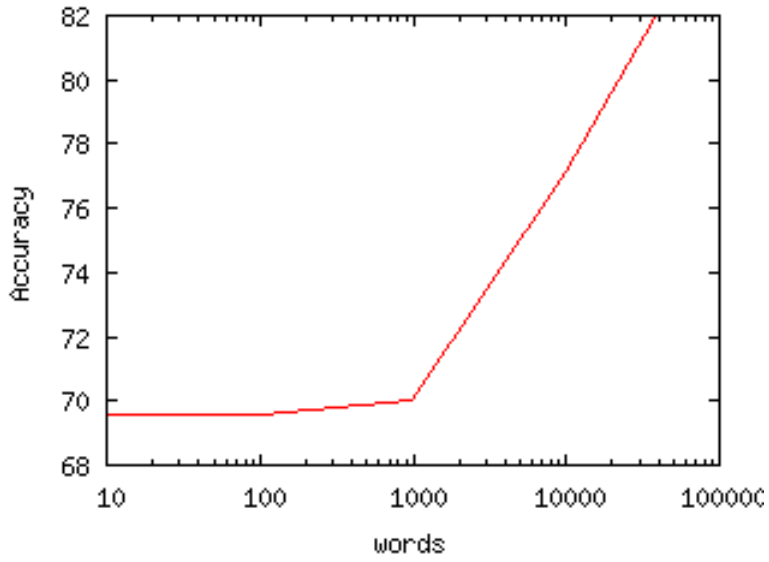
We learned a translation lexicon by comparing word positions in a parallel text: the Statenvertaling bible editions from 1637 and 1888

Words which frequently appeared in translated sentences were assumed to be translations of each other

		POS	
Text	**Baseline**	**Learned**	**Ceiling**
Bontekoe 1646	68.2%	81.9%	88.8%

(With the INL lexicon we obtained a POS accuracy of 82.0%)

Remarks on using a learned lexicon



For the Bontekoe text, the POS accuracy obtained with a learned lexicon is almost the same as for the INL lexicon and the score progression suggests there is room for improvement with larger lexicons

The learned lexicon leaves the morphological information intact

7: Learning morphological translation rules

Creating a modern version of a historical word often involves changing a few letters, like *aen* becomes *aan*

These rules can be created by an expert, or learned automatically, provided that we have a parallel lexicon available

We learned 86 rules from our learned lexicon

Examples of learned morphological rules

Frequency	Precision	Rule
895	0.903	y $\Rightarrow$ i
623	0.967	ae $\Rightarrow$ aa
346	0.989	uy $\Rightarrow$ ui
222	0.996	aen $\Rightarrow$ aan
221	0.978	hey $\Rightarrow$ hei
177	0.947	uyt $\Rightarrow$ uit
162	0.982	aer $\Rightarrow$ aar
150	0.993	$\wedge$uy $\Rightarrow$ ui
139	0.993	$\wedge$ae $\Rightarrow$ aa
107	0.930	ck\$ $\Rightarrow$ jk

7: Learning morphological translation rules

Creating a modern version of a historical word often involves changing a few letters, like *aen* becomes *aan*

These rules can be created by an expert, or learned automatically, provided that we have a parallel lexicon available

We learned 86 rules from our learned lexicon

		POS	
Text	**Baseline**	**Morphology**	**Ceiling**
Bontekoe 1646	68.2%	72.2%	88.8%

Summary of the Bontekoe scores

Method	POS accuracy	Comment
Baseline	68.2%	
Adelheid	71.4%	
Morphological rules	72.2%	
Learned lexicon	81.9%	
INL lexicon	82.0%	Looses morphology
Moses	??.?%	Impractical
Ceiling	88.8%	Requires annotation work

The Bontekoe scores in combination with our experiences suggest that the learned lexicon is the most suitable approach for Nederlab

Concluding remarks

We evaluated several methods for improving part-of-speech tagging for seventeenth century Dutch

The most suitable method is based on translating the historical words to modern variants with a learned lexicon, before doing the tagging with a modern part-of-speech tagger

The translation should be done word-by-word in order to allow for an easy linking of the tags to the historical text

Future work

1. Examining methods for increasing the learned lexicon
   (but with what data?)

2. Evaluating the learned lexicon on tags with morphological information
   (but on what data?)

3. Integration of the translation process in the Nederlab pipeline
   (expected problem: tokenization)

Corpora and lexicons

Statenvertaling edition 1637, source: dbnl.nl

Statenvertaling edition 1888, source: statenvertaling.net

Schat der Gesontheyt by Johan van Beverwijck (1663),
source: volkoomen.nl

100 manually tagged 17th century Dutch sentences,
source: Dieuwke Hupkes

INL lexicon,
source: sk.taalbanknederlands.inl.nl/LexiconService

References

Antal van den Bosch, Bertjan Busser, Walter Daelemans and Sander Canisius, An efficient memory-based morphosyntactic tagger and parser for Dutch. In Frank van Eynde, Peter Dirix, Ineke Schuurman and Vincent Vandeghinste (eds.), *Selected Papers of the 17th CLIN Meeting*, Leuven, Belgium, 2007.

Dieuwke Hupkes, *Part-of-Speech Tagging van asynchrone Bijbelteksten deel 2*. Manurscript, 10 October 2014 (in Dutch).

Max Kaufmann and Jugal Kalita, Syntactic normalization of twitter messages. In *International conference on natural language processing (ICON)*, Kharagpur, India. 2010.

Philipp Koehn, et al., Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL-2007*, ACL, 2007.

M. Rem and Hans van Halteren, *Tagging and Lemmatization Manual for the corpus van Reenen-Mulder and the Adelheid 1.0 Tagger-Lemmatizer*. Radboud University Nijmegen, 2011.

Martin Reynaert, TICCLops: Text-induced Corpus Clean-up as online processing system. In: *Proceedings of Coling 2014*, Dublin, Ireland, 2014.

THE END