Improving Part-of-Speech Tagging of Historical Text by First Translating to Modern Text

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

25 May 2016

Introduction

The project Nederlab aims at providing digital versions of historical texts written in the language Dutch

In order to make it easier to find relevant passages in the texts, syntactic annotations have been added to all sentences in the texts

Because of the large volumes of texts, these annotations have been created by software tools

The tools have been developed for processing twentieth-century text and perform poorly on historical text

The problem

Accuracy of assigning syntactic classes (part-of-speech tags) to text:

Twentieth-century text: 98.6%
Seventeenth-century text: 68.2%

Errors in the syntactic annotations make it harder to find relevant texts in the corpus

Solutions

There are two ways to improve the performance of the language tools on historical texts:

1. Build special-purpose tools for each relevant historical period
2. Translate the historical text to modern language before applying the tools

We chose the second method and compared four translation methods

1: Machine translation

1. Find a historical text and a related version in modern language
2. Have a tool (Moses) learn the relation between the two texts
3. Use the trained tool for translating other historical texts

Pro: the quality of the translation is good

Con: annotations assigned to the translation cannot be linked back to the original text: words had been deleted and inserted

2: Translation based on a historical lexicon

1. Find a lexicon with historical words and their modern equivalents
2. Look up each word of a historical text in the lexicon
3. Replace the historical word by its modern equivalent (choose the best one)

Pro: word annotations can easily be mapped back to the original text

Con: slightly worse quality of the translation and the translated words may have been simplified (for example: doing $\rightarrow$ do)

3: Translation based on a learned lexicon

1. Find a historical text and a related version in modern language
2. Have a tool learn the relations between words and their translations
3. Replace words in historical texts by their modern equivalents

Pro: word annotations can easily be mapped back to the original text and translated words are not simplified

Con: only words from the training texts can be translated

4: Translation based on learned rules

1. Find a relevant word-to-word lexicon, for example the learned lexicon
2. Have a tool learn the character relations between related words
3. Replace words in historical texts by their modern equivalents

Pro: with character replacement rules it is possible to translate word which were not present in the training lexicon

Con: you cannot translate words of which the translation is completely different

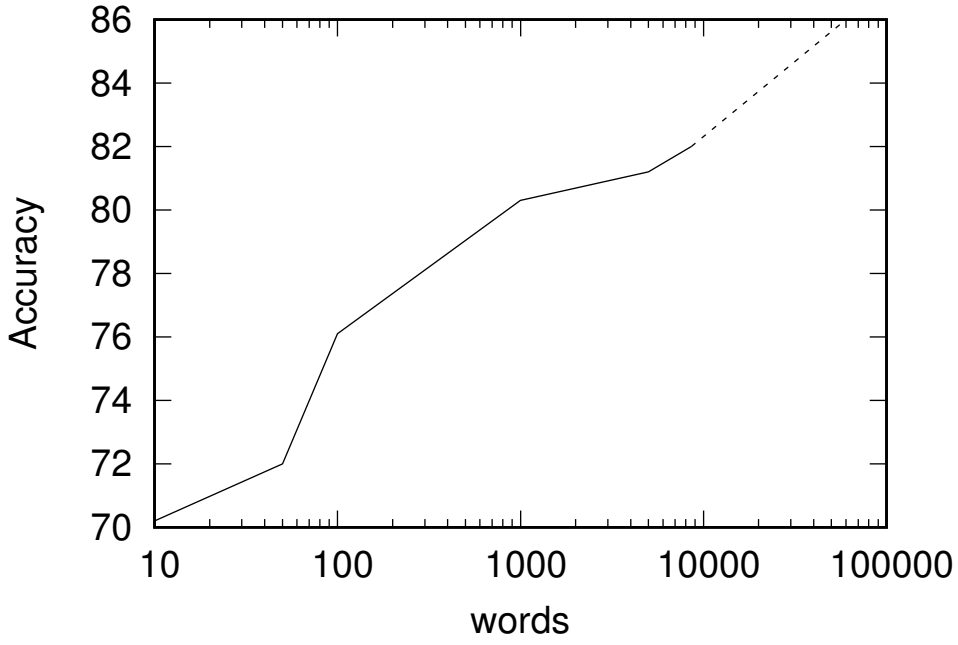
Examples of translation rules

y ⇒ i	groe**y**den ⇒ groe**i**den
ae ⇒ aa	**ae**ngesien ⇒ **aa**ngezien
uy ⇒ ui	gel**uy**t ⇒ gel**ui**d
aen ⇒ aan	**aen**gesien ⇒ **aan**gezien
hey ⇒ hei	gesont**hey**t ⇒ gezond**hei**d
uyt ⇒ uit	gel**uyt** ⇒ gel**uid**
aer ⇒ aar	w**aer** ⇒ w**aar**

Evaluation scores

Method	Translation	POS annotation
Modern tagger	0.124	68.2%
Historical tagger	0.124	71.4%
Manual translation	0.345	88.8%
Machine translation	0.283	NA
Historical lexicon	0.191	82.0%
Learned lexicon	0.219	82.1%
Orthographic rules	0.160	73.4%

Improving performances



Concluding remarks

We explored four methods for translating historical texts to modern text in order to improve the quality of automatic syntactic analysis

It was essential for the texts to be translated word-by-word

With the two lexicon-based translation methods we obtained the highest annotation qualities

Further performance improvements are possible, if the resources the lexicons were based on, are increased in size

THE END