The CLIN2017 Shared Task: Translating Historical Text

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

10 February 2017

The CLIN2017 Shared task

In several current research projects there is a demand for processing historical text with language annotation tools

However, because of language change, tools trained to process contemporary language perform badly on old texts

Building new tools for historical will take a lot of effort

Alternatively, we can translate older versions of language to modern versions before applying the tools

The CLIN2017 consist of automatically translating seventeenth-century Dutch text to modern (twenty-first-century) Dutch text

What does translation mean?

Reader: change every word that I do not understand

Computational linguist: change the words the tagger does not know

Historical linguist: leave the text alone! Just build a new analysis tool

We chose to replace all the words that did not occur in a standard lexicon with the required base syntactic category

We do not aim for semantic correctness

We do need translations with token-to-token alignments

Example translation

This seventeenth-century sentence contains many unknown words:

't Geslacht, de geboort, plaets, tydt, leven, ende wercken Van Karel van Mander, Schilder, en Poeet, Mitsgaders Zyn overlyden, ende begraeffenis.

Our target translation contains fewer unknown words:

't Geslacht , de geboorte , plaats , tijd , leven , ende werken van Karel van Mander , schilder , en poëet , mitsgaders zijn overlijden , ende begrafenis .

As a standard for modern Dutch we use the dictionary at `vandale.nl`

Data

As training data we used four editions of the Dutch Statenvertaling bible (1637, 1657, 1888 and 2010; 925,000 tokens) and a manually translated text by Steven Blankaart (1698; 1,000 words)

As test data we used 10 manually translated texts of about 1,000 words taken from different decades of the seventeenth century, and written by well-known authors (like Bredero, De Groot, Hooft, Huygens, Leeuwenhoeck and Van Riebeeck)

The texts originated from the Digital Library for Dutch Literature: `dbnl.nl`

Methods

Eight teams participated in the shared task. They used a variety of techniques for building translation systems:

- The trainable machine translation system Moses
- The word alignment tool MGIZA
- The Levenshtein distance for finding similar words
- Minimum Error Rate Training (MERT) for training
- Character machine translation
- Lexical rewrite rules
- Spelling normalization techniques

Results of the shared task

Rank	BLEU score	Team
1.	0.61022	Ljubljana / Zagreb / Geneva
2.	0.60665	Bochum
3.	0.59896	Helsinki / Uppsala
4.	0.56783	Amsterdam
5.	0.53752	Leuven
6.	0.46866	Utrecht (improved: 0.50329)
7.	0.45061	Groningen (improved: 0.51316)
8.	0.43011	Valencia
	0.33097	baseline

Effects on part-of-speech tagging quality

Rank	Accuracy	Unknown Words	Team
	0.875±0.026	0.093	ceiling (manual translation)
1.	0.856±0.029	0.148	Helsinki / Uppsala
2.	0.842±0.022	0.131	Ljubljana / Zagreb / Geneva
3.	0.836±0.027	0.129	Amsterdam
4.	0.825±0.025	0.133	Leuven
5.	0.824±0.030	0.085	Bochum
6.	0.812±0.030	0.223	Utrecht
7.	0.793±0.024	0.019	Groningen
	-	0.282	Valencia (no alignment)
	0.709±0.030	0.407	baseline

Error analysis of best Helsinki run (#2)

Word changes

16x	+1.000	ende→en
15x	+0.600	't→het
10x	+1.000	so→zo
10x	+1.000	ick→ik
7x	+1.000	sal→zal
6x	+1.000	zyn→zijn
6x	+1.000	soo→zo
6x	+1.000	hy→hij
6x	+0.000	daer→daar
413x	+0.278	...others...

Part-of-speech changes

37x	+0.595	SPEC→N
36x	+0.556	N→WW
29x	+0.414	N→ADJ
22x	+0.818	N→VNW
17x	+1.000	SPEC→VG
17x	+0.118	SPEC→ADJ
15x	+0.933	SPEC→VNW
15x	+0.133	WW→N
14x	+1.000	SPEC→BW
103x	+0.621	...others...

Concluding remarks

We have organized a shared task on translating historical text to modern text for improving automatic syntactic analysis

The team of Ljubljana/Zagreb/Geneva obtained the best score for translating text (BLEU score)

The team from Helsinki/Uppsala obtained the best score for part-of-speech accuracy

Results can be found on `ifarm.nl/clin2017st`

After the lunch you can find 5 shared task system posters in the poster session

THE END