Advancements in the Application of Neural Machine Translation for the Processing of Historical Documents

Miguel Domingo

midobal@prhlt.upv.es

Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València

ValgrAI Scientific Council Forum 2023

Ciutat Politècnica de la Innovació, July 4, 2023

Outline

Outline

Motivation

- Historical documents are an important part of our cultural heritage.
- However, due to their linguistic characteristics they are mostly limited to scholars.
- The linguistic variation in historical documents has always been a concern for scholars in humanities.
- Human language evolves with the passage of time.
- Orthography changes depending on the author and time period.
- e.g., the data in LALME[1] indicate 45 different forms recorded for the pronoun *it*, 64 for the pronoun *she* and more than 500 for the preposition *through*.

[1] Linguistic Atlas of Late Medieval English.

Outline

Language Modernization

Introduction

Goal: make historical documents more accessible to a general audience.

Language Modernization

Introduction

Goal: make historical documents more accessible to a general audience.

Original	**Modernized**
To be, or not to be? That is the question Whether tis nobler in the mind to suffer The slings and arrows of outrageous fortune, Or to take arms against a sea of troubles, And, by opposing, end them?	The question is: is it better to be alive or dead? Is it nobler to put up with all the nasty things that luck throws your way, or to fight against all those troubles by simply putting an end to them once and for all?

Language Modernization

Approaches

- Statistical machine translation (SMT).
- Neural machine translation (NMT).
  - Recurrent neural networks with long short-term memory units (LSTM).
  - Transformer.
- NMT enriched with modern documents.
  - Synthetic data generated using modern documents.

Language Modernization

Work in progress

- Adapting pre-trained large language models for this task.
- We are working with multilingual models:
  - mBART 50[2], which covers 50 languages.
  - mT5[3], which covers 100 languages.

[2]Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., ... & Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401.

[3]Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

Spelling Normalization

Introduction

Goal: achieve an orthography consistency by adapting a document's spelling to modern standards.

Spelling Normalization

Introduction

Goal: achieve an orthography consistency by adapting a document's spelling to modern standards.

Original	**Normalized**
"Nunca fuera ca**u**allero	"Nunca fuera ca**b**allero
de damas ta**m**bien seru**u**ido,	de damas ta**n** bien ser**v**ido,
como fuera don Qui**x**ote	como fuera don Qui**j**ote
quando de su aldea vino:	cuando de su aldea vino:
don**z**ellas cura**u**an de**l**,	don**c**ellas cura**b**an d**e é**l,
princesas del su ro**z**ino."	princesas del su ro**c**ino."

Spelling Normalization

Approaches

- Statistical dictionary (SD).
- SMT.
- NMT.
  - LSTM.
  - Transformer.
- Character-based (CB) SMT.
- CBNMT.
  - CBNMT.
  - SubChar (Subwords–Characters).
  - CharSub (Characters–Subwords).
- CBNMT enriched with modern documents.
  - Synthetic data generated using modern documents.

Spelling Normalization

Work in progress

- So far, we have work using only error-free transcripts.
- The field of handwriting text recognition (HTR) also faces this problem.
- We are working on combining the HTR and MT models to improve the modern transcripts.