Machine Translation Applications to Historical Documents

Miguel Domingo, Francisco Casacuberta

midobal@prhlt.upv.es, fcn@prhlt.upv.es

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

PRHLT Seminar

CPI, May 3, 2022

Outline

1. Language Modernization

2. Spelling Normalization

3. Online Demonstrator

Outline

1. Language Modernization

2. Spelling Normalization

3. Online Demonstrator

Motivation

- Historical documents are an important part of our cultural heritage.
- However, due to their linguistic characteristics they are mostly limited to scholars.

Introduction

Goal: make historical documents more accessible to a general audience.

Introduction

Goal: make historical documents more accessible to a general audience.

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

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 through backtranslation.

Experimental framework

Corpora:

- Dutch Bible (17$^{\mathrm{th}}$ century Dutch; 30K segments).
- El Quijote (17$^{\mathrm{th}}$ century Spanish; 10K segments).
- OE-ME (11$^{\mathrm{th}}$ century English; 3K segments).

Metrics:

- TER.
- BLEU.

Experimental framework

Evaluation:

- Automatic metrics.
- Human evaluation.
  - Scholars (4 Scholars specialized in classic Spanish literature).
  - Non-experts (42 participants).

Evaluation

Automatic metrics

Approach	Dutch Bible		El Quijote		OE-ME	
	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]
Baseline	57.9	12.9	44.2	36.3	91.0	2.8
SMT	11.5	77.5	**30.7**†	**58.3**†	**39.6**†	**39.6**†
NMT$_{\text{LSTM}}$	13.8	79.6	55.1	39.8	82.7	12.8
NMT$_{\text{Transformer}}$	**11.1**†	**81.7**†	38.4	49.3	54.7	27.3
Enriched NMT$_{\text{LSTM}}$	**11.1**†	**80.6**†	**31.9**†	**57.3**†	**44.3**†	**35.9**†
Enriched NMT$_{\text{Transformer}}$	18.2	70.6	36.7	51.0	47.2	31.0

All results are significantly different between all approaches except those denoted with†.

Evaluation

Scholars

- **Fluency**: how fluid does the modernized sentence sound?
- **Lexical meaning**: how correct is the lexicon of the modernized sentence?
- **Syntax**: how correct is the syntactic construction of the modernized sentence?
- **Semantic**: is the meaning of the original sentence preserved in the modernized sentence?
  - **1**: the meaning is lost.
  - **2**: a great part of the meaning is lost.
  - **3**: half the meaning is lost.
  - **4**: part of the meaning is lost.
  - **5**: the meaning remains.
- **Modernization**: how appropriate is the modernization?

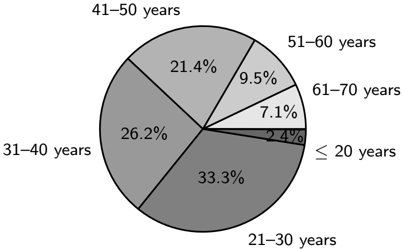
Evaluation

Scholars

	Fluency	Lexical meaning	Syntax	Semantic	Modernization
SMT	3.7	3.3	3.4	3.5	3.2
En. NMT_{LSTM}	3.7	3.3	3.4	3.5	3.2

Evaluation

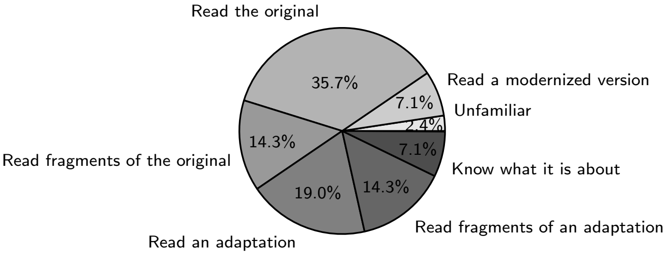
Non-experts



Age distribution.

Evaluation

Non-experts



Familiarity with *El Quijote*.

Evaluation

Non-experts

	Original	Modernized	Indifferent	Not equal
SMT	3.2	61.4	27.6	7.8
NMT	6.4	50.9	22.3	20.3

Percentage of cases in which the users selected that option.

Work in Progress

- Adapting pre-trained models for this task.
- We are working with mT5[1] since it covers 100 languages.

[1] 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.

Outline

1. Language Modernization

2. Spelling Normalization

3. Online Demonstrator

Motivation

- 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[2] indicate 45 different forms recorded for the pronoun *it*, 64 for the pronoun *she* and more than 500 for the preposition *through*.

[2] Linguistic Atlas of Late Medieval English.

Introduction

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

Introduction

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

Original

“Nunca fuera ca**u**allero
de damas ta**m**bien ser**u**ido,
como fuera don Qui**x**ote
quando de su aldea vino:
don**z**ellas cura**u**an d**e**l,
princesas del su ro**z**ino.”

Normalized

“Nunca fuera ca**b**allero
de damas ta**n** bien ser**v**ido,
como fuera don Qui**j**ote
cuando de su aldea vino:
don**c**ellas cura**b**an d**e** **é**l,
princesas del su ro**c**ino.”

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 through backtranslation.

Experimental framework

Corpora:

- Entremeses y Comedias (17th century Spanish; 35K segments).
- Quijote (17th century Spanish; 48K segments).
- Bohorič (18th century Slovene; 4K segments).
- Gaj (19th century Slovene; 13K segments).

Metrics:

- Character Error Rate (CER).
- TER.
- BLEU.

Main approaches

System	Quijote			Bohorič		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
Baseline	7.9	19.5	59.4	21.7	49.0	18.0
SD	3.9	5.5	89.3	16.2	20.7	56.1
CBSMT	**2.5**†	**3.0**†	**94.4**†	**2.4**	**8.7**	**80.4**
CBNMT$_{\text{LSTM}}$	2.7	4.3‡	93.3‡	29.4	39.5	48.7
En. CBNMT$_{\text{LSTM}}$	**2.2**†	4.0‡	93.2‡	28.6	38.3	49.5
CBNMT$_{\text{Trans.}}$	**1.9**†	**3.3**†	**93.9**†	26.2†	30.6†	60.0†
En. CBNMT$_{\text{Trans.}}$	**2.4**†	5.1	89.7	25.7†	29.8†	60.8†

All results are significantly different between all approaches except those denoted with † and ‡ (respectively).

Additional CBNMT approaches

System	Quijote			Bohorič		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
En. CBNMT$_{\text{LSTM}}$	**2.2**†	4.0†	93.2‡	28.6‡	38.3	49.5
En. SubChar$_{\text{LSTM}}$	**2.3**†	**3.3**‡	**94.9**†	29.5†	36.9	51.5
En. CharSub$_{\text{LSTM}}$	**2.3**†	4.1†	93.0‡	27.5*	39.6†	47.2
En. CBNMT$_{\text{Trans.}}$	**2.4**†	5.1	89.7	25.7	29.8‡	60.8†
En. SubChar$_{\text{Trans.}}$	**2.4**†	**3.2**‡	**94.4**†	27.3*	31.8	57.8
En. CharSub$_{\text{Trans.}}$	**2.4**†	**3.5**‡	93.9‡	**8.8**	**11.5**	**79.3**

All results are significantly different between all approaches except those denoted with †, ‡ and * (respectively).

Work in Progress

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

Outline

1. Language Modernization

2. Spelling Normalization

3. Online Demonstrator

Online Demonstrator

`https://demosmt.prhlt.upv.es/mthd/`