Enriching Character-based Neural Machine Translation with Modern Documents for Achieving an Orthography Consistency in Historical Documents

Miguel Domingo, Francisco Casacuberta

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

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

PatReCH 2019

Trento, September 9, 2019

Outline

Outline

Introduction

- The linguistic variation in historical documents has always been a concern for scholars in humanities.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.

Motivation

- Historical documents are an important part of our cultural heritage.
- Interest in effective natural language processing for these documents is on the rise.
- Achieve an orthography consistency by adapting the documents spelling to modern standards.

Example[2]:

> Bien responde la esperança
> en que engañado he viuido
> al cuydado que he tenido
> de tu estudio y tu criança!

[2] Fred F. Jehle (2001). *Works of Miguel de Cervantes in Old- and Modern-spelling*. Indiana University Purdue University Fort Wayne.

Motivation

- Historical documents are an important part of our cultural heritage.
- Interest in effective natural language processing for these documents is on the rise.
- Achieve an orthography consistency by adapting the documents spelling to modern standards.

Example[2]:

Bien responde la esperança	Bien responde la esperanza
en que engañado he viuido	en que engañado he vivido
al cuydado que he tenido	al cuidado que he tenido
de tu estudio y tu criança!	de tu estudio y tu crianza!

[2] Fred F. Jehle (2001). *Works of Miguel de Cervantes in Old- and Modern-spelling*. Indiana University Purdue University Fort Wayne.

Outline

Existing Normalization Approaches

Machine Translation (MT):

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y} \mid \mathbf{x}) \tag{1}$$

Character-based Statistical MT

Computes Eq. (1) at a character level. Words are split into characters and, then, conventional SMT is applied.

Character-based Neural MT

Neural approach to compute Eq. (1) at a character level. Words are split into characters and, then, conventional NMT is applied.

Character-based NMT Enriched with Modern Documents

- Scarce availability of parallel training data for historical documents.
- NMT approaches need an abundant quantity of parallel training data.
- We propose to use modern documents to enrich the NMT systems:
  1. We train a character-based SMT system (normalized version–original version).
  2. We translate the modern documents, obtaining a new synthetic version which captures the orthography inconsistencies that the original documents have.
  3. This new version, together with the original modern documents, conform the synthetic parallel data.
  4. We combine the synthetic data with the training dataset.
  5. We use the resulting dataset to train the enriched character-based NMT normalization system.

Outline

Corpora

- **Entremeses y Comedias**[3]: A 17$^{\text{th}}$ century Spanish collection of comedies by Miguel de Cervantes. It is composed of 16 plays, 8 of which have a very short length.
- **Quijote**[3]: The 17$^{\text{th}}$ century Spanish two-volumes novel by Miguel de Cervantes.
- **Bohorič**[4]: A collection of 18$^{\text{th}}$ century Slovene texts written in the old Bohorič alphabet.
- **Gaj**[4]: A collection of 19$^{\text{th}}$ century Slovene texts written in the Gaj alphabet.

---

[3] Fred F. Jehle (2001). *Works of Miguel de Cervantes in Old- and Modern-spelling*. Indiana University Purdue University Fort Wayne.

[4] Nikola Ljubešić et al. (2016). *Dataset of normalised Slovene text KonvNormSl 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1068.

		Entremeses y Comedias	Quijote	Bohorič	Gaj
Train	$\|S\|$	35.6K	48.0K	3.6K	13.0K
	$\|T\|$	250.0/244.0K	436.0/428.0K	61.2/61.0K	198.2/197.6K
	$\|V\|$	19.0/18.0K	24.4/23.3K	14.3/10.9K	34.5/30.7K
	$\|W\|$	52.4K	97.5K	33.0K	32.7K
Development	$\|S\|$	2.0K	2.0K	447	1.6K
	$\|T\|$	13.7/13.6K	19.0/18.0K	7.1/7.1K	25.7/25.6K
	$\|V\|$	3.0/3.0K	3.2/3.2K	2.9/2.5K	8.2/7.7K
	$\|W\|$	1.9K	4.5K	3.8K	4.5K
Test	$\|S\|$	2.0K	2.0K	448	1.6K
	$\|T\|$	15.0/13.3K	18.0/18.0K	7.3/7.3K	26.3/26.2K
	$\|V\|$	2.7/2.6K	3.2/3.2K	3.0/2.6K	8.4/8.0K
	$\|W\|$	3.3K	3.8K	3.8K	4.8K
Modern documents	$\|S\|$	500.0K	500.0K	500.0K	500.0K
	$\|T\|$	3.5M	3.5M	3.0M	3.0M
	$\|V\|$	67.3K	67.3K	84.7K	84.7K

Metrics

- Character Error Rate (CER).
- Translation Error Rate (TER).
- BiLingual Evaluation Understudy (BLEU).

- We used `sacreBLEU`[5] to ensure consistent BLEU scores.
- We applied approximate randomization tests[6], with $10,000$ repetitions and using a p-value of 0.05.

[5] Matt Post (2018). "A Call for Clarity in Reporting BLEU Scores". In: *Proceedings of the Third Conference on Machine Translation*, pp. 186–191.

[6] Stefan Riezler and John T Maxwell (2005). "On some pitfalls in automatic evaluation and significance testing for MT". . In: *Proceedings of the workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 57–64.

MT Systems

- SMT systems were trained with `Moses`.
- NMT systems were trained with `OpenNMT-py`.
- Statistical dictionary as a second baseline.

Outline

Results

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	CER	TER	BLEU	CER	TER	BLEU	CER	TER	BLEU	CER	TER	BLEU
Baseline	8.1	28.0	47.0	7.9	19.5	59.4	21.7	49.0	18.0	3.5	12.3	72.6
Statistical Dictionary	7.8	18.9	66.8	3.9	5.5	89.3	16.2	20.7	56.1	7.6	8.8	79.8
CBSMT	**1.3**	**4.4**	**91.7**	**2.5**	**3.0**	**94.4**	**2.4**	**8.7**	**80.4**	**1.4**	**5.1**	**88.3**
CBNMT	2.4	8.0	84.8	4.2	7.6	85.1	37.0	45.1	40.1	39.0	42.5	45.4
Enriched CBNMT	1.9	7.2	85.9	3.3	4.5	91.9	28.7	37.3	49.0	36.4	40.7	47.3

Example

Original: dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.

Example

Original: dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.

Normalized: dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.

Example

Original: dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.

Normalized: dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.

Statistical Dictionary: dobro manengo, da meni drugi, kakor je kristus cerkvene, in kakor ima katholshka cerkev navado drugi.

Example

Original:	dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.
Normalized:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
Statistical Dictionary:	dobro manengo, da meni drugi, kakor je kristus cerkvene, in kakor ima katholshka cerkev navado drugi.
CBSMT:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.

Example

Original:	dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.
Normalized:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
Statistical Dictionary:	dobro manengo, da meni drugi, kakor je kristus cerkvene, in kakor ima katholshka cerkev navado drugi.
CBSMT:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
CBNMT:	dobro manengo, da otže krztiti, kokor je krstiti.

Example

Original:	dobro manengo, de otshe kerstiti, koker je kristus goripostavel, inu koker ima katholshka zir kuv navado kerstiti.
Normalized:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
Statistical Dictionary:	dobro manengo, da meni drugi, kakor je kristus cerkvene, in kakor ima katholshka cerkev navado drugi.
CBSMT:	dobro manengo, da hoče krstiti, kakor je kristus goripostavil, in kakor ima katoliška cerkev n avado krstiti.
CBNMT:	dobro manengo, da otže krztiti, kokor je krstiti.
Enriched CBNMT:	dobro manengo, da otže krstiti, kakor je kriztus goripostavil, in koker ima katoliška cerkev nava

Outline

Conclusions

- We proposed a normalization method to enrich CBNMT systems using modern documents.
- We tested our proposal in different data sets, observing significant gains for all metrics.
- We compared several normalization approaches, reaching the conclusion than CBSMT systems are more suitable for this task.
- We believe that this is specially true due to the scarce availability of parallel training data when working with historical documents[7].

[7] Marcel Bollmann and Anders Søgaard (2016). "Improving historical spelling normalization with bi-directional LSTMs and multi-task learning". In: *Proceedings of the International Conference on the Computational Linguistics*, pp. 131–139.

Future Work

- Further research the use of modern documents to enrich the neural systems.
- Further investigate how to balance synthetic and real data.
- Use a data selection approach to find the most suitable data for each corpus.