A Comparison of Character-based Neural Machine Translations Techniques Applied to Spelling Normalization

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 2020

Online, January 11, 2021

Outline

1. Introduction
2. Normalization Approaches
3. Experimental Framework
4. Results
5. Conclusions

Outline

1. Introduction

2. Normalization Approaches

3. Experimental Framework

4. Results

5. Conclusions

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- The linguistic variation in historical documents has always been a concern for scholars in humanities.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- Historical documents are an important part of our cultural heritage.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]:

¿Cómo est**ay**s, Ro**z**inante, tan delgado?
Porque nunca se come, y se trabaja.
Pues ¿qué es de la ce**u**ada y de la paja?
No me de**x**a mi amo ni **v**n bocado.

¿Cómo est**ái**s, Ro**c**inante, tan delgado?
Porque nunca se come, y se trabaja.
Pues ¿qué es de la ce**b**ada y de la paja?
No me de**j**a mi amo ni **u**n bocado.

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

Outline

1. Introduction

2. Normalization Approaches

3. Experimental Framework

4. Results

5. Conclusions

Normalization Approaches

Machine Translation (MT):

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

Character-based Statistical MT (CBSMT)
Computes Eq. (1) at a character level. Words are split into characters and, then, conventional SMT is applied.

Normalization Approaches

Machine Translation (MT):

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

Character-based Neural MT

Neural approach to compute Eq. (1) at a character level.

- **CBNMT**: This technique uses a character level strategy. Words from both the source and the target are split into characters.
- **SubChar**: This technique combines a sub-word level and a character level strategies. Source words are split into sub-words and target words into characters.
- **CharSub**: This technique combines a character level and a sub-word level strategy. Source words are split into characters and target words into sub-words.

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 make use of 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

1. Introduction

2. Normalization Approaches

3. Experimental Framework

4. Results

5. Conclusions

Corpora

- **Entremeses y Comedias**[3]: A 17th 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 17th century Spanish two-volumes novel by Miguel de Cervantes.
- **Bohorič**[4]: A collection of 18th century Slovene texts written in the old Bohorič alphabet.
- **Gaj**[4]: A collection of 19th 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`.

Outline

1. Introduction

2. Normalization Approaches

3. Experimental Framework

4. Results

5. Conclusions

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
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	1.7†	12.0	82.7	2.7	4.3†	93.3‡	29.4	39.5	48.7	31.5‡	36.9	53.1
SubChar	23.3	32.8	54.1	**2.2**†	3.7	93.8‡	36.7	47.7	39.4	32.7	37.3	52.4
CharSub	5.8	18.2	75.2	3.7	5.8	89.8	67.9	83.8	5.3	37.2	48.1	36.3
Enriched CBNMT	1.7†	13.3	79.4†	**2.2**†	4.0†	93.2‡	28.6	38.3	49.5	30.5	35.4†	54.9†
Enriched SubChar	37.8	35.8	59.3	**2.3**†	**3.3**†	**94.9**†	29.5	36.9	51.5	31.5‡	35.9†	54.3†
Enriched CharSub	3.8	15.2	78.9†	**2.3**†	4.1†	93.0‡	27.5	39.6	47.2	29.4	37.2	52.3

Baseline system corresponds to considering the original document as the document to which the spelling has been normalized to match modern standards. All results are significantly different between all systems except those denoted with † and ‡ (respectively). Best results are denoted in **bold**.

Examples

Original:	¡O mal logrado moço! Salid fuera;
Normalized:	¡Oh mal logrado mozo! Salid fuera;
CBSMT:	¡Oh mal logrado mozo! Salí fuera;
CBNMT:	¡Oh mal logrado mozo! Salí fuera;
Enriched CBNMT:	¡Oh mal logrado mozo! Salí fuera;
SubChar:	gueso mal logrado mozo Salí guesto fuera;
Enriched SubChar:	¡Oh mal logrado mozo! ␣␣␣␣␣ ␣␣␣␣␣␣
CharSub:	¡Oh mal logrado mozo! allí fuera;
Enriched CharSub:	¡Oh mal logrado mozo! Salí fuera;

Examples

Original: “Para esso se yo vn buen remedio”, dixo el del Bosque;
Normalized: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CBSMT: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CBNMT: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;
Enriched CBNMT: “Para es␣o se yo un buen remedio”, dijo el del Bosque;

SubChar: “Para es␣o se yo un buen remedio”, dijo el del Bosque;
Enriched SubChar: “Para es␣o sé yo un buen remedio”, dijo el del Bosque;

CharSub: “Para es␣o se yo un buen remedio”, dijo el del Bosque;
Enriched CharSub: “Para es␣o se yo un buen remedio”, dijo el del Bosque;

Outline

1. Introduction
2. Normalization Approaches
3. Experimental Framework
4. Results
5. Conclusions

Conclusions

- We evaluated different CBNMT normalization approaches, some of which their neural models were enriched using modern documents.
- We tested our proposal in different data sets, and reached the conclusion that not all approaches are equally suited for each task.
- CBSMT systems yielded the best results for three out of the four tasks.
- We believe that this is mostly 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.
- In this work, we used a previously known method in order to assess the different CBNMT approaches under the same circumstances. We should further investigate new methods such as using a data selection approach to find the most suitable data for each corpus.