Spelling Normalization of Historical Documents by Using a Machine Translation Approach

Authors: Miguel Domingo, Francisco Casacuberta
{midobal, fcn}@prhlt.upv.es
Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València

Introduction

- Due to the lack of a spelling convention, orthography changes depending on the author and the time period.
- This represents a problem for the preservation of the cultural heritage, which strives to create a digital text version of historical documents.

Systems

We tackled spelling normalization using the following MT approaches:

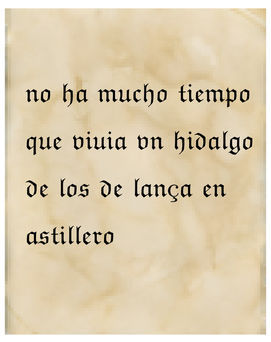
- SMT (trained with `Moses`).
- NMT (trained with `OpenNMT`).
- Character-based MT (CBSMT, CBNMT).

As baseline, we consider the original document. Additionally, we implemented a second baseline based on a statistical dictionary (SD).

Metrics

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

Spelling Normalization



Original document

Transcription

no ha mucho tiempo que viuia vn hidalgo
de los de lança en astillero

no ha mucho tiempo que vivía un hidalgo
de los de lanza en astillero

Version with updated spelling

Corpora

- **Entremeses y Comedias**: A collection of comedies by Miguel de Cervantes, written in 17$^{\text{th}}$ century Spanish. (∼40K sentences.)
- **Quijote**: The 17$^{\text{th}}$ century Spanish novel by Miguel de Cervantes. (∼50K sentences.)
- **Bohorič**: A collection of 18$^{\text{th}}$ century Slovene texts written in the Bohorič alphabet. (∼4K sentences.)
- **Gaj**: A collection of 19$^{\text{th}}$ century Slovene texts written in the Gaj alphabet. (∼13K sentences.)

Results

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	BLEU	TER	CER	BLEU	TER	CER	BLEU	TER	CER	BLEU	TER	CER
Baseline	46.1 ± 1.4	31.7 ± 1.2	12.0 ± 0.4	59.6 ± 1.2	19.4 ± 0.7	7.4 ± 0.3	16.4 ± 1.6	49.0 ± 1.5	21.7 ± 0.6	68.1 ± 1.1	12.3 ± 0.5	3.5 ± 0.1
SD	80.8 ± 1.2	8.3 ± 0.5	4.0 ± 0.3	89.7 ± 0.8	5.3 ± 0.5	3.4 ± 0.3	52.5 ± 2.0	20.7 ± 1.2	17.2 ± 0.7	75.1 ± 0.8	8.8 ± 0.4	8.7 ± 0.3
SMT	82.1 ± 1.1	8.0 ± 0.5	6.7 ± 0.2	91.1 ± 0.7	4.5 ± 0.4	5.3 ± 0.3	63.0 ± 2.1	15.1 ± 1.1	9.0 ± 0.5	$\mathbf{82.6 \pm 0.7}$	5.2 ± 0.3	2.8 ± 0.1
SMT_{BPE}	83.6 ± 1.1	7.2 ± 0.5	6.2 ± 0.2	$\mathbf{94.6 \pm 0.6}$	$\mathbf{2.8 \pm 0.3}$	4.3 ± 0.2	70.4 ± 2.0	11.7 ± 1.0	5.3 ± 0.3	$\mathbf{83.7 \pm 0.7}$	$\mathbf{1.8 \pm 0.3}$	2.7 ± 0.1
NMT	72.2 ± 1.4	15.2 ± 0.9	18.0 ± 0.8	84.4 ± 0.9	8.1 ± 0.5	10.2 ± 2.4	36.7 ± 2.0	33.9 ± 2.1	41.4 ± 1.4	50.4 ± 1.4	28.3 ± 3.3	36.0 ± 2.7
NMT_{BPE}	76.7 ± 1.3	12.4 ± 0.8	8.1 ± 0.5	92.0 ± 0.7	4.6 ± 0.4	3.8 ± 0.3	31.6 ± 2.2	43.5 ± 6.1	48.6 ± 3.6	68.0 ± 1.5	23.7 ± 3.7	19.8 ± 2.6
CBSMT	$\mathbf{91.4 \pm 0.9}$	$\mathbf{3.7 \pm 0.4}$	$\mathbf{1.2 \pm 0.1}$	$\mathbf{94.7 \pm 0.6}$	$\mathbf{2.8 \pm 0.3}$	$\mathbf{2.0 \pm 0.2}$	$\mathbf{75.5 \pm 1.8}$	$\mathbf{8.7 \pm 0.9}$	$\mathbf{2.4 \pm 0.2}$	$\mathbf{83.2 \pm 0.7}$	5.0 ± 0.3	$\mathbf{1.3 \pm 0.1}$
CBNMT	81.3 ± 1.3	8.3 ± 0.8	3.0 ± 0.6	91.0 ± 0.7	4.6 ± 0.4	2.9 ± 0.3	27.6 ± 2.4	85.2 ± 6.7	68.2 ± 4.5	40.2 ± 1.9	62.7 ± 2.9	52.5 ± 2.1

Conclusions

- SMT yielded better results than NMT.
- Character-based approaches yielded the best results for each kind of system.
- The statistical dictionary could be useful in cases in which its worth sacrificing quality to increase speed.

Future Work

- Try new character-based approaches.
- Obtain more diverse corpora from broader domains.
- Synthetic data.

Acknowledgements

The research leading to these results has received funding from the Ministerio de Economía y Competitividad (MINECO) under project CoMUN-HaT (grant agreement TIN2015-70924-C2-1-R), and Generalitat Valenciana (grant agreement PROMETEO/2018/004). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for this research.