A Machine Translation Approach for Modernizing Historical Documents Using Backtranslation

Miguel Domingo, Francisco Casacuberta

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

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

IWSLT 2018

Bruges, October 30, 2018

Outline

Outline

Introduction

- Human language evolves with the passage of time.
- The lack of an spelling convention makes orthography to change from one document to another.
- Historical documents are hard to comprehend by contemporary people.
- They are only accessible to scholars specialized in the time period in which a certain document was written.

Motivation

- Help in the preservation of the cultural heritage.
- Make historical documents accessible to a broader audience.

Example[1]:

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:

[1] John Crowther (2004). *No Fear Shakespeare: Sonnets*. SparkNotes.

Motivation

- Help in the preservation of the cultural heritage.
- Make historical documents accessible to a broader audience.

Example[1]:

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date:

Shall I compare you to a summer day?
You're lovelier and milder.
Rough winds shake the pretty buds of May,
and summer doesn't last nearly long enough.

[1] John Crowther (2004). *No Fear Shakespeare: Sonnets*. SparkNotes.

Outline

Machine Translation

Statistical Machine Translation

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

Neural Machine Translation

Neural approach to compute Eq. (1).

Backtranslation[2]

- To build a parallel corpus from monolingual data.
- Useful in resources-poor scenarios.

[2] Rico Sennrich et al. (2015). "Improving neural machine translation models with monolingual data". In: *arXiv preprint arXiv:1511.06709*.

Outline

Corpora

- Scarce availability of parallel training data for historical documents[3].
- We made use of the Dutch Bible[4].
- We built two additional corpora.
- We collected monolingual data from free-of-right sources.

---

[3] 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.

[4] Erik Tjong Kim Sang et al. (2017). "The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation". In: *Computational Linguistics in the Netherlands Journal* 7, pp. 53–64.

- Corpora:
  - Don Quijote: original version, modern version[5].
  - El Conde Lucanor: original version, modern version[6].
- Steps:
  - Split document into sentences:
    - Remove empty lines.
    - Transform document into a single sentence.
    - Split into sentences by adding line breaks to relevant punctuation (i.e., dots, quotation marks, etc).
  - Ensure consistency of symbols (e.g., quotation marks).
  - Align documents using `Hunalign`[7].

[5] Andrés Trapiello (2015). *Don Quijote de la Mancha Puesto en castellano actual íntegra y fielmente por Andrés Trapiello*. Ediciones Destino.

[6] Luis López Nieves (2002). *El Conde Lucanor*. Biblioteca Digital Ciudad Seva.

[7] D. Varga et al. (2005). "Parallel Corpora for Medium Density Languages". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 590–596.

		Dutch Bible	El Quijote	El Conde Lucanor
Train	$\|S\|$	35.2K	10K	-
	$\|T\|$	870.4/862.4K	283.3/283.2K	-
	$\|V\|$	53.8/42.8K	31.7/31.3K	-
Development	$\|S\|$	2000	2000	-
	$\|T\|$	56.4/54.8K	53.2/53.2K	-
	$\|V\|$	9.1/7.8K	10.7/10.6K	-
Test	$\|S\|$	5000	2000	2252
	$\|T\|$	145.8/140.8K	41.8/42.0K	62.0/56.7K
	$\|V\|$	10.5/9.0K	8.9/9.0K	7.4/8.6K
Monolingual	$\|S\|$	4.1M	567.0K	-
	$\|T\|$	88.3M	9.5M	-
	$\|V\|$	2.0M	470.4K	-

Metrics

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

Confidence intervals ($p = 0.05$) were computed for all metrics by means of bootstrap resampling[8].

[8] Philipp Koehn (2004). "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 388–395.

MT Systems

- SMT systems were trained with Moses.
- NMT systems were trained with OpenNMT.
- BPE was applied to both kind of systems.

Outline

Results

System	Dutch Bible		El Quijote		El Conde Lucanor	
	BLEU	TER	BLEU	TER	BLEU	TER
Baseline	13.5 ± 0.3	57.0 ± 0.3	36.5 ± 0.8	43.3 ± 1.1	5.8 ± 0.3	89.6 ± 1.0
Baseline_2	50.8 ± 0.4	26.5 ± 0.3	-	-	-	-
SMT	$\mathbf{80.1 \pm 0.5}$	$\mathbf{9.9 \pm 0.3}$	$\mathbf{58.9 \pm 1.0}$	$\mathbf{29.4 \pm 1.2}$	$\mathbf{8.4 \pm 0.3}$	$\mathbf{83.8 \pm 1.0}$
NMT	38.0 ± 0.6	51.7 ± 2.2	37.4 ± 1.2	51.5 ± 2.0	2.7 ± 0.2	99.5 ± 2.0
$\text{NMT}_{\text{Synthetic}}$	17.4 ± 0.5	65.6 ± 1.7	45.2 ± 1.3	50.6 ± 3.5	3.1 ± 0.2	165.1 ± 8.2

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

$\text{NMT}_{\text{Synthetic}}$: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fierded que suele pintarse.

El Quijote

Original: Y, leuantandose, dexó de comer, y fue a quitar la cubierta de la primera imagen, que mostro ser la de San Iorge puesto a cauallo, con vna serpiente enroscada a los pies, y la lana atrauessada por la boca, con la fierea que suele pintarse.

Modernized: Y dejando de comer, se levantó y fue a quitar la cubierta de la primera imagen, que resultó ser la de san Jorge a caballo, con una serpiente enroscada a los pies y la lanza atravesándole la boca, con la fiereza que suele pintarse.

SMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fiereza que suele pintarse.

NMT: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera ancen, que mostró ser la de San Marorge a los pies y la lanza ahabesada por la boca;

$\text{NMT}_{\text{Synthetic}}$: Y levantándose, dejó de comer, y fue a quitar la cubierta de la primera imagen, que mostró ser la de San Jorge puesto a caballo, con una serpiente enroscada a los pies, y la lanza atravesada por la boca, con la fierded que suele pintarse.

El Conde Lucanor

Original: -Señor conde Lucanor -dixo Patronio-, vien entiendo que el mío consejo non vos faze grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, fazerlo he luego.

Modernized: -Señor Conde Lucanor -dijo Patronio-, bien sé que mi consejo no os hace mucha falta, pero, como confiáis en mí,

---

SMT: - Señor conde Lucanor -dijo Patroniorosa, vien entiendo que el mío consejo non vos face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos aconseje en ello, ferlo he luego .

NMT: Señor conde Olcanor dijo dijo Pacasos dijo en entiendo que el mío consejo non os fazo felimengua y vuestra merced es que vos diga lo que en esto entiendo.

NMT$_{\text{Synthetic}}$: -Señor conde Lucanor -dijo Patronio, vien entiendo que el mío consejo non es face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, también yo he dicho.

El Conde Lucanor

Original:	-Señor conde Lucanor -dixo Patronio-, vien entiendo que el mío consejo non vos faze grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, fazerlo he luego.
Modernized:	-Señor Conde Lucanor -dijo Patronio-, bien sé que mi consejo no os hace mucha falta, pero, como confiáis en mí,
SMT:	- Señor conde Lucanor -dijo Patroniorosa, vien entiendo que el mío consejo non vos face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos aconseje en ello, ferlo he luego .
NMT:	Señor conde Olcanor dijo dijo Pacasos dijo en entiendo que el mío consejo non os fazo felimengua y vuestra merced es que vos diga lo que en esto entiendo.
NMT$_{\text{Synthetic}}$:	-Señor conde Lucanor -dijo Patronio, vien entiendo que el mío consejo non es face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, también yo he dicho.

El Conde Lucanor

Original:	-Señor conde Lucanor -dixo Patronio-, vien entiendo que el mío consejo non vos faze grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, fazerlo he luego.
Modernized:	-Señor Conde Lucanor -dijo Patronio-, bien sé que mi consejo no os hace mucha falta, pero, como confiáis en mí,
SMT:	- Señor conde Lucanor -dijo Patroniorosa, vien entiendo que el mío consejo non vos face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos aconseje en ello, ferlo he luego .
NMT:	Señor conde Olcanor dijo dijo Pacasos dijo en entiendo que el mío consejo non os fazo felimengua y vuestra merced es que vos diga lo que en esto entiendo.
NMT$_{\text{Synthetic}}$:	-Señor conde Lucanor -dijo Patronio, vien entiendo que el mío consejo non es face grant mengua, pero vuestra voluntad es que vos diga lo que en esto entiendo, et vos conseje sobre ello, también yo he dicho.

Outline

Conclusions

- Tested with 3 historical datasets from 3 different time periods and 2 different languages.
- SMT yielded the best results.
- Improvements of 22 to 67 BLEU points and 14 to 48 TER points.
- NMT suffered from the scarce training data.
- Backtranslation was only able to improve results for one dataset, and only in terms of BLEU.

Future Work

- Research the relation between monolingual and training data.
- Explore the use of data selection techniques for constructing the monolingual corpus.