An Interactive Machine Translation Framework for Modernizing the Language of Historical Documents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

IbPRIA 2022

Aveiro, May 4, 2022

Outline

Outline

Introduction

- Historical documents are part of our cultural heritage.
- However, due to their linguistic characteristics they are mostly limited to scholars.
- Language modernization helps non-experts to understand their content, but it is not error-free.
- The interactive machine translation framework can help scholars to generate error-free modernization.

Outline

Language Modernization

Goal: make historical documents more accessible to a general audience.

Language Modernization

Goal: make historical documents more accessible to a general audience.

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

Language Modernization

Approaches:

$$\hat{y}_1^{\hat{I}} = \arg\max_{y_1^I} Pr(y_1^I \mid x_1^J)$$

- Statistical machine translation (SMT).
- Neural machine translation (NMT).
  - Recurrent neural networks with long short-term memory units (LSTM).
  - Transformer.

Outline

Interactive Machine Translation

Goal: collaborative framework in which human and machine work together to produce the final high-quality translations.

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.

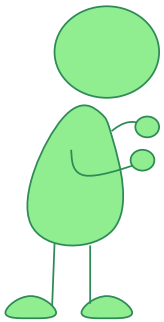
All things he foresees and knows, and
of all nations language he understands.

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.



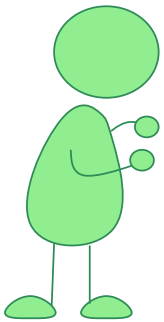
All things he foresees and knows, and
of all nations language he understands.

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.



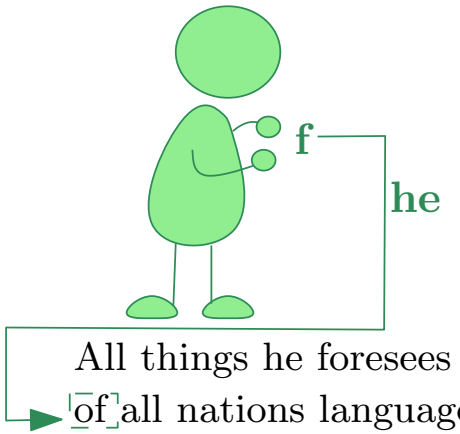
All things he foresees and knows, and
of all nations language he understands.

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.

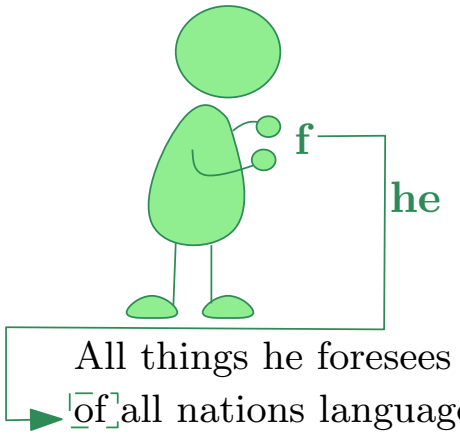


Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.



All things he foresees and knows, and
of all nations language he understands.

All things he foresees and knows, and
he understands of all nations language.

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Suffix generation:

$$\hat{y}_{i+1}^{\hat{I}} = \arg\max_{I, y_{i+1}^{I}} Pr(y_{i+1}^{I} \mid x_1^J, f = \tilde{y}_1^i) = \arg\max_{I, y_{i+1}^{I}} Pr(\tilde{y}_1^i \, y_{i+1}^{I} \mid x_1^J)$$

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Suffix generation:

$$\hat{y}_{i+1}^{\hat{I}} = \arg\max_{I, y_{i+1}^{I}} Pr(y_{i+1}^{I} \mid x_1^J, f = \tilde{y}_1^i) = \arg\max_{I, y_{i+1}^{I}} Pr(\tilde{y}_1^i \, y_{i+1}^{I} \mid x_1^J)$$

MT fundamental equation:

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^{I}} Pr(y_1^{I} \mid x_1^J)$$

Interactive Machine Translation

Segment-based IMT

Interactive Machine Translation

Segment-based IMT

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.

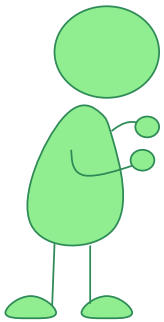
All things he foresees and knows, and
of all nations language he understands.

Interactive Machine Translation

Segment-based IMT

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.



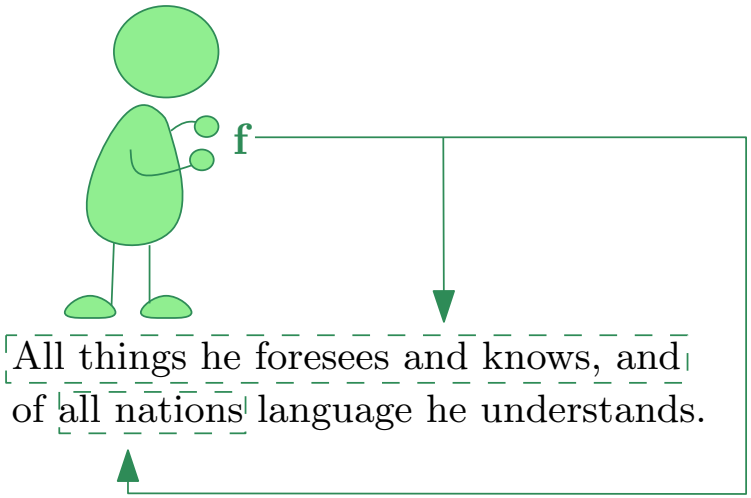
All things he foresees and knows, and
of all nations language he understands.

Interactive Machine Translation

Segment-based IMT

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.

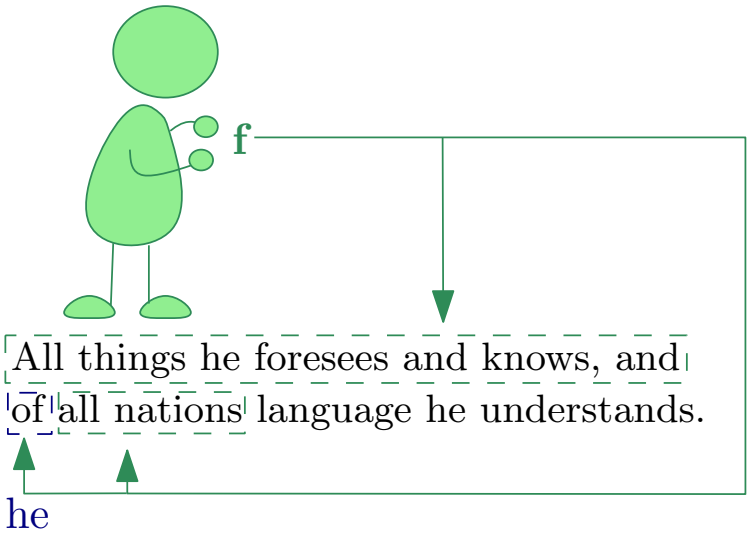


Interactive Machine Translation

Segment-based IMT

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.

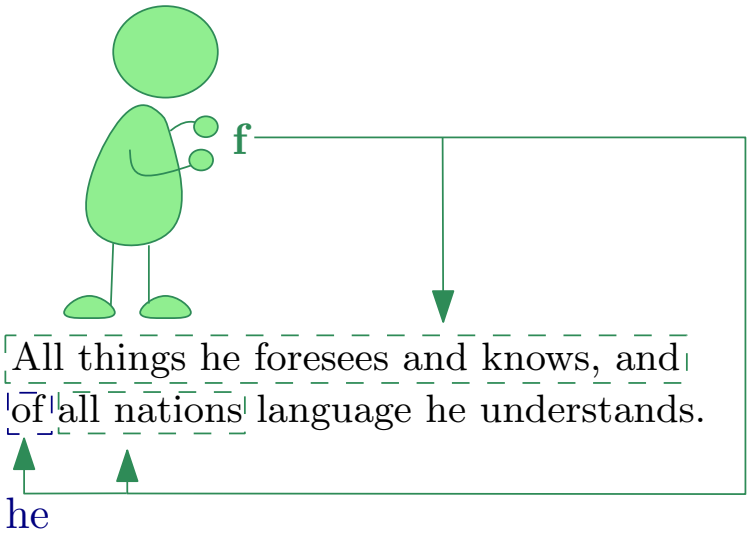


Interactive Machine Translation

Segment-based IMT

Source: Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

Target translation: All things he foresees and knows, and he understands the tongues of all nations.



Outline

Corpora

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

Metrics

- Word Stroke Ratio (WSR).
- Mouse Action Ratio (MAR).
- Translation Error Rate (TER).
- BiLingual Evaluation Understudy (BLEU).

- We applied approximate randomization tests[1], with $10,000$ repetitions and using a p-value of 0.05.

[1] 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.

Outline

Results

Corpus	Approach	Modernization quality		Prefix-based		Segment-based	
		TER [↓]	BLEU [↑]	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]
Dutch Bible	SMT	11.5	77.5	14.3	4.4	**9.0**	**10.8**
	NMT_{LSTM}	50.7†	43.4	42.6‡	9.2	42.6‡	50.9
	$\text{NMT}_{\text{Transformer}}$	50.3†	35.8	49.2‡	10.4	49.2‡	48.3
El Quijote	SMT	30.7	58.3	38.8	10.9	**22.0**	**19.7**
	NMT_{LSTM}	42.9	50.4	68.9‡	11.8	68.9‡	47.8
	$\text{NMT}_{\text{Transformer}}$	47.3	46.1	73.2‡	13.4	73.2‡	50.5
OE-ME	SMT	39.6	39.6	58.2	15.5	**28.2**	**26.1**
	NMT_{LSTM}	56.4	30.3	72.1‡	12.8†	72.1‡	59.5
	$\text{NMT}_{\text{Transformer}}$	58.9	28.2	73.5‡	13.3†	73.5‡	49.5

All results are significantly different between all approaches except those denoted with †. Given the same approach, all results are significantly different between the different IMT protocols except those denoted with ‡.

Examples

source (x): Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

target translation ($\hat{y}$): All things he foresees and knows, and he understands the tongues of all nations.

IT-0	**System**	All things he foresceawa and knows, and of all nations language he understands.
IT-1	**User**	All things he **foresees** and knows, and of all nations language he understands.
	System	All things he foresees and knows, and of all nations language he understands.
IT-2	**User**	All things he foresees and knows, and **he** all nations language he understands.
	System	All things he foresees and knows, and he understands of all nations language.
IT-3	**User**	All things he foresees and knows, and he understands **the** all nations language.
	System	All things he foresees and knows, and he understands the beginning of all nations language.
IT-4	**User**	All things he foresees and knows, and he understands the **tongues** of all nations language.
	System	All things he foresees and knows, and he understands the tongues all.
IT-5	**User**	All things he foresees and knows, and he understands the tongues **of**
	System	All things he foresees and knows, and he understands the tongues of all.
IT-6	**User**	All things he foresees and knows, and he understands the tongues of all **nations.**
	System	All things he foresees and knows, and he understands the tongues of all nations .
END	**User**	All things he foresees and knows, and he understands the tongues of all nations.

Example of a prefix-based IMT session.

Examples

source (x): Ealle ing he foresceawa and wt, and ealra eoda gereord he cann.

target translation (ŷ): All things he foresees and knows, and he understands the tongues of all nations.

IT-0	**System**	All things he foresceawa and knows, and of all nations language he understands.
IT-1	**User**	All things he **foresees** and knows, and of all nations language he understands.
	System	All things he foresceawa foresees and knows, and language he understand of all nations .
IT-2	**User**	All things he foresees and knows, and he understand **the** of all nations .
	System	All things he foresees and knows, and he understand the language of all nations .
IT-3	**User**	All things he foresees and knows, and he understand the **tongues** of all nations .
	System	All things he foresees and knows, and he understand the tongues of all nations .
END	**User**	All things he foresees and knows, and he understands the tongues of all nations.

Example of a segment-based IMT session.

Outline

Conclusions and Future Work

- We deployed the interactive framework into language modernization to help scholar to generate error-free modernizations.
- We deployed two different protocols: SMT and NMT.
- While IMT always succeeds in reducing the human effort, the SMT approach yielded the best results.
- We evaluated our proposal under simulated conditions. In a future work, we would like to perform a human evaluation with the help of scholars.