Historical Documents Modernization

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

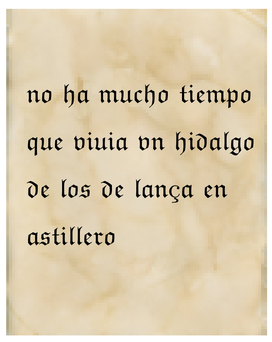
Introduction

- Language evolution makes historical documents hard to comprehend by contemporary people.
- Frequently, this problem limits its accessibility to scholars specialized in the time period in which they were written.
- Adapting the language to modern standards (using a translation approach) could help to break this barrier and increase their accessibility to a broader audience.

Tasks

- **Standard spelling**: update document's spelling to match current standards.
- **Document modernization**: translate the document into a modern version of its original language.

Standar Spelling



Original document

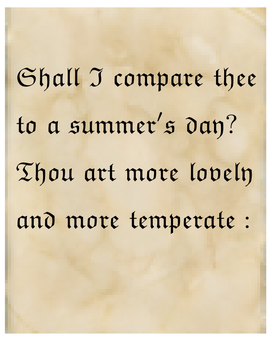
Transcription

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

Document Modernization



Original document

Transcription

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:

Shall I compare you to a summer day?
You're lovelier and milder.

Modern version

Approach

We approached both tasks as a translation task. The main problem arisen was the lack of suitable training data. To solve it, we made use of a data selection technique (**infrequent n-grams**) to filter the available out-of-domain corpora:

$$i(\mathbf{x}) = \sum_{\mathbf{m} \in X} \min(1, R(\mathbf{m}))t$$

- X: set of n-grams.
- $R(\mathbf{m})$: counts of $\mathbf{m}$ in $\mathbf{x}$.
- t: infrequency threshold.

Systems

- Standard SMT system (trained with `Moses`).
- Use of Byte Pair Encoding to reduce vocabulary problems.
- Additional language model (from external data).

Corpora

CLIN2017 Shared Task on Translating Historical Text (Dutch):

- **Bible** (books from different versions of the Dutch Bible):
  - 1637–1888 train and test (fragment from another Bible book).
  - 1637–2010 train.
  - 1657–1888 train.
  - 1657–2010 train.
- **Dutch Literature** (collection of fragments from Dutch literature):
  - 17^{th}–21^{st} century development (small text) and test.

19^{th} and 21^{st} century works from the *Digitale Bibliotheek voor de Nederlandse letteren* (to enrich language models).

Results

Standard spelling (train: all versions of Bible, test: Dutch literature):

System	Original corpora		Data selection	
	BLEU	TER	BLEU	TER
Baseline	29.9 ± 1.8	32.4 ± 1.1	-	-
SMT	48.1 ± 1.8	22.0 ± 0.8	$\mathbf{49.9 \pm 1.8}$	$\mathbf{20.2 \pm 0.8}$
+ LM_2	49.4 ± 1.8	21.2 ± 0.8	49.8 ± 1.8	20.9 ± 0.8
SMT_{BPE}	48.6 ± 1.6	24.2 ± 0.9	49.2 ± 1.6	23.7 ± 0.8
+ LM_2	47.9 ± 1.7	25.5 ± 0.9	49.9 ± 1.7	23.7 ± 0.8

Best results achieved using **standard SMT** and **data selection**.

Document modernization (train and test: 1637–1888 Bible):

System	BLEU	TER
Baseline	13.5 ± 0.3	57.0 ± 0.3
$Baseline_2$	50.8 ± 0.4	26.5 ± 0.3
SMT	64.8 ± 0.4	17.0 ± 0.3
+ LM_2	65.1 ± 0.4	17.3 ± 0.3
SMT_{BPE}	64.8 ± 0.4	17.4 ± 0.3
+ LM_2	$\mathbf{66.7 \pm 0.4}$	$\mathbf{16.2 \pm 0.3}$

Best results achieved using **BPE** and an **additional language model**.

Conclusions

- Document's language quality increased (with respects to the modern language).
- Tested on the task of standardizing document's spelling.
- BPE improved modernization but not updating the spelling.
- Filtering the train corpus improved results.

Future Work

- Experiment with more corpora.
- Historical manuscripts. (They present extra difficulties such as abbreviations particular to each author.)
- Neural Machine Translation.

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), Generalitat Valenciana under project ALMAMATER (grant agreement PROMETEOII/2014/030) and Universitat Politècnica de València under grant FPI (2014).