Some Contributions to Interactive Machine Translation and to the Applications of Machine Translation for Historical Documents

Miguel Domingo

Supervised by Prof. Francisco Casacuberta

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

Ph.D. defense

DSIC, January 28, 2022

Outline

1. Interactive Machine Translation. (Chapter 3.)
2. Historical Document Processing. (Chapters 4 and 5.)
3. IMT for the Processing of Historical Documents. (Chapter 6.)
4. Conclusions

Outline

1. Interactive Machine Translation. (Chapter 3.)

2. Historical Document Processing. (Chapters 4 and 5.)

3. IMT for the Processing of Historical Documents. (Chapter 6.)

4. Conclusions

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: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

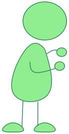
the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

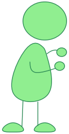


the commission found that the measures relating to contracts temporaires inférieurs bourses to two years

Interactive Machine Translation
Prefix-based interactive machine translation (IMT)

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration



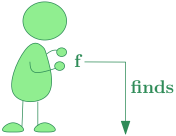
the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration



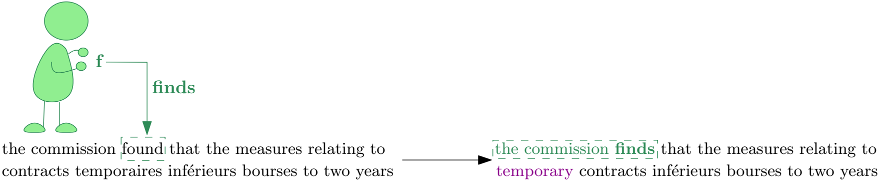
the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration



the commission found that the measures relating to contracts temporaires inférieurs bourses to two years

the commission **finds** that the measures relating to temporary contracts inférieurs bourses to two years

Interactive Machine Translation

Prefix-based interactive machine translation (IMT)

Suffix generation:

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

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: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

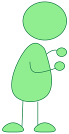
Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years

Interactive Machine Translation
Segment-based IMT

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration



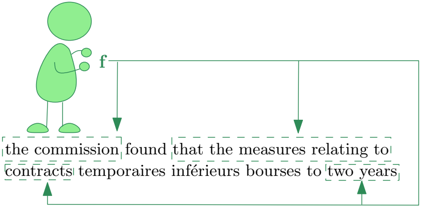
the commission found that the measures relating to contracts temporaires inférieurs bourses to two years

Interactive Machine Translation

Segment-based IMT

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

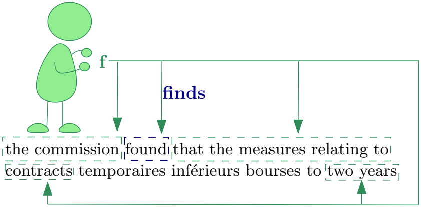

f
the commission found that the measures relating to
contracts temporaires inférieurs bourses to two years

Interactive Machine Translation

Segment-based IMT

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration

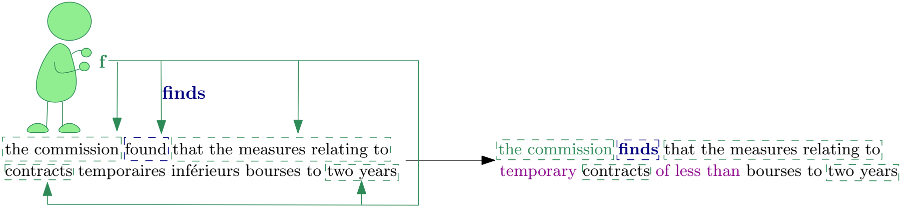


the commission found that the measures relating to contracts temporaires inférieurs bourses to two years

Interactive Machine Translation
Segment-based IMT

Source: la commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans

Target translation: the commission finds that the measures relating to temporary contracts of less than two years duration


f
finds
the commission found that the measures relating to contracts temporaires inférieurs bourses to two years
the commission finds that the measures relating to temporary contracts of less than bourses to two years

Interactive Machine Translation
Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Interactive Machine Translation
Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests

Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation:

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: If you have been exposed , you should consult go your doctor for tests

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion:

Interactive Machine Translation
Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should}}$ ~~consult~~ $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests

Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Word correction:

Interactive Machine Translation

Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation
Segment-based IMT: user actions

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

User actions:

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation
Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

User actions:

Segment validation: inserting a new segment $\tilde{\mathbf{f}}_i$ in $\tilde{\mathbf{f}}_1^N$.

Words deletion: $\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{your doctor for tests}}$

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

User actions:

Segment validation: inserting a new segment $\tilde{\mathbf{f}}_i$ in $\tilde{\mathbf{f}}_1^N$.

Words deletion: merging $\tilde{\mathbf{f}}_i$, $\tilde{\mathbf{f}}_{i+1}$ into a new one.

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

User actions:

Segment validation: inserting a new segment $\tilde{\mathbf{f}}_i$ in $\tilde{\mathbf{f}}_1^N$.

Words deletion: merging $\tilde{\mathbf{f}}_i$, $\tilde{\mathbf{f}}_{i+1}$ into a new one.

Word correction: inserting a new one-word validated segment $\tilde{\mathbf{f}}_i$ in $\tilde{\mathbf{f}}_1^N$.

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

$$\widehat{\mathbf{h}}_0^{N+1} = \widehat{\mathbf{h}}_0, \ldots, \widehat{\mathbf{h}}_{N+1}$$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

$$\widehat{\mathbf{h}}_0^{N+1} = \widehat{\mathbf{h}}_0, \ldots, \widehat{\mathbf{h}}_{N+1}$$

Translation segments generation:

$$\widehat{\mathbf{h}}_0^{N+1} = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0^{N+1} \mid x_1^J, \tilde{\mathbf{f}}_1^N) = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N, \mathbf{h}_{N+1} \mid x_1^J)$$

Interactive Machine Translation

Segment-based IMT: formalization

$$\tilde{\mathbf{f}}_1^N = \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$$

$$\widehat{\mathbf{h}}_0^{N+1} = \widehat{\mathbf{h}}_0, \ldots, \widehat{\mathbf{h}}_{N+1}$$

Translation segments generation:

$$\widehat{\mathbf{h}}_0^{N+1} = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0^{N+1} \mid x_1^J, \tilde{\mathbf{f}}_1^N) = \underset{\mathbf{h}_0^{N+1}}{\arg\max}\, Pr(\mathbf{h}_0, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N, \mathbf{h}_{N+1} \mid x_1^J)$$

Suffix generation:

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

Interactive Machine Translation
Segment-based IMT: implementation

Our proposal relies on the XML scheme of *Moses* decoder (Koehn et al., 2007).

Interactive Machine Translation
Segment-based IMT: implementation

Our proposal relies on the XML scheme of *Moses* decoder (Koehn et al., 2007).

Sentence to translate: *La commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans.*

Interactive Machine Translation

Segment-based IMT: implementation

Our proposal relies on the XML scheme of *Moses* decoder (Koehn et al., 2007).

Sentence to translate: *La commission a constaté que les mesures relatives aux contrats temporaires inférieurs à deux ans.*

```
<x translation=“The commission” >La commission</x>
<x translation=“finds” >a constaté</x>
<x translation=“that the measures relating to
contracts” >que les mesures relatives aux contrats</x>
temporaires inférieurs à <x translation=“two years” >deux
ans</x>
```

Interactive Machine Translation
Segment-based IMT with active prediction

Given a source sentence $x_1^J = x_1, \ldots, x_J$ and its translation hypothesis $y_1^I = y_1, \ldots, y_I$, the confidence value of a word y_i ($c(y_i)$) is given by:

$$c(y_i) = \max_{1 \leq j \leq J} p(y_i \mid x_j)$$

Lexicon probabilities given by IBM Model 1 (Brown et al., 1993) or hidden Markov alignment models (Vogel et al., 1996).

Interactive Machine Translation

Segment-based IMT: experimental framework

Corpora:

- EMEA. Medical domain. Fr–En, De–En. 1M segments.
- EU. Legal domain. Es–En, Fr–En. 200K/1M segments.
- TED. Public speeches. Zh–En, Es–En. 150K segments.
- Xerox. Technical domain. Fr–En, Es–En. 50K segments.
- Europarl. Legal domain. Fr–En, De–En. 2M segments.

Interactive Machine Translation
Segment-based IMT: experimental framework

Metrics:

- Word Stroke Ration (WSR) (Tomás and Casacuberta, 2006).
- Mouse Action Ration (MAR) (Barrachina et al., 2009).
- Translation Error Rate (TER) (Snover et al., 2006).
- BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002).

Interactive Machine Translation

Segment-based IMT: experimental framework

User simulation:

- Prefix-based: hypothesis and reference comparison to detect the leftmost wrong word.
- Segment-based:
  - Longest common subsequence (Apostolico and Guerra, 1987) between hypothesis and reference.
  - Check if any pair of consecutive validated segments should be merged into a single segment.
  - Hypothesis and reference comparison to detect the leftmost wrong word.

Interactive Machine Translation

Segment-based IMT: evaluation

Main approaches

				Prefix-based		Segment-based	
Corpus	**Language**	**BLEU [↑]**	**TER [↓]**	**WSR [↓]**	**MAR [↓]**	**WSR [↓]**	**MAR [↓]**
EMEA	De–En	23.4	57.6	70.9	14.1	31.0	24.4
	En–De	15.7	64.8	74.9	12.0	35.6	23.1
EU	Es–En	47.3	40.8	45.6	10.2	30.5	16.0
	En–Es	47.9	41.1	44.6	9.7	31.9	14.8
TED	Zh–En	11.7	76.2	83.1	22.4	36.1	35.8
	En–Zh	8.7	83.3	86.3	55.7	60.0	80.0
Xerox	De–En	32.2	54.6	62.7	15.1	29.2	26.9
	En–De	24.1	64.5	68.3	12.6	32.7	23.6
Europarl	De–En	19.2	61.1	73.3	17.7	34.4	30.8
	En–De	15.3	68.4	75.0	15.0	33.1	25.9

Interactive Machine Translation

Segment-based IMT: evaluation

Active prediction

				Segment-based with active prediction					
		Segment-based		IBM$_1$		HMM		Random	
Corpus	**Language**	**WSR [↓]**	**MAR [↓]**	**WSR [↓]**	**MAR [↓]**	**WSR [↓]**	**MAR [↓]**	**WSR [↓]**	**MAR [↓]**
EMEA	De–En	31.0	24.4	30.3	24.3	30.7	24.6	30.0	24.1
	En–De	35.6	23.1	35.0	22.6	35.2	22.6	34.7	22.6
EU	Es–En	30.5	16.0	30.7	17.6	31.2	17.2	31.0	17.0
	En–Es	31.9	14.8	31.2	16.7	31.6	16.0	31.7	15.8
TED	Zh–En	36.1	35.8	35.8	35.4	35.9	35.4	34.9	35.0
	En–Zh	60.0	80.0	60.3	85.5	60.9	83.3	60.9	81.8
Xerox	De–En	29.2	26.9	29.3	26.7	29.2	26.6	29.0	26.5
	En–De	32.7	23.6	32.1	22.6	32.3	22.5	32.0	22.7
Europarl	De–En	34.4	30.8	34.3	30.7	34.5	30.7	33.6	30.2
	En–De	33.1	25.9	32.6	25.4	32.6	25.4	32.1	25.3

Interactive Machine Translation
Neural IMT (INMT) vs IMT

		Prefix-based						Segment-based					
		INMT$_{\text{RNN}}$		INMT$_{\text{Trans.}}$		IMT		INMT$_{\text{RNN}}$		INMT$_{\text{Trans.}}$		IMT	
		WSR [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
TED	Zh–En	54.9	14.2	60.1	14.3	83.1	22.4	51.2	21.2	49.2	20.4	36.1	35.8
	En–Zh	68.1	28.9	66.7	29.6	86.3	55.7	58.4	64.2	56.6	62.5	60.0	80.0
Xerox	De–En	38.4	9.4	42.2	10.0	62.7	15.1	35.1	13.3	39.9	14.1	29.2	26.9
	En–De	55.1	10.8	56.5	11.2	68.3	12.6	50.9	14.9	54.7	16.0	32.7	23.6

Outline

1. Interactive Machine Translation. (Chapter 3.)

2. Historical Document Processing. (Chapters 4 and 5.)

3. IMT for the Processing of Historical Documents. (Chapter 6.)

4. Conclusions

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	The question is: is it better to be alive or dead?
Whether tis nobler in the mind to suffer	Is it nobler to put up
The slings and arrows of outrageous fortune,	with all the nasty things that luck throws your way,
Or to take arms against a sea of troubles,	or to fight against all those troubles
And, by opposing, end them?	by simply putting an end to them once and for all?

Language Modernization

Approaches:

- Statistical machine translation (SMT).
- Neural machine translation (NMT).
  - Recurrent neural networks with long short-term memory units (LSTM).
  - Transformer.
- NMT enriched with modern documents.
  - Synthetic data generated through backtranslation.

Language Modernization
Experimental framework

Corpora:

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

Metrics:

- TER.
- BLEU.

Language Modernization

Experimental framework

Evaluation:

- Automatic metrics.
- Human evaluation.
  - Scholars (4 Scholars specialized in classic Spanish literature).
  - Non-experts (42 participants).

Language Modernization

Evaluation: automatic metrics

Approach	**Dutch Bible**		**El Quijote**		**OE-ME**	
	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]	TER [↓]	BLEU [↑]
Baseline	57.9	12.9	44.2	36.3	91.0	2.8
SMT	11.5	77.5	**30.7**†	**58.3**†	**39.6**†	**39.6**†
NMT$_{\text{LSTM}}$	13.8	79.6	55.1	39.8	82.7	12.8
NMT$_{\text{Transformer}}$	**11.1**†	**81.7**†	38.4	49.3	54.7	27.3
Enriched NMT$_{\text{LSTM}}$	**11.1**†	**80.6**†	**31.9**†	**57.3**†	**44.3**†	**35.9**†
Enriched NMT$_{\text{Transformer}}$	18.2	70.6	36.7	51.0	47.2	31.0

All results are significantly different between all approaches except those denoted with†.

Language Modernization

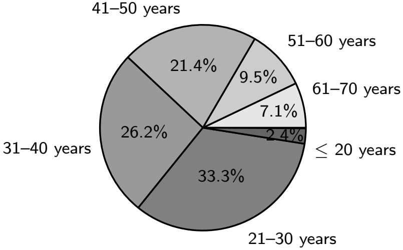
Evaluation: scholars

- **Fluency**: how fluid does the modernized sentence sound?
- **Lexical meaning**: how correct is the lexicon of the modernized sentence?
- **Syntax**: how correct is the syntactic construction of the modernized sentence?
- **Semantic**: is the meaning of the original sentence preserved in the modernized sentence?
    - **1**: the meaning is lost.
    - **2**: a great part of the meaning is lost.
    - **3**: half the meaning is lost.
    - **4**: part of the meaning is lost.
    - **5**: the meaning remains.
- **Modernization**: how appropriate is the modernization?

Language Modernization
Evaluation: scholars

	Fluency	Lexical meaning	Syntax	Semantic	Modernization
SMT	3.7	3.3	3.4	3.5	3.2
En. NMT_{LSTM}	3.7	3.3	3.4	3.5	3.2

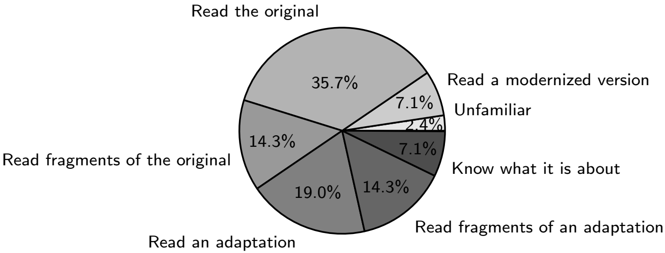
Language Modernization
Evaluation: non-experts



Age distribution.

Language Modernization

Evaluation: non-experts



Familiarity with *El Quijote*.

Language Modernization
Evaluation: non-experts

	Original	Modernized	Indifferent	Not equal
SMT	3.2	61.4	27.6	7.8
NMT	6.4	50.9	22.3	20.3

Percentage of cases in which the users selected that option.

Spelling Normalization

Goal: achieve an orthography consistency by adapting a document's spelling to modern standards.

Spelling Normalization

Goal: achieve an orthography consistency by adapting a document's spelling to modern standards.

Original

"Nunca fuera cauallero
de damas tambien seruido,
como fuera don Quixote
quando de su aldea vino:
donzellas curauan del,
princesas del su rozino."

Normalized

"Nunca fuera caballero
de damas tan bien servido,
como fuera don Quijote
cuando de su aldea vino:
doncellas curaban de él,
princesas del su rocino."

Spelling Normalization

Approaches:

- Statistical dictionary (SD).
- SMT.
- NMT.
  - LSTM.
  - Transformer.
- Character-based (CB) SMT.
- CBNMT.
  - CBNMT.
  - SubChar (Subwords–Characters).
  - CharSub (Characters–Subwords).
- CBNMT enriched with modern documents.
  - Synthetic data generated through backtranslation.

Spelling Normalization
Experimental framework

Corpora:

- Entremeses y Comedias (17$^{\text{th}}$ century Spanish; 35K segments).
- Quijote (17$^{\text{th}}$ century Spanish; 48K segments).
- Bohorič (18$^{\text{th}}$ century Slovene; 4K segments).
- Gaj (19$^{\text{th}}$ century Slovene; 13K segments).

Metrics:

- Character Error Rate (CER).
- TER.
- BLEU.

Spelling Normalization

Main approaches

System	**Quijote**			**Bohorič**		
	CER	TER	BLEU	CER	TER	BLEU
	[↓]	[↓]	[↑]	[↓]	[↓]	[↑]
Baseline	7.9	19.5	59.4	21.7	49.0	18.0
SD	3.9	5.5	89.3	16.2	20.7	56.1
CBSMT	**2.5**†	**3.0**†	**94.4**†	**2.4**	**8.7**	**80.4**
CBNMT$_{\text{LSTM}}$	2.7	4.3‡	93.3‡	29.4	39.5	48.7
En. CBNMT$_{\text{LSTM}}$	**2.2**†	4.0‡	93.2‡	28.6	38.3	49.5
CBNMT$_{\text{Trans.}}$	**1.9**†	**3.3**†	**93.9**†	26.2†	30.6†	60.0†
En. CBNMT$_{\text{Trans.}}$	**2.4**†	5.1	89.7	25.7†	29.8†	60.8†

All results are significantly different between all approaches except those denoted with † and ‡ (respectively).

Spelling Normalization
Additional CBNMT approaches

System	**Quijote**			**Bohorič**		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
En. CBNMT$_{\text{LSTM}}$	**2.2**†	4.0†	93.2‡	28.6‡	38.3	49.5
En. SubChar$_{\text{LSTM}}$	**2.3**†	**3.3**‡	**94.9**†	29.5†	36.9	51.5
En. CharSub$_{\text{LSTM}}$	**2.3**†	4.1†	93.0‡	27.5*	39.6†	47.2
En. CBNMT$_{\text{Trans.}}$	**2.4**†	5.1	89.7	25.7	29.8‡	60.8†
En. SubChar$_{\text{Trans.}}$	**2.4**†	**3.2**‡	**94.4**†	27.3*	31.8	57.8
En. CharSub$_{\text{Trans.}}$	**2.4**†	**3.5**‡	93.9‡	**8.8**	**11.5**	**79.3**

All results are significantly different between all approaches except those denoted with†, ‡ and * (respectively).

Outline

1. Interactive Machine Translation. (Chapter 3.)
2. Historical Document Processing. (Chapters 4 and 5.)
3. IMT for the Processing of Historical Documents. (Chapter 6.)
4. Conclusions

Language Modernization

Goal: Help scholars to generate error-free modernizations.

Language Modernization

Goal: Help scholars to generate error-free modernizations.

Approaches:

- SMT.
- Enriched NMT.
  - LSTM.
  - Transformer.

IMT:

- Prefix-based.
- Segment-based.

Language Modernization

Goal: Help scholars to generate error-free modernizations.

Approaches:

- SMT.
- Enriched NMT.
  - LSTM.
  - Transformer.

IMT:

- Prefix-based.
- Segment-based.

Online demonstrator: `http://demosmt.prhlt.upv.es/mthd/`.

Language Modernization

Experimental framework

Corpora:

- Dutch Bible (17$^{\text{th}}$ century Dutch).
- El Quijote (17$^{\text{th}}$ century Spanish).
- OE-ME (11$^{\text{th}}$ century English).

Metrics:

- WSR.
- MAR.
- TER.
- BLEU.

Language Modernization

Evaluation

		Modernization quality		Prefix-based		Segment-based	
Corpus	Approach	TER [↓]	BLEU [↑]	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]
El Quijote	SMT	30.7	58.3	38.8	10.9	**22.0**	**19.7**
	En. NMT_{LSTM}	42.9	50.4	68.9‡	11.8	68.9‡	47.8
	En. $\text{NMT}_{\text{Transformer}}$	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**
	En. NMT_{LSTM}	56.4	30.3	72.1‡	12.8†	72.1‡	59.5
	En. $\text{NMT}_{\text{Transformer}}$	58.9	28.2	73.5‡	13.3†	73.5‡	49.5

All results are significantly different between all approaches except those denoted with † and ‡ (respectively).

Spelling Normalization

Goal: Help scholars to generate error-free normalizations.

Spelling Normalization

Goal: Help scholars to generate error-free normalizations.

Approaches:

- CBSMT.
- Enriched CBNMT.
    - LSTM.
    - Transformer.

IMT:

- Prefix-based.
- Segment-based.

Spelling Normalization

Goal: Help scholars to generate error-free normalizations.

Approaches:

- CBSMT.
- Enriched CBNMT.
  - LSTM.
  - Transformer.

IMT:

- Prefix-based.
- Segment-based.

Online demonstrator: `http://demosmt.prhlt.upv.es/mthd/`.

Spelling Normalization
Experimental framework

Corpora:

- Entremeses y Comedias (17$^{\mathrm{th}}$ century Spanish).
- Quijote (17$^{\mathrm{th}}$ century Spanish).
- Bohorič (18$^{\mathrm{th}}$ century Slovene).
- Gaj (19$^{\mathrm{th}}$ century Slovene).

Metrics:

- Key Stroke Ratio (KSR) (Tomás and Casacuberta, 2006).
- MAR.
- CER.
- TER.
- BLEU.

Spelling Normalization

Evaluation

		Normalization quality			Prefix-based		Segment-based	
Corpus	**Approach**	**CER [↓]**	**TER [↓]**	**BLEU [↑]**	**KSR [↓]**	**MAR [↓]**	**KSR [↓]**	**MAR [↓]**
Entremeses y Comedias	CBSMT	1.3†	4.4	91.7	**0.9**‡	**4.1**	0.7‡	6.7
	En. CBNMT$_{\text{LSTM}}$	3.5	9.4	84.9	1.9‡	2.1†	1.9‡	3.3
	En. CBNMT$_{\text{Transformer}}$	1.5†	6.5	87.2	1.4‡	2.1†	1.4‡	3.4
Quijote	CBSMT	2.5†	3.0†	94.4†	1.4†‡	3.7	1.1†‡	5.3
	En. CBNMT$_{\text{LSTM}}$	2.6†	4.3	93.9†	**1.4**†	**1.4**†‡	1.4†‡	2.1
	En. CBNMT$_{\text{Transformer}}$	2.2†	3.7†	94.4†	**1.5**†‡	**1.4**†	1.5†‡	2.1

All results are significantly different between all approaches except those denoted with† and ‡ (respectively).

Outline

1. Interactive Machine Translation. (Chapter 3.)

2. Historical Document Processing. (Chapters 4 and 5.)

3. IMT for the Processing of Historical Documents. (Chapter 6.)

4. Conclusions

Scientific contributions

IMT:

- Developed a new protocol to allow the user to validate the correct parts of a translation hypothesis.
- Wide experimentation that showcases a substantial decrease of the typing effort.
- Tested an active prediction protocol to assist the user in the correction step.
- Applied IMT to two task related with the processing of historical documents.

Scientific contributions

Language modernization:

- Proposed several modernization approaches based on SMT and NMT.
- Conducted a wide experimentation, which counted with the help of 4 scholars and 42 volunteers.

Spelling normalization:

- Proposed several normalization approaches based on SMT, NMT, CBSMT and CBNMT.
- Evaluated our approaches using different datasets from different time periods and languages.

Publications derived from the thesis

- *Modernizing historical documents: A user study*. PRL. **JCR Q2**.
- *Interactive neural machine translation*. CSL. **Second author; JCR Q2**.
- *Segment-based interactive-predictive machine translation*. MTJ. **Peer-reviewed journal**.
- *The CLIN27 shared task: Translating historical text to contemporary language for improving automatic linguistic annotation*. CLIN. **Alphabetical order**; **Peer-reviewed journal**.
- *Two demonstrations of the machine translation applications to historical documents*. ICPR. **CORE B**.
- *Spelling normalization of historical documents by using a machine translation approach*. EAMT. **CORE B**.

Publications derived from the thesis

- *Historical documents modernization*. EAMT. **CORE B**.
- *Interactive-predictive translation based on multiple word-segments*. EAMT. **CORE B**. **Best paper award**.
- *A machine translation approach for modernizing historical documents using back translation*. IWSLT. **Peer-reviewed workshop**.
- *A comparison of character-based neural machine translations techniques applied to spelling normalization*. PatReCH. **Peer-reviewed workshop**.
- *Enriching character-based neural machine translation with modern documents for achieving an orthography consistency in historical documents*. PatReCH. **Peer-reviewed workshop**.

Other publications

- *How much does tokenization affect neural machine translation?* CICLing. **CORE B**.
- *A user study of the incremental learning in NMT*. EAMT. **CORE B**.
- *Demonstration of a neural machine translation system with online learning for translators*. ACL. **CORE A+**.
- *Incremental adaptation of NMT for professional post-editors: A user study*. MT Summit. **CORE B**.

Future work

IMT:

- Improve how the system deals with user corrections.
- Develop new protocols to assist the user in the validation step.

Language modernization:

- Tackle the main problems that were pointed out during the evaluation: punctuation, diacritical marks, etc.

Spelling normalization:

- Better profit from modern documents to enrich the systems.
- Human evaluation.
- Try new neural architectures.

Bibliography

Apostolico, A. and Guerra, C. (1987). The longest common subsequence problem revisited. Algorithmica, 2:315336.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Toms, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. Computational Linguistics, 35:328.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263311.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 177180.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 311318.

Bibliography

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the Association for Machine Translation in the Americas, pages 223231.

Tomás, J. and Casacuberta, F. (2006). Statistical phrase-based models for interactive computer-assisted translation. In Proceedings of the International Conference on Computational Linguistics/Association for Computational Linguistics, pages 835841.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In Proceedings of the Conference on Computational Linguistics, volume 2, pages 836841.

Interactive Machine Translation

Segment-based IMT: XML generation

Interactive Machine Translation

Segment-based IMT: XML generation

User actions:

Segment validation:

Words deletion:

Word correction:

Interactive Machine Translation

Segment-based IMT: XML generation

User actions:

Segment validation: for each validated segment, align the target words with the source words (phrase alignments).

Words deletion:

Word correction:

Interactive Machine Translation
Segment-based IMT: XML generation

User actions:

Segment validation: for each validated segment, align the target words with the source words (phrase alignments).

Words deletion: merge the segments into a single tag.

Word correction:

Interactive Machine Translation

Segment-based IMT: XML generation

User actions:

Segment validation: for each validated segment, align the target words with the source words (phrase alignments).

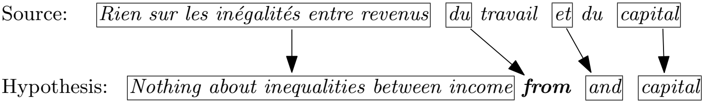
Words deletion: merge the segments into a single tag.

Word correction: compute alignment probability (using hidden Markov alignment models[1]) between new word and all non-validated source words.

[1] Stephan Vogel et al. (1996). "HMM-based Word Alignment in Statistical Translation". In: *Proceedings of the Conference on Computational Linguistics*. Vol. 2, pp. 836–841.

Interactive Machine Translation
Segment-based IMT: XML generation

Segment reorders



Source: *Rien sur les inégalités entre revenus* *du* *travail* *et* *du* *capital*

Hypothesis: *Nothing about inequalities between income* ***from*** *and* *capital*

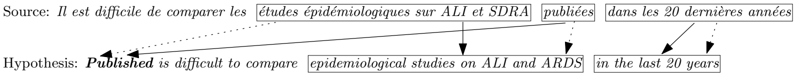
XML: <x translation="Nothing about the inequalities between income" > Rien sur les inégalités entre revenus</x><wall/> <x translation="from" >du</x><wall/> travail <x translation="and" > et</x><wall/> du <x translation="capital" >capital</x><wall/>

Translation: *Nothing about inequalities between income* *from* *work* *and* *capital*

Interactive Machine Translation

Segment-based IMT: XML generation

Segment reorders


Source: *Il est difficile de comparer les* *études épidémiologiques sur ALI et SDRA* *publiées* *dans les 20 dernières années*

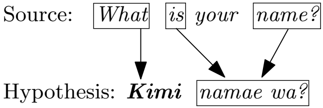
Hypothesis: ***Published*** *is difficult to compare* *epidemiological studies on ALI and ARDS* *in the last 20 years*

XML: Il est difficile de comparer les
<x translation=“Published” >études épidémiologiques sur ALI et
SDRA</x><wall/> <x translation= “epidemiological studies on ALI and
ARDS” >publiées</x><wall/> <x translation=“in the last 20
years” >dans les dernières 20 années</x><wall/>

Translation: *It is difficult to compare the* *Published* *epidemiological studies on ALI and ARDS* *in the last 20 years*

Interactive Machine Translation
Segment-based IMT: XML generation

Non-consecutive corresponding sources



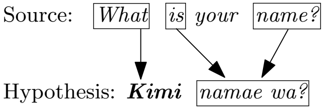
XML: `<x translation="Kimi" >What</x> <x translation="wa" >is</x> your <x translation="namae?" >name?</x>`

Translation: *Kimi* *wa* *no* *namae?*

Interactive Machine Translation

Segment-based IMT: XML generation

Non-consecutive corresponding sources


Source: *What* *is* *your* *name?*
Hypothesis: ***Kimi*** *namae wa?*

XML: <x translation="Kimi" >What</x> <x translation="wa" >is</x> your <x translation="namae?" >name?</x>

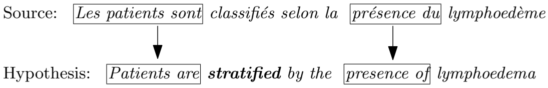
Translation: *Kimi* *wa* *no* *namae?*

<u>Solution</u>:
XML: <x translation="Kimi" >What</x> <x translation="namae wa?" >is</x> your <x translation="" >name?</x>

Translation: *Kimi* *no* *namae wa?*

Interactive Machine Translation
Segment-based IMT: XML generation

Words without corresponding source segment



Source: *Les patients sont* *classifiés selon la* *présence du* *lymphoedème*

Hypothesis: *Patients are* ***stratified*** *by the* *presence of* *lymphoedema*

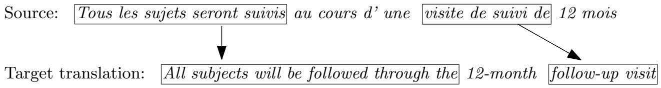
XML: <x translation=“Patients are” >Les patients sont</x> classifiés selon la <x translation=“presence of” >présence du</x> lymphoedème <x translation=“stratified” >.</x>

Translation: *Patients are* *classifiés* *stratified* *by the* *presence of* *lymphoedema*

Interactive Machine Translation

Segment-based IMT: XML generation

Spurious words



Source: *Tous les sujets seront suivis* *au cours d' une* *visite de suivi de* *12 mois*

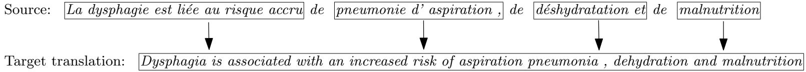
Target translation: *All subjects will be followed through the* *12-month* *follow-up visit*

Hypothesis: *All subjects will be followed through the* *course of a* *12-month* *12 months* *follow-up visit*

User feedback: *All subjects will be followed through the 12-month follow-up visit*

Interactive Machine Translation
Segment-based IMT: XML generation

Spurious words



Source: *La dysphagie est liée au risque accru* *de* *pneumonie d' aspiration ,* *de* *déshydratation et* *de* *malnutrition*

Target translation: *Dysphagia is associated with an increased risk of aspiration pneumonia , dehydration and malnutrition*

Hypothesis: *Dysphagia is associated with* *an* *increased risk of aspiration pneumonia , dehydration* *and malnutrition* *of of of*

User feedback: *Dysphagia is associated with* *an* *increased risk of aspiration pneumonia , dehydration* *and malnutrition* #

Interactive Machine Translation

Segment-based IMT: experimental framework

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Interactive Machine Translation

Segment-based IMT: experimental framework

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: $2 + 1 + 2 = 5$

Interactive Machine Translation

Segment-based IMT: experimental framework

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: $2 + 1 + 2 = 5$

Words deletion: $\boxed{\text{If you have been exposed , you should }\cancel{\text{consult}}\text{ go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: 1

Interactive Machine Translation

Segment-based IMT: experimental framework

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: $2 + 1 + 2 = 5$

Words deletion: $\boxed{\text{If you have been exposed , you should \sout{consult} go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: 1

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
Mouse actions: 1
Word strokes: 1

Interactive Machine Translation

Segment-based IMT: experimental framework

Reference: If you have been exposed , you should go to your doctor for tests
Hypothesis: If you have been exposed , you should consult go your doctor for tests

Segment validation: $\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: $2 + 1 + 2 = 5$

Words deletion: $\boxed{\text{If you have been exposed , you should } \cancel{\text{consult}} \text{ go}}$ $\boxed{\text{your doctor for tests}}$
Mouse actions: 1

Word correction: $\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
Mouse actions: 1
Word strokes: 1

Total mouse actions: 7
Total word strokes: 1

Interactive Machine Translation

Segment-based IMT: evaluation

Main approaches

				Prefix-based		Segment-based	
Corpus	Language	BLEU [↑]	TER [↓]	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]
EMEA	Fr–En	30.5	48.6	57.8	12.4	33.6	21.6
	En–Fr	29.8	52.6	58.4	12.5	41.7	21.7
	De–En	23.4	57.6	70.9	14.1	31.0	24.4
	En–De	15.7	64.8	74.9	12.0	35.6	23.1
EU	Es–En	47.3	40.8	45.6	10.2	30.5	16.0
	En–Es	47.9	41.1	44.6	9.7	31.9	14.8
	Fr–En	52.1	36.2	37.3	7.5	26.3	14.4
	En–Fr	51.3	38.6	38.8	7.3	29.4	12.8
TED	Zh–En	11.7	76.2	83.1	22.4	36.1	35.8
	En–Zh	8.7	83.3	86.3	55.7	60.0	80.0
	Es–En	36.5	42.7	51.1	12.9	31.7	22.9
	En–Es	31.3	47.7	53.2	12.3	36.7	22.8
Xerox	Es–En	52.2	31.8	35.8	10.5	20.0	20.4
	En–Es	60.8	27.3	28.3	7.9	21.9	14.3
	De–En	32.2	54.6	62.7	15.1	29.2	26.9
	En–De	24.1	64.5	68.3	12.6	32.7	23.6
Europarl	Fr–En	26.5	51.4	58.7	13.9	30.2	30.3
	En–Fr	26.5	55.6	61.4	13.5	31.5	28.4
	De–En	19.2	61.1	73.3	17.7	34.4	30.8
	En–De	15.3	68.4	75.0	15.0	33.1	25.9

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

Interactive Machine Translation
Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go}}$ consult your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	If you have been exposed , you should **go** your doctor for tests
	MT	If you have been exposed , you should go consult your doctor for tests
IT-2	**User**	If you have been exposed , you should go **to** your doctor for tests
	MT	If you have been exposed , you should go to consult your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go}}$ consult your doctor for tests
IT-2	**User**	$\boxed{\text{If you have been exposed , you should go}}$ **to** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go to}}$ consult your doctor for tests
IT-3	**User**	$\boxed{\text{If you have been exposed , you should go to}}$ **your** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go to your}}$ doctor for tests

Interactive Machine Translation
Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go}}$ consult your doctor for tests
IT-2	**User**	$\boxed{\text{If you have been exposed , you should go}}$ **to** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go to}}$ consult your doctor for tests
IT-3	**User**	$\boxed{\text{If you have been exposed , you should go to}}$ **your** your doctor for tests
	MT	$\boxed{\text{If you have been exposed , you should go to your}}$ doctor for tests
END	**User**	If you have been exposed , you should go to your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation ($\hat{\mathbf{y}}$): If you have been exposed , you should go to your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests

Interactive Machine Translation
Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	If you have been exposed , you should **go** your doctor for tests
	MT	If you have been exposed , you should consult go your doctor for tests

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation ($\hat{y}$): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
IT-2	**User**	$\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{to}}$ $\boxed{\text{your doctor for tests}}$

Interactive Machine Translation

Segment-based IMT: evaluation

source (x): Si vous avez été exposé , vous devriez consulter votre médecin pour des tests

target translation (ŷ): If you have been exposed , you should go to your doctor for tests

IT-0	**MT**	If you have been exposed , you should consult your doctor for tests
IT-1	**User**	$\boxed{\text{If you have been exposed , you should}}$ **go** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should}}$ consult $\boxed{\text{go}}$ $\boxed{\text{your doctor for tests}}$
IT-2	**User**	$\boxed{\text{If you have been exposed , you should go}}$ **to** $\boxed{\text{your doctor for tests}}$
	MT	$\boxed{\text{If you have been exposed , you should go}}$ $\boxed{\text{to}}$ $\boxed{\text{your doctor for tests}}$
END	**User**	If you have been exposed , you should go to your doctor for tests

Interactive Machine Translation
Segment-based IMT: evaluation

Active prediction

				Segment-based with active prediction					
		Segment-based		IBM$_1$		HMM		Random	
Corpus	Language	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]	WSR [↓]	MAR [↓]
EMEA	Fr–En	33.6	21.6	35.1	23.4	35.5	22.9	35.7	22.8
	En–Fr	41.7	21.7	41.2	23.3	41.8	22.5	41.9	22.0
	De–En	31.0	24.4	30.3	24.3	30.7	24.6	30.0	24.1
	En–De	35.6	23.1	35.0	22.6	35.2	22.6	34.7	22.6
EU	Es–En	30.5	16.0	30.7	17.6	31.2	17.2	31.0	17.0
	En–Es	31.9	14.8	31.2	16.7	31.6	16.0	31.7	15.8
	Fr–En	26.3	14.4	26.9	15.7	27.2	15.5	27.2	15.4
	En–Fr	29.4	12.8	29.4	13.8	29.6	13.7	29.6	13.5
TED	Zh–En	36.1	35.8	35.8	35.4	35.9	35.4	34.9	35.0
	En–Zh	60.0	80.0	60.3	85.5	60.9	83.3	60.9	81.8
	Es–En	31.7	22.9	32.0	24.7	32.3	24.4	32.2	24.2
	En–Es	36.7	22.8	36.6	24.7	37.1	24.0	37.1	23.7
Xerox	Es–En	20.0	20.4	20.1	20.4	20.1	20.5	19.9	20.1
	En–Es	21.9	14.3	22.3	15.2	22.6	14.9	22.6	14.7
	De–En	29.2	26.9	29.3	26.7	29.2	26.6	29.0	26.5
	En–De	32.7	23.6	32.1	22.6	32.3	22.5	32.0	22.7
Europarl	Fr–En	30.2	30.3	29.8	29.7	29.8	29.7	29.4	29.6
	En–Fr	31.5	28.4	30.9	27.7	31.1	27.6	30.4	27.5
	De–En	34.4	30.8	34.3	30.7	34.5	30.7	33.6	30.2
	En–De	33.1	25.9	32.6	25.4	32.6	25.4	32.1	25.3

Interactive Machine Translation

Neural IMT (INMT) vs IMT

		Prefix-based						**Segment-based**					
		INMT$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**		**INMT**$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**	
		WSR [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]	**WSR** [↓]	**MAR** [↓]
TED	Zh–En	54.9	14.2	60.1	14.3	83.1	22.4	51.2	21.2	49.2	20.4	36.1	35.8
	En–Zh	68.1	28.9	66.7	29.6	86.3	55.7	58.4	64.2	56.6	62.5	60.0	80.0
Xerox	Es–En	30.7	7.2	37.4	8.3	35.8	10.5	29.1	12.5	35.5	13.2	20.0	20.4
	En–Es	28.4	7.3	32.1	8.0	28.3	7.9	22.7	7.5	30.2	12.7	21.9	14.3
	De–En	38.4	9.4	42.2	10.0	62.7	15.1	35.1	13.3	39.9	14.1	29.2	26.9
	En–De	55.1	10.8	56.5	11.2	68.3	12.6	50.9	14.9	54.7	16.0	32.7	23.6

		Translation quality					
		INMT$_{\text{RNN}}$		**INMT**$_{\text{Trans.}}$		**IMT**	
		BLEU [↑]	**TER** [↓]	**BLEU** [↑]	**TER** [↓]	**BLEU** [↑]	**TER** [↓]
TED	Zh–En	13.7	75.7	11.5	76.7	11.7	76.2
	En–Zh	9.3	76.7	8.2	77.6	8.7	83.3
Xerox	Es–En	59.0	28.6	53.9	32.1	52.2	31.8
	En–Es	63.5	27.5	60.5	28.3	60.8	27.3
	De–En	36.2	51.1	31.3	54.9	32.2	54.6
	En–De	25.4	63.0	23.2	64.3	24.1	64.5

Language Modernization
Evaluation: scholars

Scholar	**SMT approach**				
	Fluency	**Lexical meaning**	**Syntax**	**Semantic**	**Modernization**
Scholar$_1$	5.0	4.3	4.3	4.6	3.9
Scholar$_2$	2.1	1.9	2.0	2.1	2.0
Scholar$_3$	3.2	3.1	2.9	2.9	3.1
Scholar$_4$	4.5	3.9	4.6	4.3	4.0
Average	3.7	3.3	3.4	3.5	3.2
	Enriched NMT$_{\text{LSTM}}$ approach				
	Fluency	**Lexical meaning**	**Syntax**	**Semantic**	**Modernization**
Scholar$_1$	4.8	4.0	4.0	4.1	4.0
Scholar$_2$	2.0	1.9	1.9	1.9	1.9
Scholar$_3$	3.3	3.2	2.9	3.0	3.1
Scholar$_4$	3.8	3.5	3.7	3.7	3.5
Average	3.7	3.3	3.4	3.5	3.2

Language Modernization

Evaluation: non-experts

Select the sentence which is easier for you to read and comprehend:

○ Riose don Quixote, y pidio que quitassen otro lieno, debaxo del qual se descubrio la imagen del patron de las Españas a cauallo, la espada ensangrentada, atropellando moros y pisando cabeças, y, en viendola, dixo don Quixote:

○ Se rió don Quijote, y pidió que quitasen otro lienzo, debajo del cual se descubrió la imagen del patrón de las Españas a caballo, la espada ensangrentada, atropellando moros y pisando cabezas y viéndola, dijo don Quijote:

○ Indifferent.

○ Both sentences do not have the same meaning.

Spelling Normalization

Main approaches

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
SD	7.8	18.9	66.8	3.9	5.5	89.3	16.2	20.7	56.1	7.6	8.8	79.8
SMT	6.7	8.0	82.1	5.3‡	4.5	91.1	9.0	15.1	63.0	2.8	5.2	82.6
NMT$_{\text{LSTM}}$	18.0	15.2	72.2	10.2	8.1	84.4	41.4	33.9	36.7	36.0	28.3	50.4
NMT$_{\text{Trans.}}$	27.5	43.9	34.3	5.5‡	18.5	60.6	43.2	66.4	12.6	12.0	18.4	68.8
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$_{\text{LSTM}}$	1.7‡	12.0	82.7	2.7	4.3‡	93.3‡	29.4	39.5	48.7	31.5	36.9	53.1
En. CBNMT$_{\text{LSTM}}$	1.7‡	13.3	79.4	**2.2**†	4.0‡	93.2‡	28.6	38.3	49.5	30.5	35.4	54.9
CBNMT$_{\text{Trans.}}$	**1.4**†	6.1	88.0	**1.9**†	**3.3**†	**93.9**†	26.2†	30.6†	60.0†	29.9†	32.1†	60.0†
En. CBNMT$_{\text{Trans.}}$	**1.1**†	5.1	89.7	**2.4**†	5.1	89.7	25.7†	29.8†	60.8†	30.0†	32.0†	60.2†

All results are significantly different between all approaches except those denoted with† and ‡ (respectively).

Spelling Normalization
Additional CBNMT approaches

System	Entremeses y Comedias			Quijote			Bohorič			Gaj		
	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]	CER [↓]	TER [↓]	BLEU [↑]
CBNMT$_{\text{LSTM}}$	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$_{\text{LSTM}}$	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$_{\text{LSTM}}$	5.8	18.2	75.2	3.7	5.8	89.8	67.9	83.8	5.3	37.2	48.1	36.3
En. CBNMT$_{\text{LSTM}}$	1.7†	13.3	79.4†	**2.2**†	4.0†	93.2‡	28.6‡	38.3	49.5	30.5†	**35.4**‡	**54.9**‡
En. SubChar$_{\text{LSTM}}$	37.8	35.8	59.3	**2.3**†	**3.3**‡	**94.9**†	29.5†	**36.9**	**51.5**	31.5	**35.9**‡	**54.3**‡
En. CharSub$_{\text{LSTM}}$	3.8	15.2	78.9†	**2.3**†	4.1†	93.0‡	**27.5***	39.6†	47.2	**29.4**	37.2†	52.3†
CBNMT$_{\text{Trans.}}$	**1.4**‡	6.1	88.0	**1.9**†	**3.3**‡	93.9‡	26.2	30.6‡	60.0†	29.9	32.1*	60.0*
SubChar$_{\text{Trans.}}$	21.2	33.1	64.8	2.6‡	**3.7**‡	93.5‡	28.6‡	33.4	55.2	30.9†	32.7*	59.2*
CharSub$_{\text{Trans.}}$	12.2	42.4	72.1	3.2	4.8	91.4	59.1	68.8	14.9	9.1	11.6	79.1
En. CBNMT$_{\text{Trans.}}$	**1.1**‡	**5.1**	**89.7**	**2.4**†	5.1	89.7	25.7	29.8‡	60.8†	30.0†	32.0*	60.2*
En. SubChar$_{\text{Trans.}}$	43.2	56.5	66.4	**2.4**†	**3.2**‡	**94.4**†	27.3*	31.8	57.8	30.6†	32.6*	59.1*
En. CharSub$_{\text{Trans.}}$	11.9	41.8	72.5	**2.4**†	**3.5**‡	93.9‡	**8.8**	**11.5**	**79.3**	**6.5**	**7.2**	**87.2**

All results are significantly different between all approaches except those denoted with†, ‡ and * (respectively).