Segment-Based Interactive Machine Translation

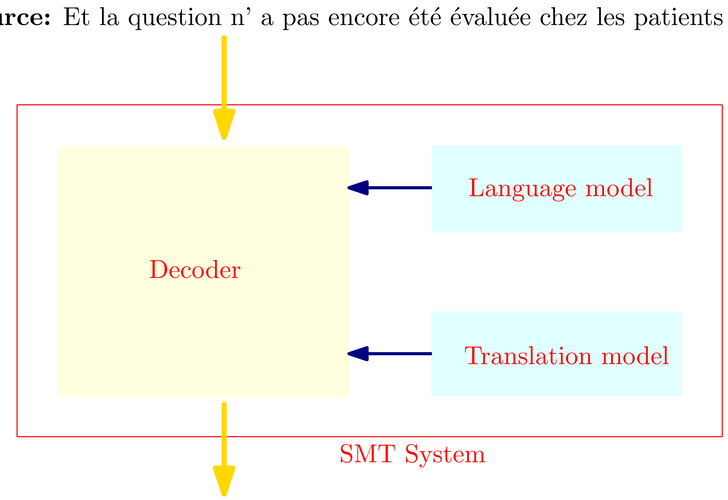
Authors: *Miguel Domingo*, Álvaro Peris **Supervisor:** Francisco Casacuberta
Ph.D. Program: Ph.D. in Computer Science
PRHLT Research Center {midobal, lvapeab, fcn}@prhlt.upv.es

Introduction

- Statistical Machine Translation (SMT) is still not good enough.
- Human correction is needed to improve translation quality.
- Interactive Machine Translation (IMT) aims at reducing the effort of this process.
- Prefix-based IMT was an interesting contribution to the field.

Statistical Machine Translation

Source: Et la question n' a pas encore été évaluée chez les patients atteints de cancer gastrique



Translation: And the issue has not been investigated among patients with gastric cancer

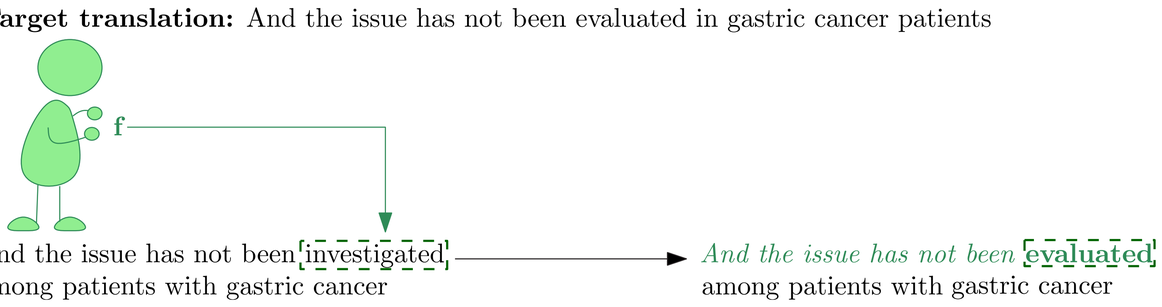
SMT main equation:

$$\tilde{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y}) \cdot Pr(\mathbf{x} \mid \mathbf{y})$$

$\mathbf{x}$: source
$\mathbf{y}$: translation
$Pr(\mathbf{y})$: language model
$Pr(\mathbf{x} \mid \mathbf{y})$: translation model

Prefix-Based IMT

Source: Et la question n' a pas encore été évaluée chez les patients atteints de cancer gastrique
Target translation: And the issue has not been evaluated in gastric cancer patients

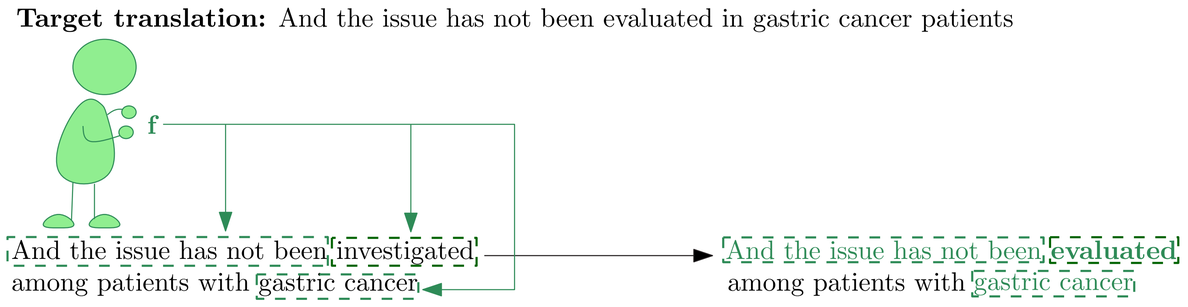


Prefix-based IMT session:

0	**S**	And the issue has not been investigated among patients with gastric cancer
1	**U**	And the issue has not been **evaluated** among patients with gastric cancer
	S	*And the issue has not been evaluated* with patients
2	**U**	*And the issue has not been evaluated* **in** patients
	S	*And the issue has not been evaluated in* patients
3	**U**	*And the issue has not been evaluated in* **gastric**
	S	*And the issue has not been evaluated in gastric* cancer patients
E	**U**	And the issue has not been evaluated in gastric cancer patients

Segment-Based IMT

Source: Et la question n' a pas encore été évaluée chez les patients atteints de cancer gastrique
Target translation: And the issue has not been evaluated in gastric cancer patients



Segment-based IMT session:

0	**S**	And the issue has not been investigated among patients with gastric cancer
1	**U**	And the issue has not been investigated among patients with gastric cancer
		And the issue has not been **evaluated** among patients with gastric cancer
	S	And the issue has not been evaluated in gastric cancer patients
E	**U**	And the issue has not been evaluated in gastric cancer patients

Statistical Framework

SMT fundamental equation:

$$\tilde{\mathbf{y}} = \arg\max_{\mathbf{y}} Pr(\mathbf{y} \mid \mathbf{x})$$

Segment-based IMT:

$$\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_N = \arg\max_{\mathbf{h}_1, \ldots, \mathbf{h}_N} Pr(\hat{\mathbf{f}}_1 \mathbf{h}_1, \ldots, \hat{\mathbf{f}}_N \mathbf{h}_N \mid \mathbf{x})$$

$\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_N$: sequence of N segments validated by the user (feedback signal).
$\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_N$: sequence of new translation segments.

$$\tilde{\mathbf{y}} = \hat{\mathbf{f}}_1 \tilde{\mathbf{h}}_1, \ldots, \hat{\mathbf{f}}_N \tilde{\mathbf{h}}_N$$

Search in the space of the translation, constrained by the sequence of segments $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_N$.

Experiments

- Simulated environment.
- Corpora from different domains.
- Comparison of prefix-based against segment-based IMT.
- Measuring human effort (WSR and MAR)

Results

Corpus	Language	BLEU	Prefix-Based		Segment-Based	
			WSR (%)	MAR (%)	WSR (%)	MAR (%)
EMEA	Fr–En	31.3	57.8	12.4	34.4	18.8
	En–Fr	30.2	58.4	12.5	40.4	16.3
EU	Es–En	48.2	45.6	10.2	28.3	15.0
	En–Es	48.7	44.6	9.7	29.8	13.5
TED	Zh–En	11.7	83.1	22.4	54.1	28.3
	En–Zh	8.7	86.3	55.7	59.2	72.4
Xerox	Es–En	54.5	35.8	10.5	23.2	16.9
	En–Es	62.2	28.3	7.9	22.1	12.5

- Substantial reduction of the typing effort (up to 29 points of WSR).
- Slight increase in the number of mouse actions (from 4 up to 6.5 points of MAR).

Conclusions

- New IMT approach that breaks down the prefix constraint.
- The user can select all correct words from each translation hypothesis.
- Human effort effectively reduced in a simulated environment.
- Future work: experiments with real users.

Acknowledgements

The research leading to these results has received funding from the Ministerio de Economía y Sostenibilidad (MINECO) under project SmartWays (grant agreement RTC-2014-1466-4), and Generalitat Valenciana under project ALMAMATER (grant agreement PROMETEOII/2014/030).