A Hyper-parameter Settings

`LSTM`

Configurations	Values
Embedding dimension D	512
Hidden dimension	512
Encoder layer	2
Decoder layer	2
Optimizer	SGD
Initial learning rate	1.0
Gradient clipping	5.0
Dropout rate	0.3
Mini-batch size	128 sentences

`Transformer`

Configurations	Values
Embedding dimension D	512
Hidden dimension	512
Number of attention heads	8
Encoder layer	6
Decoder layer	6
Optimizer	Adam
Warmup steps	4000
Gradient clipping	–
Dropout rate	0.1
Mini-batch size	3500 tokens

Table 5: Model and optimization configurations: we basically followed the recommended hyper-parameters introduced in Luong et al. (2015) (`LSTM`) and Vaswani et al. (2017) (`Transformer`), respectively.

Detailed NMT configurations used in our experiments are shown in Table 5. For `Transformer` model, we used the "base" setting described in Vaswani et al. (2017).

B Actual Translation Examples

Input	`et puis il faut leur dire la vrit sur l' entrepreneuriat .`
Reference	`and then you have to tell them the truth about entrepreneurship .`
Baseline (`Transformer`)	`and then they have to tell the truth about entrepreneurship .`
Proposed (`Transformer+VAT w/ BT`)	`and then you have to tell them the truth about entrepreneurship .`
Input	`et ils m' ont laisse partir . c' tait un miracle .`
Reference	`and they let me go . it was a miracle .`
Baseline (`Transformer`)	`and they left me . it was a miracle .`
Proposed (`Transformer+VAT w/ BT`)	`and they let me go . it was a miracle .`
Input	`mais je lutte pour maintenir cette perspective dans ma vie quotidienne .`
Reference	`but I struggle to maintain this perspective in my daily life .`
Baseline (`Transformer`)	`but I 'm fighting to maintain that perspective in my daily life .`
Proposed (`Transformer+VAT w/ BT`)	`but I struggle to keep that perspective in my daily life .`

Table 6: Example translation from French→English (test2013).

Table 6 shows an example of an improved translation in French→English setting.