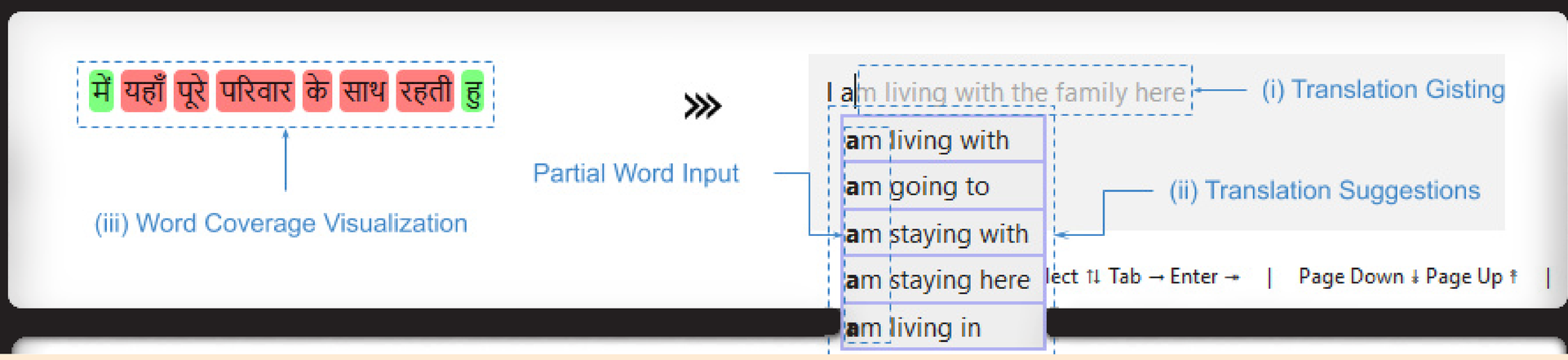
Interactive Neural Machine Translation

Sebastin Santy[1] **Sandipan Dandapat**[2] **Monojit Choudhury**[1] **Kalika Bali**[1]

[1]Microsoft Research, Bangalore, India [2]Microsoft R&D, Hyderabad, India

{t-sesan, sadandap, monojitc, kalikab}@microsoft.com



Aim

Assist translators by providing translation suggestions on the fly.

Advantages

- **Faster Turnaround**
  The gisting and suggestions help the translator breeze through translation tasks with minimal typing.
- **High Translation Quality**
  Language is inherently divergent and human translators cannot quickly enumerate all acceptable variants of a translation. On the other hand, machine translation has not yet reached human quality, though it can provide a number of variants. Combine the individual strengths, to produce high quality translations.
- **Amateur Translators**
  Expert translators are scarce. Take help of bilingual speakers who have native proficiency in these languages for translation tasks by providing suggestions and gisting.

Method

- **Seq2Seq decoder**: Conditional probability of generating output token y_t, at time step t, given the full input sequence $\mathbf{x}$ and the previously output tokens $y_1, ..., y_{t-1}$ is:

$$p(y_t|y_1, ..., y_{t-1}, \mathbf{x}) = g(y_{t-1}, s_t, c_t) \quad (1)$$

  - $g \rightarrow$ non-linearity function
  - $s_t \rightarrow$ hidden state
  - $c_t \rightarrow$ context vector: weighted average of all encoder hidden states with weights generated by the attention mechanism
- **INMT decoder**: Condition based on the partial input from the human translator $\{y'_1, ..., y'_{t-1}\}$ instead of default Seq2Seq output $\{y_1, ..., y_{t-1}\}$:

$$p(y_t|y'_1, ..., y'_{t-1}, \mathbf{x}) = g(y'_{t-1}, s_t, c_t) \quad (2)$$

- **Sparsemax Attention** is used to aid one-to-one source-target word mapping for word coverage visualization.
- **Beam Search** is used to produce multiple suggestions based on the partial input. It selects the most probable full translation for a given input sentence. If and when the translator diverges from this full translation, a new beam search is conducted from the partial input prefix till end of sentence is encountered.

Interface Overview

- **Translation Gisting**
  Prime the translator with a quick translation to reduce cognitive load. Spotting errors in the gisting is much easier, than trying to mentally structure the translations.
- **Translation Suggestions**
  Gist might not be the correct translation. Provide bi-gram suggestions which the translator can choose instead of the full gist.
- **Word Coverage Visualization**
  Show one-to-one source-target word mapping. This will help in understanding how much translation is completed.
- **Transliteration**
  Non-European languages have non-Latin script. Amateur translators usually use English keyboards to type. Provide character-wise transliteration, as each character triggers the engine to give new outputs.

Experiments

BLEU % - Measure the BLEU score of the generated gist after a certain fraction - x% of words of the intended translation has been provided. Table 1 shows the average BLEU score for each language pair at different values of x.

	Data Size	0%	10%	20%	40%
bn-en	1.1M	25.31	27.54	35.68	54.03
hi-en	1.5M	40.64	42.06	47.90	62.18
ml-en	897K	19.76	21.95	29.84	49.88
ta-en	428K	18.71	20.90	27.05	44.55
te-en	104K	11.92	14.57	21.17	41.98

Multi-BLEU Score with x% of partial input

Keystroke Reduction - Algorithmically compare minimum number of keystrokes required when typing interactively versus the same when manually typing. Reduction of around 30% keystrokes is observed for all the above mentioned languages.

System Overview

OpenNMT (PyTorch); JQuery (JS); Django;

Try it out! https://aka.ms/inmt