Enhancing Language Models in Statistical Machine Translation with Backward N-grams and Mutual Information Triggers

Deyi Xiong, Min Zhang, Haizhou Li
dyxiong@i2r.a-star.edu.sg

Background

Previous efforts on language models in statistical machine translation can be roughly categorized into two directions:

- **Large language models**:
  - built on billions or trillions of words in a distributed manner
  - yield better results but at the cost of huge storage and high computation

- **Syntax-based language models**:
  - trained on syntactic parse trees (constituent trees of dependency trees)
  - capture long-distance dependencies

Our Method

Learn additional information from training data to capture richer contexts and long-distance dependencies without resorting any linguistic resources such as a syntactic parser.

Backward LM

- **Model**
  \[ P(w_n^t) = \prod_{l=1}^{n} P(w_l | w_{l+1} \cdots w_n) \approx \prod_{l=1}^{n} P(w_l | w_{l+1} \cdots w_{n-1}) \]

- **Training**
  - invert the order in each sentence in the training data
  - then use any language tools to train the backward language model on the inverted training data

- **Decoding**
  - Three functions
    \[ P(x, y) = \prod_{i=1}^{n-1} P(x_i, y_{i+1}, \ldots, y_n) \]
    \[ C(x, y) = \begin{cases} 1 & \text{if } y \geq x, \\ 0 & \text{otherwise} \end{cases} \]
    \[ R(x, y) = \begin{cases} 1 & \text{if } y > x, \\ 0 & \text{otherwise} \end{cases} \]
  - CKY decoding with BTG rules

MI Trigger Model

- **Model**
  - Trigger pair: an ordered 2-tuple \((x, y)\)
  - \(x \rightarrow y\)

- **Pointwise mutual information**
  \[ PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \]

- **Trigger Model**
  \[ MI(w_i^t) = \prod_{i=1}^{n} \prod_{j=1}^{n} \exp(PMI(w_i, w_j)) \]

Experiments

- **Experimental setup**
  - Baseline: a BTG-based system (Xiong et al., 2006)
  - bilingual training data: nearly 100M words
  - 5-gram LMs (forward/backward) trained on Xinhua section of Gigaword corpus (306M words)
  - NIST MT-03/MT-04&05 as dev set and test sets respectively

- **Results**

Conclusions

- The BLM and MI trigger model collectively achieve up to 1 BLEU point on Chinese-English translation
- Both BLM and MI trigger model are able to capture useful information to improve translation quality