Decoding with a Document-level Language Model

Approximate the objective function by sentence-level translation model, document-level language model, and sentence-level language model scores.

\[
\hat{y} = \arg \max_y \log p(y | x, c^{(y)}) = \arg \max_y \log p(c^{(y)} | x, y) p(y | x)
\]

\[
\text{Assuming } x \text{ and } y \text{ are semantically similar}
\]

\[
y = \arg \max_y \log [p(y | x) + \log p(y | c^{(y)}) - \log p(y | p)]
\]

Objective

C−Score \(y, x, c^{(y)}\) = \( \log P_{S-T,M}(y | x) + \log P_{D-L,M}(y | c^{(y)}) - \log P_{S-T,M}(y) \)

Source Sent. \(x\) \rightarrow Sentence-level TM \(\hat{y}\) \rightarrow Context \(c^{(y)}\) \rightarrow Document-level LM \(\hat{y}\) \rightarrow Output

Decoding Strategy

Ranker with C-Score (§ 2.2.1)

Generate n-best hypotheses by sentence-level decoding and select the one that maximizes C-Score

End-to-End Neural Network

Context (previous translation)

Comprehensible hypotheses

Sentence-level TM

Document-level LM

PMI

Decompose the sentence C-Score into token-wise C-Score and perform beam search

\[
C\text{-Score}(y, x, c^{(y)}) = \sum \log P_{S-T,M}(y_j | x, y_{<j}) + \log P_{D-L,M}(y_j | c^{(y)}_{<j}) - \log P_{S-T,M}(y_j | y_{<j})
\]

Context-aware Beam Search (§ 2.2.2)

A: Sentence-level Transformer fails to reflect context in translation (low target-side PMI).
B: Decoding with C-Score with T=1 suffers from over-correction.

Analysis: How models change source-target PMI correlation?

Each point stands for a pair of PMI: \(\text{PMI}(c^{(y_i)}, y_{<i})\), \(\text{PMI}(c^{(y_i)}, y_{<i})\) for the i-th sentence pair in a dev set where model is the reference \(y_i\) and outputs of S-TM, proposed beam search with \(T=4\), and proposed beam search with \(T=4\).

Conclusion


Experiments

Settings

<table>
<thead>
<tr>
<th>Data</th>
<th>OpenSubtitles2018 (English → Russian, parallel: 6M, monolingual: 30M)</th>
</tr>
</thead>
</table>

Overall translation performance measured by BLEU score

- Transformer w/ BT: Sentence-level TM
- Transformer w/o context: DocReranker
- Transformer w/ context: DocRepair
- Transformer w/ context, proposed beam search: Ours (Base)
- Transformer w/ context, proposed beam search: Ours (DocReranker)

Only ours (rerank) and BayesDocReranker achieved significant improvements over Transformer. BayesDocReranker performed almost as well without context.

Evaluation of the ability to capture context [Vozia+2019]

Models guess the correct translation out of several candidates based on the translation score

- Deixis (person deixis), lex.c (lexical cohesion), ell.infl (influence of Russian nouns caused by ellipsis), and ell.vp (verb ellipses in English text not allowed in Russian)

- Models: Deixis, lex.c, ell.infl, ell.vp

- Scores: C-Score with T=1 suffers from over-correction.

- Analysis: How models change source-target PMI correlation?

- Each point stands for a pair of PMI: \(\text{PMI}(c^{(y_i)}, y_{<i})\), \(\text{PMI}(c^{(y_i)}, y_{<i})\) for the i-th sentence pair in a dev set where model is the reference \(y_i\) and outputs of S-TM, proposed beam search with \(T=4\), and proposed beam search with \(T=4\).