Overview

decoding strategies covered so far
- greedy search
- sampling
- beam search
- ensemble decoding

today
- vocabulary selection
- better greedy decoding
- reranking (right-to-left and reconstruction)
- constrained decoding
- simultaneous translation

Refresher
beam search
- maintain list of $K$ hypotheses (beam)
- at each time step, expand each hypothesis $k$: $p(y^k_t | S, y^k_{<t})$
- select $K$ hypotheses with highest total probability:

$$\prod_k p(y^k_t | S, y^k_{<t})$$

Refresher

time complexity of beam search

$O(|V|kt)$

- $|V|$: network vocabulary size
- $k$: beam size
- $t$: number of time steps
Vocabulary Selection Strategies

goal: reduce |V|

[Jean et al., 2015]
at decoding time, select a subset of the target vocabulary for softmax and search:
- fixed set of most common target words
- top translations of each source word according to IBM model

[L’Hostis et al., 2016]
- empirical comparison of different vocabulary selection strategies
- using IBM model (word alignment) performs best

Better Greedy Decoding

goal: reduce k (improve quality of greedy decoding)

sequence-level knowledge distillation [Kim and Rush, 2016]
- train teacher network on original training data
- train student network to mimic teacher

sequence-level knowledge distillation [Kim and Rush, 2016]
- word-level KD: minimize cross-entropy to teacher distribution
- sequence-level: teacher translates training set (with beam search)
  - KD: use 1-best translation as new reference
  - interpolation: use translation that is closest to reference (BLEU) as new reference
Better Greedy Decoding

sequence-level knowledge distillation [Kim and Rush, 2016]

- experimental settings:
  - English → German WMT 2014 data
  - large teacher network (4 layers; hidden layer size 1000)
  - small student network (2 layers; hidden layer size 500)

<table>
<thead>
<tr>
<th>model</th>
<th>BLEU (K=1)</th>
<th>BLEU (K=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>teacher baseline (4x1000)</td>
<td>17.7</td>
<td>19.5</td>
</tr>
<tr>
<td>sequence-level interpolation</td>
<td>19.6</td>
<td>19.8</td>
</tr>
<tr>
<td>student baseline (2x500)</td>
<td>14.7</td>
<td>17.6</td>
</tr>
<tr>
<td>word-level KD</td>
<td>15.4</td>
<td>17.7</td>
</tr>
<tr>
<td>sequence-level KD</td>
<td>18.9</td>
<td>19.0</td>
</tr>
<tr>
<td>sequence-level interpolation</td>
<td>18.5</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Reranking

phrase-based SMT

- common in phrase-based SMT with linear framework
- compute expensive features only for $k$-best translations

neural MT

- if previous predictions are incorrect, predictions may be less reliable
  → rerank with model trained to decode right-to-left
  [Liu et al., 2016, Sennrich et al., 2016]
- without coverage model, we may delete or repeat parts of source text
  → rerank with reconstruction cost ($p(S|T)$)
  [Li and Jurafsky, 2016, Tu et al., 2016]

Reconstruction

Example 1 (under-translation)

Source
Dieser Zustand erhöht vier bis fünf Mal das Risiko, dass eine transitorische ischämische Attacke (TIA) oder Schlaganfall vorkommt.

Reference
This condition increases your risk by about four to five times of having a transient ischaemic attack (TIA) or stroke.

Translation
This condition increases the risk of transient ischaemic attack (TIA) or stroke.

Reranking Example 1

<table>
<thead>
<tr>
<th>lcost</th>
<th>rcost</th>
<th>Translation</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.85</td>
<td>2.20</td>
<td>this condition increases the risk of transient ischaemic attack (TIA) or stroke.</td>
<td>7</td>
</tr>
<tr>
<td>4.93</td>
<td>2.02</td>
<td>this condition increases the risk of transient ischaemic attacks (TIA) or stroke.</td>
<td>6</td>
</tr>
<tr>
<td>5.36</td>
<td>2.28</td>
<td>this condition increases the risk of a transient ischaemic attack (TIA) or stroke.</td>
<td>9</td>
</tr>
<tr>
<td>6.67</td>
<td>0.44</td>
<td>this condition increases four to five times the risk that transient ischaemic attack (TIA) or stroke.</td>
<td>1</td>
</tr>
<tr>
<td>5.13</td>
<td>2.22</td>
<td>this condition increases the risk of transient ischaemic attack (TIA) or stroke.</td>
<td>10</td>
</tr>
<tr>
<td>6.95</td>
<td>0.44</td>
<td>this situation increases four to five times the risk that transient ischaemic attack (TIA) or stroke.</td>
<td>2</td>
</tr>
</tbody>
</table>
Constrained Decoding

why?
- force translation of terminology
- interactive machine translation

Prefix-Constrained Decoding

- cumbersome in phrase-based MT
- very natural in neural MT
- standard decoding:
  \[
  p(T|S) = \prod_{i=1}^{n} p(y_i|y_{i-1}, x_1, \ldots, x_m)
  \]

- prefix-constrained decoding:
  \[
  \text{PRE} = y_1, \ldots, y_j
  \]
  \[
  p(T|S, \text{PRE}) = \prod_{i=j+1}^{n} p(y_i|y_{i-1}, x_1, \ldots, x_m)
  \]
  - simple change to decoding algorithm; no changes to model/training

Constrained Decoding

arbitrary constraints
- how can we decode with more general constraints?
- keep track of how many constraints hypothesis fulfills
- finished hypothesis is only valid if all constraints are fulfilled
- challenge: hypotheses that fulfill constraints must survive pruning

Grid Beam Search [Hokamp and Liu, 2017]

- core idea: eliminate competition between hypotheses that fulfill different number of constraints
- 2d grid (each box is one beam):
  - x axis: number of time steps
  - y axis: number of constraint tokens matched
Simultaneous Translation

objectives in simultaneous translation:

- maximize translation quality
- minimize latency

to minimize latency, system may start translating before full input has been seen

---

Don’t Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation

Grissom II et al., 2014

- actions:
  - commit partial translation
  - wait for more words
  - predict the next or final source word
- goal: learn a policy that maximizes latency-bleu:

\[ Q(x, y) = \frac{1}{T} \sum_t \text{BLEU}(y_t, r) + T \cdot \text{BLEU}(y_T, r) \]

---

Figure 2: A simultaneous translation from source (German) to target (English). The agent chooses to wait until after (3). At this point, it is sufficiently confident to predict the final verb of the sentence (4). Given this additional information, it can now begin translating the sentence into English, constraining future translations (5). As the rest of the sentence is revealed, the system can translate the remainder of the sentence.
[Gu et al., 2017]:
- unidirectional encoder
- simple action space: read or write

Simultaneous Neural Machine Translation

Bibliography I


Bibliography II


