THE UNIVERSITY of EDINBURGH

Machine Translation

14: Advanced Decoding Techniques

Rico Sennrich

University of Edinburgh

Overview

decoding strategies covered so far

- greedy search
- sampling
- beam search
- ensemble decoding

today

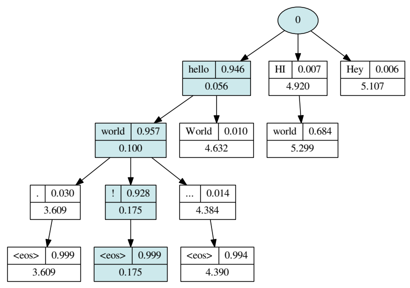
- vocabulary selection
- better greedy decoding
- reranking (right-to-left and reconstruction)
- constrained decoding
- simultaneous translation

Refresher

beam search

- maintain list of K hypotheses (beam)
- at each time step, expand each hypothesis k: $p(y_i^k|S, y_{<i}^k)$
- select K hypotheses with highest total probability:

$$\prod_i p(y_i^k|S, y_{<i}^k)$$



$K = 3$

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

Vocabulary Selection Strategies

[L'Hostis et al., 2016]

- empirical comparison of different vocabulary selection strategies
- using IBM model (word alignment) performs best

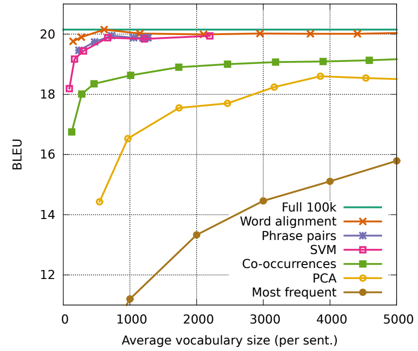


Figure 2: BLEU vs. vocabulary size for different selection strategies.

Better Greedy Decoding

goal: reduce k (improve quality of greedy decoding)

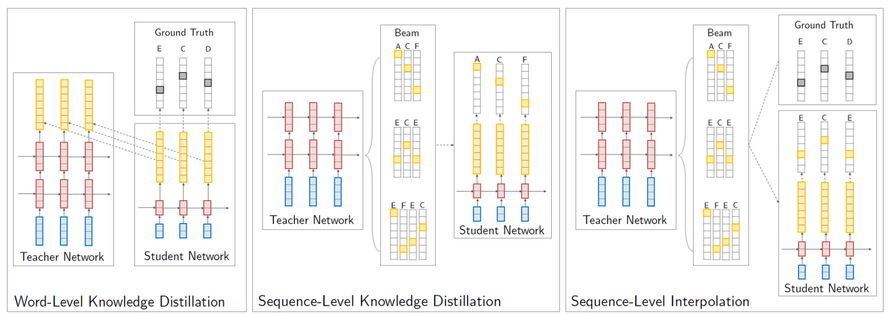
sequence-level knowledge distillation [Kim and Rush, 2016]

- train teacher network on original training data
- train student network to mimic teacher

Better Greedy Decoding

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)

model	BLEU (K=1)	BLEU (K=5)
teacher baseline (4x1000)	17.7	19.5
sequence-level interpolation	**19.6**	**19.8**
student baseline (2x500)	14.7	17.6
word-level KD	15.4	17.7
sequence-level KD	**18.9**	**19.0**
sequence-level interpolation	18.5	18.7

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.

slide credit: Phil Williams

Reconstruction

Reranking Example 1

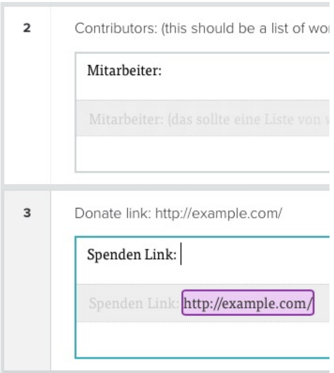
lcost	rcost	Translation	Rank'
4.85	2.20	this condition increases the risk of transient ischaemic attack (TIA) or stroke .	7
4.93	2.02	this condition increases the risk of transient ischaemic attacks (TIA) or stroke .	6
5.36	2.28	this condition increases the risk of a transient ischaemic attack (TIA) or stroke .	9
6.67	0.44	this condition increases **four to five times** the risk that transient ischaemic attack (TIA) or	1
5.13	2.22	this condition increases the risk of transient ischemic attack (TIA) or stroke .	10
6.95	0.44	this situation increases **four to five times** the risk that transient ischaemic attack (TIA) or stroke	2

slide credit: Phil Williams

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:

$$p(T|S) = \prod_{i=1}^{n} p(y_i|y_1, \dots, y_{i-1}, x_1, \dots, x_m)$$

- prefix-constrained decoding:

$$\text{PRE} = y_1, \dots, y_j$$

$$p(T|S, \text{PRE}) = \prod_{i=j+1}^{n} p(y_i|y_1, \dots, y_{i-1}, x_1, \dots, x_m)$$

- simple change to decoding algorithm; no changes to model/training

Constrained Decoding

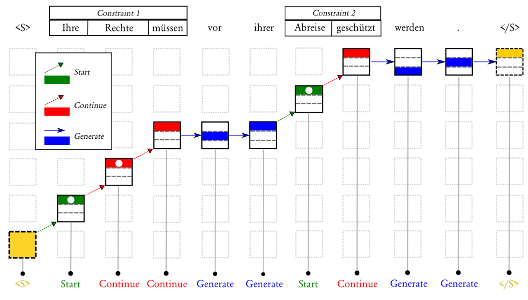
arbitrary constraints

- 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

Constrained Decoding

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



Input: Rights protection should begin before their departure .

Constrained Decoding

Grid Beam Search [Hokamp and Liu, 2017]

- very general:
  - agnostic to model architecture
  - requires no source-side information
  - requires no retraining
- constraints must be in-vocabulary: use subword-level model
- problem: high computational complexity: $O(|V|ktc)$
  (k: beam size; t: length; c: # constraint tokens)

Simultaneous Translation

objectives in simultaneous translation:

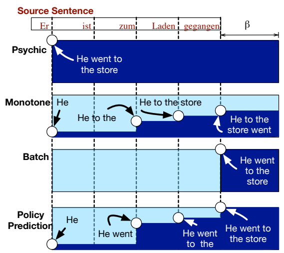
1. maximize translation quality
2. minimize latency

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**:

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



Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation [Grissom II et al., 2014]

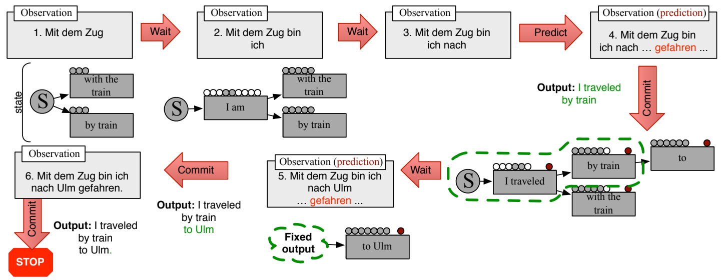
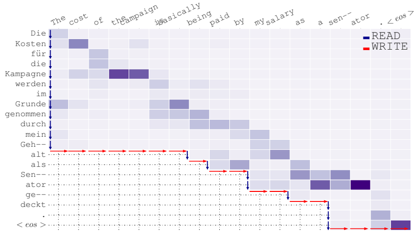
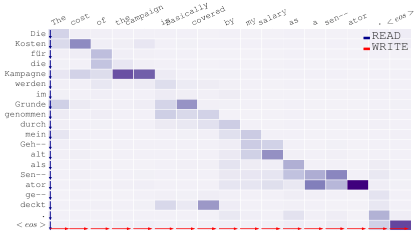


Figure 2: A simultaneous translation from source (German) to target (English). The agent chooses to wait until after (3). At this point, it is sufficiently confident to predict the final verb of the sentence (4). Given this additional information, it can now begin translating the sentence into English, constraining future translations (5). As the rest of the sentence is revealed, the system can translate the remainder of the sentence.

Simultaneous Neural Machine Translation



(a) Simultaneous Neural Machine Translation



(b) Neural Machine Translation

[Gu et al., 2017]:

- unidirectional encoder
- simple action space: **read** or **write**

Bibliography I

Grissom II, A., He, H., Boyd-Graber, J., Morgan, J., and Daumé III, H. (2014).
Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation.
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Gu, J., Neubig, G., Cho, K., and Li, V. O. (2017).
Learning to Translate in Real-time with Neural Machine Translation.
In
Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Pape
pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

Hokamp, C. and Liu, Q. (2017).
Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search.
In
Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - A
pages 1535–1546.

Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2015).
On Using Very Large Target Vocabulary for Neural Machine Translation.
In
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference o
pages 1–10, Beijing, China. Association for Computational Linguistics.

Kim, Y. and Rush, A. M. (2016).
Sequence-Level Knowledge Distillation.
In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Bibliography II

L'Hostis, G., Grangier, D., and Auli, M. (2016).
Vocabulary Selection Strategies for Neural Machine Translation.
ArXiv e-prints.

Li, J. and Jurafsky, D. (2016).
Mutual Information and Diverse Decoding Improve Neural Machine Translation.
CoRR, abs/1601.00372.

Liu, L., Utiyama, M., Finch, A., and Sumita, E. (2016).
Agreement on Target-bidirectional Neural Machine Translation .
In NAACL HLT 16, San Diego, CA.

Sennrich, R., Haddow, B., and Birch, A. (2016).
Edinburgh Neural Machine Translation Systems for WMT 16.
In Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 368–373, Berlin, Germany.

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2016).
Neural Machine Translation with Reconstruction.
CoRR, abs/1611.01874.