Investigations on Phrase-based Decoding with Recurrent Neural Network Language and Translation Models

Tamer Alkhouli, Felix Rietig, and Hermann Ney
surname@cs.rwth-aachen.de

Tenth Workshop on Statistical Machine Translation
Lisbon, Portugal
18.09.2015

Human Language Technology and Pattern Recognition
Chair of Computer Science 6
Computer Science Department
RWTH Aachen University, Germany

Motivation

- Neural networks (NNs) were applied successfully in machine translation
- NN translation models applied in n-best rescoring:
  - Feedforward NNs (FFNNs): [Le & Allauzen$^+$ 12]
  - Recurrent NNs (RNNs): [Hu & Auli$^+$ 14, Sundermeyer & Alkhouli$^+$ 14]
- NN translation models (TMs) in phrase-based decoding:
  - FFNNs: [Devlin & Zbib$^+$ 14]
  - Recurrent neural networks (RNNs): this work
- Neural machine translation
  - [Sutskever & Vinyals$^+$ 14, Bahdanau & Cho$^+$ 15]

Motivation

Directly related:

- RNN language models (LMs) in phrase-based decoding: [Auli & Gao 14]
- Caching for RNN LM in speech recognition: [Huang & Zweig+ 14]
- Word-based RNN TMs: [Sundermeyer & Alkhouli+ 14]

This work:

- Integration of RNN LMs and TMs into phrase-based decoding
- Caching to allow a flexible choice between translation quality and speed
- Phrase-based decoding vs. rescoring with RNN LMs and TMs

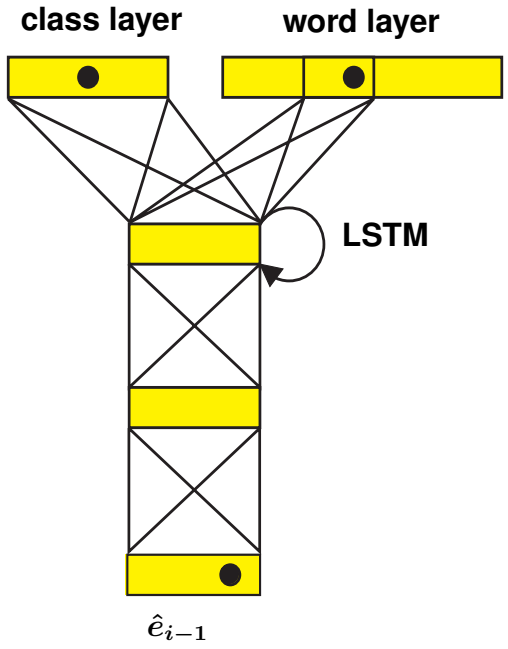
Recurrent Neural Network Language Models

- RNN LM computes the probability of the target sequence $e_1^I = e_1 \ldots e_i \ldots e_I$

$$p(e_1^I) = \prod_{i=1}^{I} p(e_i|e_1^{i-1})$$

  - unbounded context encoded in the RNN state

- Evaluation of $p(e_i|e_1^{i-1})$
  1. word embedding lookup
  2. advance hidden state
  3. compute full raw output layer
  4. normalize output layer



Phrase-based Decoding

Phrase-based decoding

- Graph of search states representing partial hypotheses
- Search state stores n-gram LM history
- States are expanded and pruned (beam search)
- States storing the same information are merged (state recombination)
  - higher LM orders lead to fewer recombinations

RNN LM Integration

RNN LM integration into phrase-based decoding

- Naïve integration: store the complete LM history in the search state
  - state recombination is reduced radically
  - less variety within the beam

- Alternative proposed by [Auli & Galley+ 13]
  - store the RNN state in the search state
  - ignore the RNN state during recombination
  - approximate RNN evaluation

RNN LM Integration

- **This work (similar to [Huang & Zweig$^+$ 14])**
  - **store the RNN state in a global cache**
  - **caching order: m recent words as caching key**
  - **store the truncated history in the search state**
  - **ignore the added information during recombination**

- **Why bother?**
  - **cache avoids redundant computations across search states**
  - **control the trade-off between accuracy and speed**

RNN LM Integration

- Large caching order of $m = 30$ used
- All entries share the same translation quality
- Class-factored output layer (2000 classes)
- RNN LM for IWSLT 2013 German$\rightarrow$English
- 1 hidden layer with 200 Long short term memory (LSTM) nodes

Cache	Speed [words/second]
none	0.03
RNN state	0.05
RNN state + norm. factor	0.19
RNN state + norm. factor + word prob.	0.19

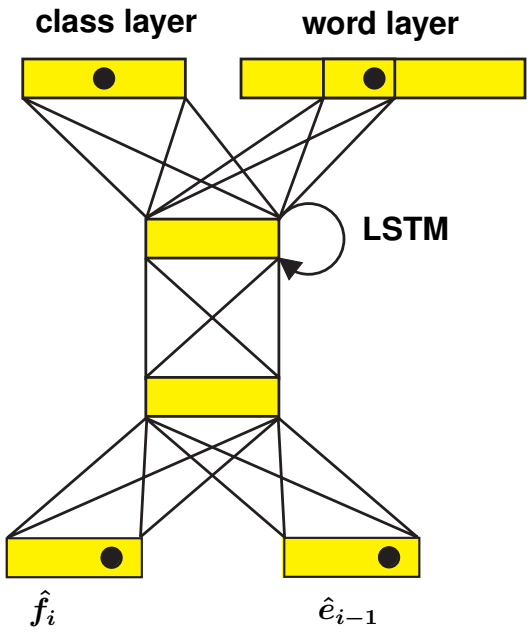
RNN TM Integration

- Source sentence $f_1^I = f_1 \ldots f_i \ldots f_I$
- Target sentence $e_1^I = e_1 \ldots e_i \ldots e_I$
- One-to-one alignment using IBM 1 models [Sundermeyer & Alkhouli+ 14]

- RNN Joint Model (JM)

$$p(e_1^I|f_1^I) \approx \prod_{i=1}^{I} p(e_i|e_1^{i-1}, \boldsymbol{f_1^i})$$

  - $\boldsymbol{f_i}$ in addition to e_{i-1} as input
- Same caching strategies as RNN LM

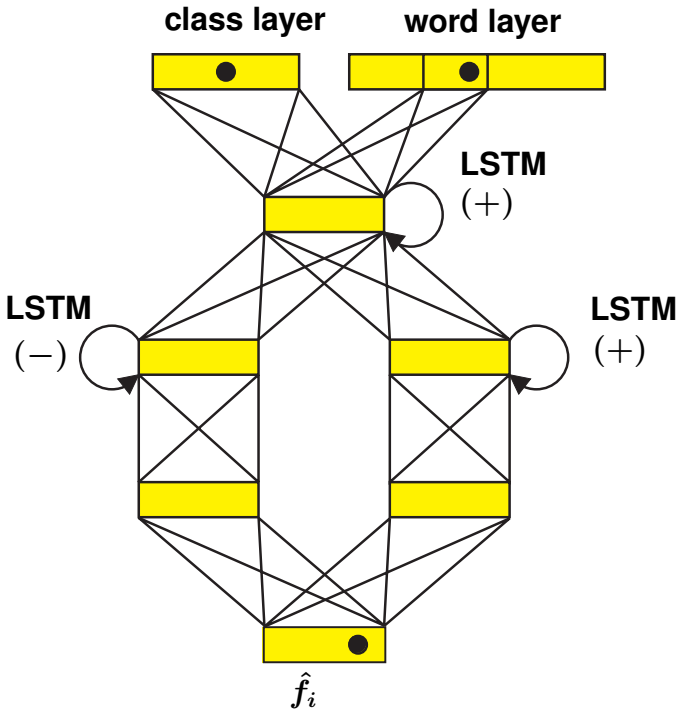


RNN TM Integration

- **Bidirectional translation model (BTM)**

$$p(e_1^I|f_1^I) \approx \prod_{i=1}^{I} p(e_i|f_1^I)$$

  - **split sentence at position i**
  - **RNN states for past and future source context**



- **Exact evaluation during decoding**

Experimental Setups

	IWSLT		BOLT	
	German	English	Arabic	English
Sentences	4.32M		921K	
Run. Words	108M	109M	14M	16M
Vocabulary	836K	792K	285K	203K
NN Sentences	138K		921K	
NN Vocabulary	41K	30K	139K	87K

Experimental Setups

Baseline: standard phrase-based decoder *Jane* [Wuebker & Huck$^+$ 12]

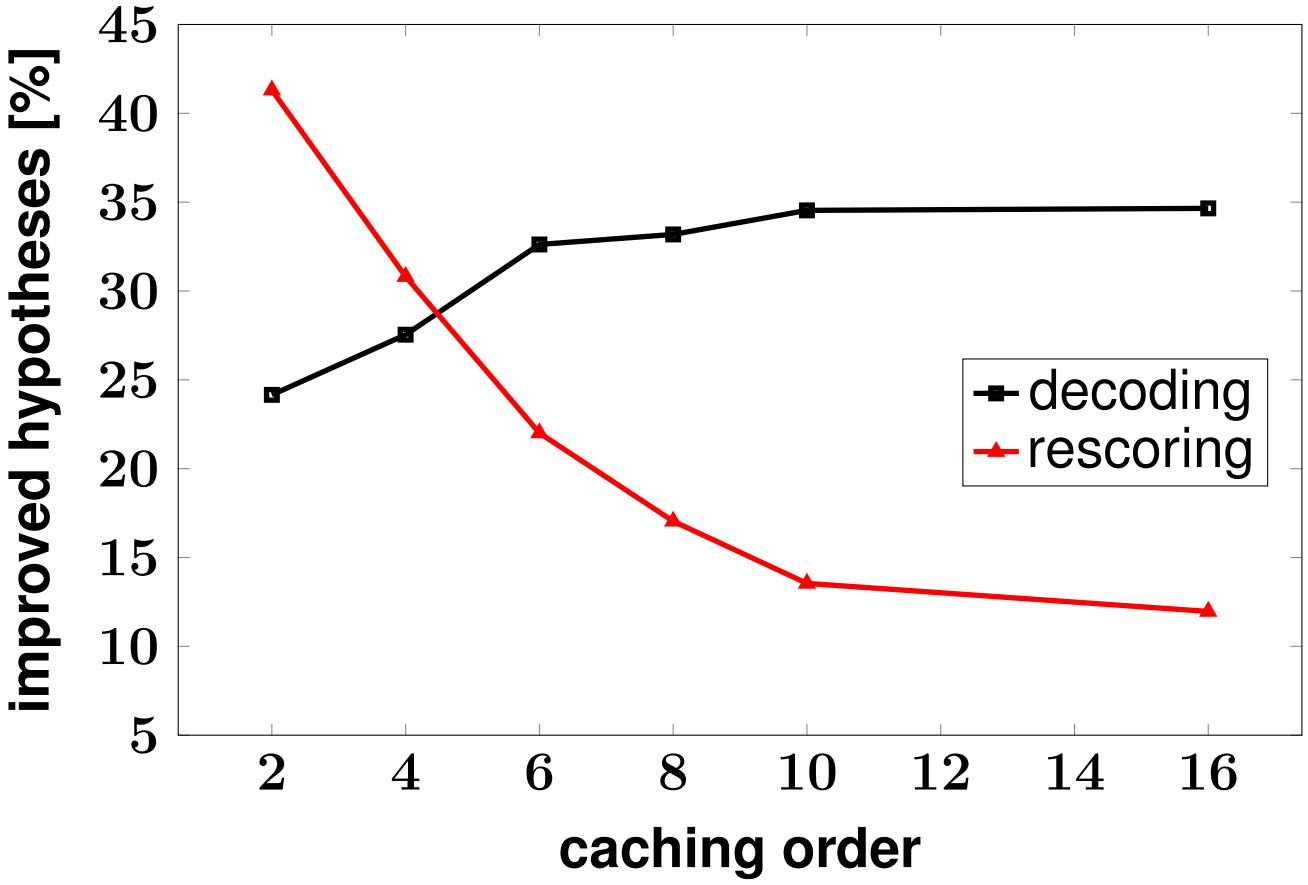
- **Hierarchical reordering model [Galley & Manning 08]**
- **IWSLT: 7-gram word class LM [Wuebker & Peitz$^+$ 13]**

NN setups

- **BTM: 1 projection and 3 LSTM layers**
- **JM and LM: 1 projection and 1 LSTM layers**
- **Class-factored output layer with 2000 classes**

Search Quality: Decoding vs. Rescoring

- RNN LM
- Rescoring hypotheses are fixed



Caching Order vs. Translation Quality

- RNN LM
- IWSLT German→English

Caching Order	BLEU [%] dev	test
2	33.1	30.8
4	33.4	31.2
6	33.9	31.6
8	33.9	31.5
16	34.0	31.5
30	33.9	31.5
-	33.9	31.5

Results: IWSLT 2013 German→English

- **LM caching order: 8, JM caching order: 5**

	test	
	BLEU [%]	TER [%]
baseline	30.6	49.2
LM Rescoring	31.5	48.6
LM Decoding	31.6	48.3
BTM Rescoring	32.2	47.8
BTM Decoding	32.3	47.3
JM Rescoring	31.6	48.3
JM Decoding	31.6	48.2

Results: IWSLT 2013 German→English

- **LM caching order: 8, JM caching order: 5**

	test	
	BLEU [%]	TER [%]
baseline	30.6	49.2
LM Rescoring	31.5	48.6
LM Decoding	31.6	48.3
+ LM Rescoring	31.9	48.4
BTM Rescoring	32.2	47.8
BTM Decoding	32.3	47.3
JM Rescoring	31.6	48.3
JM Decoding	31.6	48.2
+ JM Rescoring	31.8	47.9

Results: BOLT Arabic→English

- **LM caching order:** 8**, JM caching order:** 10
- `test1`**:** 1510 **segments**

	test1	
	BLEU [%]	TER [%]
baseline	23.9	59.7
LM Rescoring	24.3	59.3
LM Decoding	24.6	59.0
BTM Rescoring	24.7	58.9
BTM Decoding	24.8	58.9
JM Rescoring	24.4	59.0
JM Decoding	24.5	59.0

Results: BOLT Arabic→English

- **LM caching order:** 8**, JM caching order:** 10
- `test1`**:** 1510 **segments**

	test1	
	BLEU [%]	TER [%]
baseline	23.9	59.7
LM Rescoring	24.3	59.3
LM Decoding	24.6	59.0
+ LM Rescoring	25.0	58.8
BTM Rescoring	24.7	58.9
BTM Decoding	24.8	58.9
JM Rescoring	24.4	59.0
JM Decoding	24.5	59.0
+ JM Rescoring	24.5	59.0

Conclusion

- Approximate and exact RNNs in phrase-based decoding
- Caching speeds up translation
- RNNs in decoding perform at least as good as in n-best rescoring
- Recombination error leads to approximate RNN scores

Future work:

- Make recombination dependent on RNN state
- Standalone decoding with alignment-based RNNs

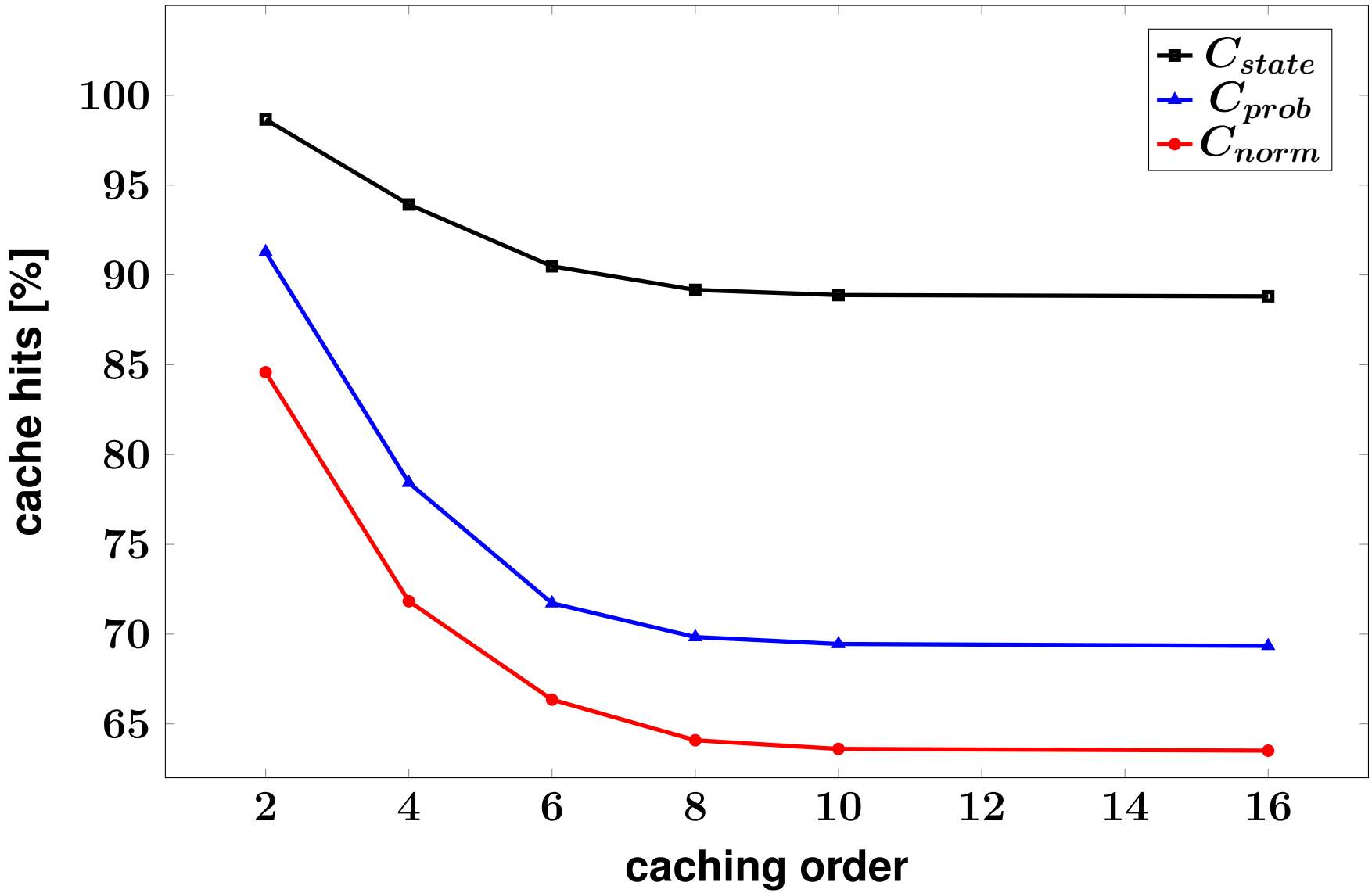
Thank you for your attention

Tamer Alkhouli

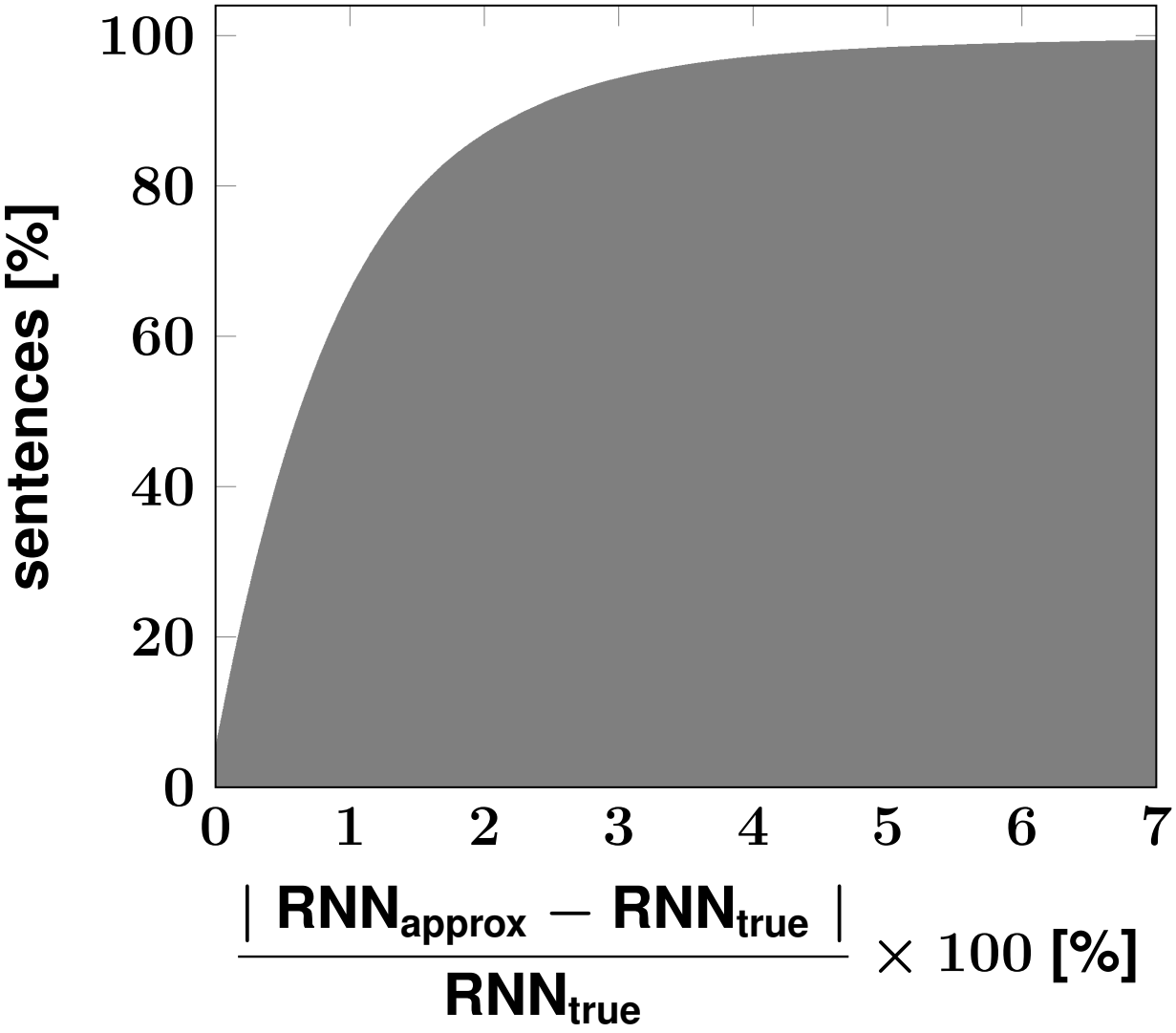
surname@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

Appendix: Caching Strategies



Appendix: RNN Relative Error



References

[Auli & Galley+ 13] M. Auli, M. Galley, C. Quirk, G. Zweig: Joint Language and Translation Modeling with Recurrent Neural Networks. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1044–1054, Seattle, USA, Oct. 2013. 6

[Auli & Gao 14] M. Auli, J. Gao: Decoder Integration and Expected BLEU Training for Recurrent Neural Network Language Models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 136–142, Baltimore, MD, USA, June 2014. 3

[Bahdanau & Cho+ 15] D. Bahdanau, K. Cho, Y. Bengio: Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*, San Diego, Calefornia, USA, May 2015. 2

[Devlin & Zbib+ 14] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, J. Makhoul: Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1370–1380, Baltimore, MD, USA, June 2014. 2

[Galley & Manning 08] M. Galley, C.D. Manning: A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 848–856, Honolulu, Hawaii, USA, October 2008. 12

[Hu & Auli+ 14] Y. Hu, M. Auli, Q. Gao, J. Gao: Minimum Translation Modeling with Recurrent Neural Networks. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 20–29, Gothenburg, Sweden, April 2014. 2

[Huang & Zweig+ 14] Z. Huang, G. Zweig, B. Dumoulin: Cache Based Recurrent Neural Network Language Model Inference for First Pass Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6404–6408, Florence, Italy, May 2014. 3, 7

[Le & Allauzen+ 12] H.S. Le, A. Allauzen, F. Yvon: Continuous Space Translation Models with Neural Networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 39–48, Montreal, Canada, June 2012. 2

[Sundermeyer & Alkhouli+ 14] M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney: Translation Modeling with Bidirectional Recurrent Neural Networks. In *Conference on Empirical Methods on Natural Language Processing*, pp. 14–25, Doha, Qatar, Oct. 2014. 2, 3, 9

[Sutskever & Vinyals+ 14] I. Sutskever, O. Vinyals, Q.V.V. Le: Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, pp. 3104–3112, Montréal, Canada, December 2014. 2

[Vilar & Stein+ 10] D. Vilar, D. Stein, M. Huck, H. Ney: Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pp. 262–270, Uppsala, Sweden, July 2010.

[Wuebker & Huck+ 12] J. Wuebker, M. Huck, S. Peitz, M. Nuhn, M. Freitag, J.T. Peter, S. Mansour, H. Ney: Jane 2: Open Source Phrase-based and Hierarchical Statistical Machine Translation. In *International Conference on Computational Linguistics*, pp. 483–491, Mumbai, India, Dec. 2012. 12

[Wuebker & Peitz+ 13] J. Wuebker, S. Peitz, F. Rietig, H. Ney: Improving Statistical Machine Translation with Word Class Models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1377–1381, Seattle, USA, Oct. 2013. 12