Extended Translation Models in Phrase-based Decoding

Andreas Guta, Joern Wuebker, Miguel Graça,
Yunsu Kim and Hermann Ney
surname@cs.rwth-aachen.de

Tenth Workshop on Statistical Machine Translation (WMT)
Lisbon, Portugal
18.09.2015

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

Introduction

Phrase-based translation models
[Och & Tillmann^{+} 99, Zens & Och^{+} 02, Koehn & Och^{+} 03]

- **phrases extracted from alignments obtained using GIZA++ [Och & Ney 03]**
- **estimation as relative frequencies of phrase pairs**
- **drawbacks:**
  - **single-word phrases translated without any context**
  - **uncaptured dependencies beyond phrase boundaries**
  - **difficulties with long-range reorderings**

Related Work

- bilingual language models [Niehues & Herrmann^{+} 11]
  - atomic source phrases, no reordering context
- reordering model based on sequence labeling [Feng & Peter^{+} 13]
  - modeling only reorderings
- operation sequence model (OSM) [Durrani & Fraser^{+} 13]
  - n-gram model based on minimal translation units
- neural network models for extended translation context
  - rescoring [Le & Allauzen^{+} 12, Sundermeyer & Alkhouli^{+} 14]
  - decoding [Devlin & Zbib^{+} 14, Auli & Gao 14, Alkhouli & Rietig^{+} 15]
  - stand-alone models [Sutskever & Vinyals^{+} 14, Bahdanau & Cho^{+} 15]
- joint translation and reordering models [Guta & Alkhouli^{+} 15]
  - word-based and simpler reordering approach than OSM
  - count models and neural networks (NNs)

This Work

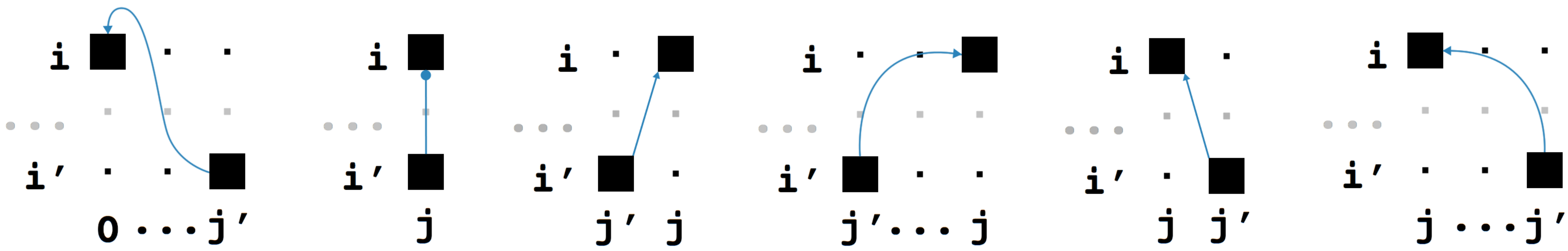
- develop two variants of extended translation models (ETM)
  - extend IBM models by a bilingual word pair and a reordering operation
  - integrated into log-linear framework of phrase-based decoding
  - explicit treatment of multiple alignments and unaligned words
- benefits:
  - lexical and reordering context for single-word phrases
  - dependencies across phrase boundaries
  - long-range source dependencies
- first step: implementation as smoothed count models
- the long-term goal:
  - application as stand-alone models in decoding
  - retraining the word alignments

Extended Translation Models

- **source sentence** $f_1^J = f_1 \dots f_j \dots f_J$
- **target sentence** $e_1^I = e_1 \dots e_i \dots e_I$
- **inverted alignment** b_1^I **with** $b_i \subseteq \{1 \dots J\}$
  - **unaligned source positions** b_0
- **empty words** f_0, e_0

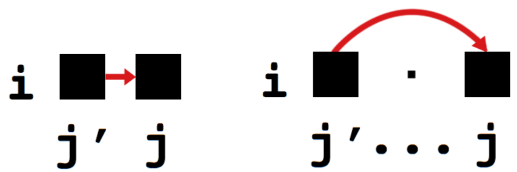
Jump Classes

- **generalizing alignments to**
  - **jump classes for source positions aligned to subsequent target positions**



insert (↓) stay (•) forward (→) jump forward (↷) backward (←) jump backward (↶)

  - **jump classes source positions aligned to the same target position**



forward (→) jump forward (↷)

Extended Inverse Translation Model (EiTM)

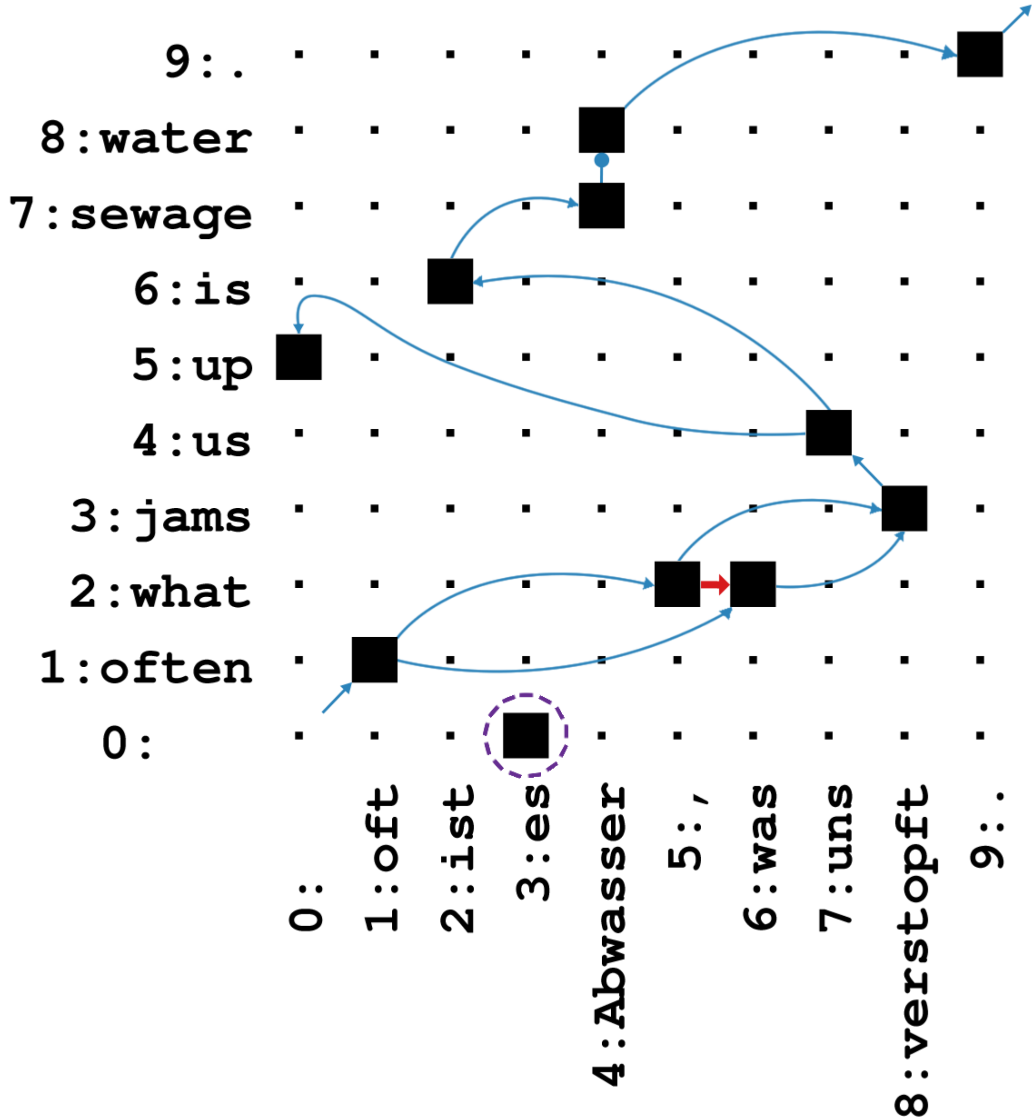
- **EiTM models the inverse probability $p(f_1^J|e_1^I)$**

$$p(f_1^J|e_1^I) = \max_{b_1^I} \left\{ \prod_{i=1}^{I} \left(\underbrace{p(f_{b_i}|e_{i'}, e_i, f_{b_{i'}}, b_{i'}, b_i)}_{\text{lexicon model}} \cdot \underbrace{p(b_i|e_{i'}, e_i, f_{b_{i'}}, b_{i'})}_{\text{alignment model}} \right) \cdot \underbrace{p(f_{b_0}|e_0)}_{\text{deletion model}} \right\}$$

- **current source words f_{b_i} and target word e_i**
- **previous source words $f_{b_{i'}}$ and target word $e_{i'}$**

- **generalize aligments $b_{i'}, b_i$ to jump classes**
- **multiple source predecessors j' in $b_{i'}$ or b_i**
  - **average probabilities over all j'**

EiTM Example



Extended Direct Translation Model (EdTM)

- further aim: model $p(e_1^I|f_1^J)$ as well
- first approach by using the EiTM:
  - swap source and target corpora
  - invert also the alignment
- drawback:
  - source words not translated in monotone order
  - source word preceding a phrase might have not been translated yet
  - its last aligned predecessor and corresponding aligned target words generally unknown
- dependencies beyond phrase boundaries cannot be captured
- develop the EdTM
  - swap source and target corpora, but keep b_1^I
  - incorporate dependencies beyond phrase boundaries

Extended Direct Translation Model (EdTM)

- EdTM models the direct probability $p(e_1^I|f_1^J)$

$$p(e_1^I|f_1^J) = \max_{b_1^I} \left\{ \prod_{i=1}^{I} \left(\underbrace{p(e_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}, b_i)}_{\text{lexicon model}} \cdot \underbrace{p(b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'})}_{\text{alignment model}} \right) \cdot \underbrace{p(e_0|f_{b_0})}_{\text{deletion model}} \right\}$$

- differences to EiTM
  - lexicon model: swapped e_i and f_{b_i}
  - alignment model: dependence on f_{b_i} (instead of e_i)
  - deletion model: swapped e_0 and f_{b_0}

Count Models and Smoothing

How to train the derived EdTM and EiTM models?

- **estimate Viterbi alignment using GIZA++ [Och & Ney 03]**
- **compute relative frequencies**
- **apply interpolated Kneser-Ney smoothing [Chen & Goodman 98]**

Integration into Phrase-based Decoding

- phrase-based decoder Jane 2 [Wuebker & Huck^{+} 12]
- log-linear model combination [Och & Ney 04]
  - tuning with minimum error rate training (MERT) [Och 03]

- annotation of phrase-table entries with word alignments
- extended translation models integrated as up to 4 additional features:
  - EdTM and EiTM
  - Source→Target and Target→Source
- search state extension:
  - store the source position aligned to the last translated target word

- context beyond phrase boundaries only in Source→Target direction

Experimental Setups

	IWSLT German	IWSLT English	IWSLT English	IWSLT French	BOLT Chinese	BOLT English	BOLT Arabic	BOLT English
Sentences								
full data	4.32M		26.05M		4.08M		0.92M	
indomain	138K		185K		67.8K		0.92M	
Run. Words	108M	109M	698M	810M	78M	86M	14M	16M
Vocabulary	836K	792K	2119K	2139K	384K	817K	285K	203K

- phrase-based systems
  - phrasal and lexical models (both directions)
  - word and phrase penalties
  - distortion model
  - 4- / 5-gram language model (LM)
  - 7-gram word class LM [Wuebker & Peitz^{+} 13]
  - hierarchical reordering model (HRM) [Galley & Manning 08]

Results: IWSLT 2014 German→English

	test2010	
	BLEU [%]	TER [%]
phrase-based system + HRM	**30.7**	**49.3**
+ EiTM (Source↔Target)	**31.4**	**48.3**
+ EdTM (Source↔Target)	**31.6**	**48.1**
+ EiTM (Source→Target) + EdTM (Source→Target)	**31.6**	**48.2**
+ EiTM (Source↔Target) + EdTM (Source↔Target)	**31.8**	**48.2**

Results: Comparison to OSM

- all results measured in BLEU [%]

	IWSLT		BOLT	
	De→En	En→Fr	Zh→En	Ar→En
phrase-based system + HRM	30.7	33.1	17.0	24.0
+ ETM	31.8	33.9	17.5	24.4
+ 7-gram OSM	31.8	34.5	17.6	24.1

Conclusion

- integration of extended translation models into phrase-based decoding
  - lexical and reordering context beyond phrase boundaries
  - multiple and empty alignments
  - relative frequencies with interpolated Kneser-Ney smoothing
- improving phrase-based systems including HRM
  - by up to 1.1% BLEU and TER
  - by 0.7% BLEU on average for four large-scale tasks
- competitive to a 7-gram OSM
  - 0.1% BLEU less improvement on average on top of phrase-based systems including the HRM
- long-term goals:
  - retraining the alignments: joint optimization
  - stand-alone decoding without phrases

Thank you for your attention

Andreas Guta

`surname@cs.rwth-aachen.de`

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

References

[Alkhouli & Rietig$^+$ 15] T. Alkhouli, F. Rietig, H. Ney: Investigations on Phrase-based Decoding with Recurrent Neural Network Language and Translation Models. In *Proceedings of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, pp. 294–303, Lisbon, Portugal, Sept. 2015. 3

[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. 3

[Chen & Goodman 98] S.F. Chen, J. Goodman: An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, 63 pages, Aug. 1998. 11

[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. 3

[Durrani & Fraser$^+$ 13] N. Durrani, A. Fraser, H. Schmid, H. Hoang, P. Koehn: Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 399–405, Sofia, Bulgaria, August 2013. 3

[Feng & Peter$^+$ 13] M. Feng, J.T. Peter, H. Ney: Advancements in Reordering Models for Statistical Machine Translation. In *Annual Meeting of the Assoc. for Computational Linguistics*, pp. 322–332, Sofia, Bulgaria, Aug. 2013. 3

[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*, EMNLP '08, pp. 848–856, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. 13

[Guta & Alkhouli$^+$ 15] A. Guta, T. Alkhouli, J.T. Peter, J. Wuebker, H. Ney: A Comparison between Count and Neural Network Models Based on Joint Translation and Reordering Sequences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1401–1411, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. 3

[Koehn & Och+ 03] P. Koehn, F.J. Och, D. Marcu: Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pp. 127–133, Edmonton, Alberta, 2003. 2

[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. 3

[Niehues & Herrmann+ 11] J. Niehues, T. Herrmann, S. Vogel, A. Waibel: *Proceedings of the Sixth Workshop on Statistical Machine Translation*, chapter Wider Context by Using Bilingual Language Models in Machine Translation, pp. 198–206. 2011. 3

[Och 03] F.J. Och: Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 160–167, Sapporo, Japan, July 2003. 12

[Och & Ney 03] F.J. Och, H. Ney: A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, Vol. 29, No. 1, pp. 19–51, March 2003. 2, 11

[Och & Ney 04] F.J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417–449, Dec. 2004. 12

[Och & Tillmann+ 99] F.J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 20–28, University of Maryland, College Park, MD, June 1999. 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. 3

[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, 2014. 3

[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. 13

[Zens & Och+ 02] R. Zens, F.J. Och, H. Ney: Phrase-Based Statistical Machine Translation. In *25th German Conf. on Artificial Intelligence (KI2002)*, pp. 18–32, Aachen, Germany, Sept. 2002. 2