QCRI's Machine Translation Systems for IWSLT'16

Nadir Durrani Fahim Dalvi Hassan Sajjad Stephan Vogel
Arabic Language Technologies
Qatar Computing Research Institute, HBKU

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Motivation

- Can NMT beat current state-of-the-art?
  - for Arabic-English language pairs

Teams

Phrase-based vs. Neural MT

Road Map

- Data Preparation
- Systems
  - Phrase-based
  - Neural
- Conclusion

Domain Adaptation

- How to best utilize large out-domain data?

Parallel Corpus	Tokens (en)	Helps TED tests?
TED	4.7M	
UN	489M	Harmful
QED	1.6M	No affect
OPUS	184M	?

- QED Test Sets
  - One combined system or separate systems?

Data Preparation

- Preprocessing
  - All arabic data segmented and normalized using MADAMIRA (Rambow et al. 2009)
  - English data tokenized using moses tokenizer
  - English→Arabic target data detokenized using Mada detokenizer (Kholy et al. 2010)
- Evaluation
  - avg. BLEU score on IWSLT test11-14

Phrase based Machine Translation

Phrase based System

Base Setup

- Framework: Moses (Koehn et al. 2007)
- Fast Aligner (Dyer et al. 2013)
- Default Moses parameters
- Lexicalized Reordering Model (Galley and Manning. 2008)
- Operation Sequence Model (Durrani et al. 2013)
- Neural Network Joint Model (Devlin et al. 2014)
- Kneser-Ney Smoothing Interpolated LM
- K-batch Mira (Cherry and Foster 2012)

Phrase based System

Key Experiments

- Data selection
  - Large out-of-domain data
    - not entirely relevant to the in-domain data
      - e.g. complete UN data hurts
  - Select a subset of the out-of-domain data
    - cross entropy difference (Axelrod et al. 2011)
    - +0.5 using MML (3.75% ~680K sentences)
    - +0.4 using Back-off Phrase-table
    - Opus was very helpful (+1.2)

Δ:+1.7

Phrase based System

Key Experiments

- Neural Network Joint Model (Devlin et al. 2014)
  - Baseline trained on TED corpus only (+0.7)

- NNJM Adaptation
  - Trained for 25 epochs on UN and OPUS data
  - Finetuned for 25 epochs on in-domain data (+0.2)

Δ:+0.9

Phrase based System

Key Experiments

- Baseline Operation Sequence Model
  - trained from concatenated parallel corpus
- Interpolated OSM (+0.6)
  - Train OSM models from each parallel corpus
  - Interpolate to minimize perplexity on tuning
- Class-based OSM (+0.1)

Δ:+0.7

Phrase based System

Results

Train	Avg. BLEU	Description
TED (baseline)	28.6	
TED + QED + UN	27.3 (-1.3)	Concatenation
TED + Back-off PT(QED,UN)	29.1 (+0.5)	
TED + MML (QED,UN)	29.2 (+0.6)	
TED + MML (QED,UN) + OPUS	30.4 (+1.8)	
Interpolated LM	30.9 (+2.3)	
Interpolated OSM	31.5 (+2.9)	
NNJM	32.1 (+3.5)	Train on concatenation
NNJM-Opus	32.3 (+3.7)	Train on OPUS, fine tune on TED
Class-based OSM	32.4 (+3.8)	
Drop-OOV	**32.6 (+4.0)**	

Phrase based System

Key Experiments

- QED Test-set
  - Phrase-table trained on concatenation
  - Use TED weights but replace TED with QED to be in-domain
    - for Language Model
    - for Interpolated OSM
  - NNJM: Fine-tuning with QED instead of TED

- English-to-Arabic Systems
  - Replicated what worked in Ar->En direction

Neural Machine Translation

Neural System

Base Setup

- Framework: Nematus (Sennrich et al. 2016)
- Bidirectional encoder model with attention
- BPE to avoid unknown words problem
- 1024 LSTM units in the encoder
- Batch size of 80
- Maximum sentence length of 80
- Dropout for only in-domain data

Neural System

Baseline

- Baseline system trained only on TED data

System	Avg. BLEU	Description
Phrase based	28.6	-

32.6
Phrase Based Best

Neural System
Baseline

- Baseline system trained only on TED data

System	Avg. BLEU	Description
Phrase based	28.6	-
Neural	**25.2**	-

32.6
Phrase Based Best

Neural System

Replicate best data selection

- Best MML settings that worked for the phrase-based system: 3.75% selected UN data

System	Avg. BLEU	Description
Phrase based MML 3.75%	29.2	**Data:** Selected UN + TED

32.6
Phrase
Based
Best

Neural System
Replicate best data selection

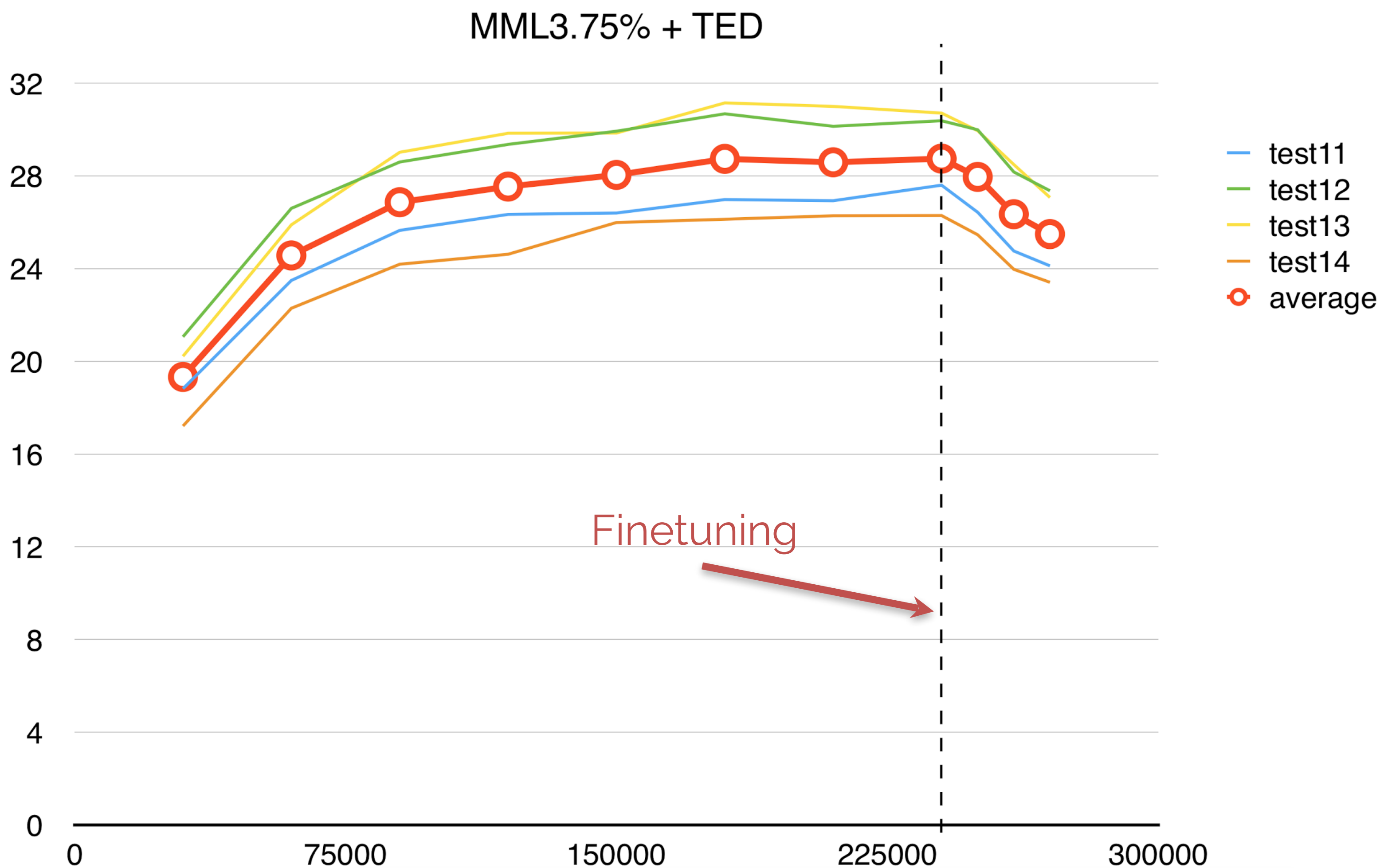
- Best MML settings that worked for the phrase-based system: 3.75% selected UN data

System	Avg. BLEU	Description
Phrase based MML 3.75%	29.2	**Data:** Selected UN + TED
Neural MML 3.75%	28.8	**Data:** Selected UN + TED

32.6
Phrase Based Best

Neural System

Why more data?



Neural System

Use more data

- Take the second best MML settings
  - UN10% (hurts in phrase-based by 0.4 points)

Train	Avg. BLEU	Description
Phrase based Baseline	28.6	**Data:** TED only
Phrase based MML 3.75%	29.2	**Data:** Selected UN + TED
Phrase based MML 10%	28.2	**Data:** Selected UN + TED
Neural MML 3.75%	28.8	**Data:** Selected UN + TED
Neural MML 10%	29.1	**Data:** Selected UN + TED

  - beats 3% but takes more time
  - be patient

32.6
Phrase
Based
Best

Neural System
Use all UN data

- Forget about selection, use all of the UN data

System	Avg. BLEU	Description
Phrase based best	32.6	**Data:** TED + QED + UN-MML + OPUS
Phrase based all UN	27.3	**Data:** UN + TED
Neural all UN	30.3	**Data:** UN + TED

Neural System

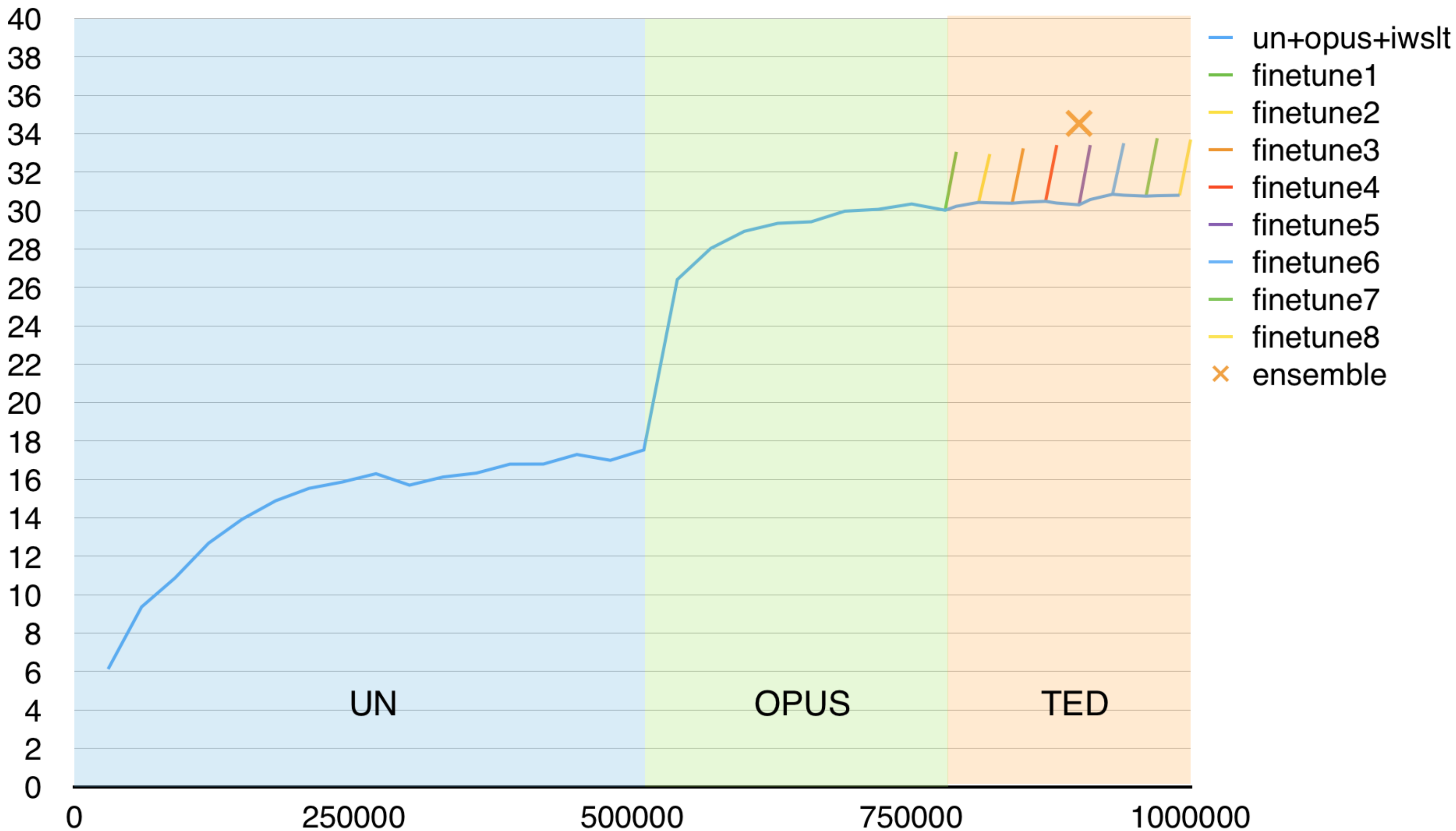
Final system

- Add subtitle (OPUS) data

System	Avg. BLEU	Description
Phrase based best	32.6	**Data:** TED + QED + UN-MML + OPUS
Neural individual	33.7	**Data:** UN -> OPUS -> TED
Neural ensemble	**34.6**	Ensemble of eight models

Neural System

NMT improvement lifetime



Neural System

English to Arabic direction

- Spent considerably less time on this direction because of computational limitations
- Replicated most of the training process from the other direction
- QED Systems: Finetune with QED data as in-domain

Neural System

Other Experiments

- Finetuning variants
  - Layer Freezing
- Dropout
- Data concatenation in base model
- BPE model training data selection

Conclusions

Other Experiments

- NMT is SOTA for Arabic-English language pair
  - have not utilized monolingual data yet (+3.0 BLEU, Sennrich et al. 2016)
- More data is better for NMT
  - as long as you have time
  - our best NMT system is trained on around 42M parallel sentences
- Adaptation is very cumbersome in Phrase Based systems
- Human effort involved in Neural MT is considerable less

Acknowledgment

- Rico Sennrich, Alexandra Birch and Marcin Junczys-Dowmunt (University of Edinburgh)
- Texas A&M Qatar for providing computational support

Thank you

References

- P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation,” in *Proceedings of the Association for Computational Linguistics (ACL’07)*, Prague, Czech Republic, 2007.
- D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *ICLR*, 2015. [Online]. Available: http://arxiv.org/pdf/1409.0473v6.pdf
- R. Sennrich, B. Haddow, and A. Birch, “Edinburgh neural machine translation systems for wmt 16,” in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 371–376. [Online]. Available: http://www.aclweb.org/anthology/W16-2323
- M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, “The united nations parallel corpus v1.0,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz˘, Slovenia, May 23-28, 2016*, 2016.
- H.Sajjad,F.Guzman,P.Nakov,A.Abdelali,K.Murray, F. A. Obaidli, and S. Vogel, “QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic spoken language translation,” in *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT-13)*, December 2013.
- P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*. European Language Resources Association (ELRA), May 2016.
- A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, “The AMARA corpus: Building parallel language resources for the educational domain,” in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, May 2014.
- A. Axelrod, X. He, and J. Gao, “Domain adaptation via pseudo in-domain data selection,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP ’11, Edinburgh, United Kingdom, 2011.
- N. Durrani, A. Fraser, H. Schmid, H. Hoang, and 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)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 399–405. [Online]. Available: http://www.aclweb.org/anthology/P13-2071

References

- N. Durrani, H. Sajjad, S. Joty, A. Abdelali, and S. Vogel, “Using joint models for domain adaptation in statistical machine translation,” in *Proceedings of the Fifteenth Machine Translation Summit (MT Summit XV)*. Florida, USA: AMTA
- N.Durrani,P.Koehn,H.Schmid,and A.Fraser,“Investigating the usefulness of generalized word representations in smt,” in *Proceedings of the 25th Annual Conference on Computational Linguistics*, ser. COLING’14, Dublin, Ireland, 2014, pp. 421–432.
- M.-T. Luong and C. D. Manning, “Stanford neural machine translation systems for spoken language domain,” in *International Workshop on Spoken Language Translation*, Da Nang, Vietnam, 2015.
- K. Heafield and A. Lavie, “CMU system combination in WMT 2011,” in *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 145–151. [Online]. Available: http://kheafield.com/professional/avenue/wmt 2011.pdf
- F. Guzman, H. Sajjad, S. Vogel, and A. Abdelali, “The AMARA corpus: Building resources for translating the web’s educational content,” in *Proceedings of the 10th International Workshop on Spoken Language Technology (IWSLT-13)*, December 2013.
- O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. Jimeno Yepes, P. Koehn, V. Logacheva, C. Monz, M. Negri, A. Neveol, M. Neves, M. Popel, M. Post, R. Rubino, C. Scarton, L. Specia, M. Turchi, K. Verspoor, and M. Zampieri, “Findings of the 2016 conference on machine translation,” in *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 131–198. [Online]. Available: http://www.aclweb.org/anthology/W/W16/W16-2301
- A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, “MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic,” in *Proceedings of the Language Resources and Evaluation Conference*, ser. LREC ’14, Reykjavik, Iceland, 2014, pp. 1094–1101.
- A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: A fast and furious segmenter for arabic,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 11–16. [Online]. Available: http://www.aclweb.org/anthology/N16-3003
- A. Birch, M. Huck, N. Durrani, N. Bogoychev, and P. Koehn, “Edinburgh SLT and MT system description for the IWSLT 2014 evaluation,” in *Proceedings of the 11th International Workshop on Spoken Language Translation*, ser. IWSLT ’14, Lake Tahoe, CA, USA, 2014.

References

- C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of ibm model 2," in *Proceedings of NAACL'13*, 2013.
- K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197. [Online]. Available: http://kheafield.com/professional/avenue/kenlm.pdf
- M. Galley and C. D. Manning, "A Simple and Effective Hierarchical Phrase Reordering Model," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, October 2008, pp. 848–856. [Online]. Available: http://www.aclweb.org/anthology/D08-1089
- N. Durrani, H. Schmid, A. Fraser, P. Koehn, and H. Schu ̈tze, "The Operation Sequence Model – Combining N-Gram-based and Phrase-based Statistical Machine Translation," *Computational Linguistics*, vol. 41, no. 2, pp. 157–186, 2015.
- J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.
- S. Joty, H. Sajjad, N. Durrani, K. Al-Mannai, A. Abdelali, and S. Vogel, "How to Avoid Unwanted Pregnancies: Domain Adaptation using Neural Network Models," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.
- C. Cherry and G. Foster, "Batch tuning strategies for statistical machine translation," in *Proceedings of the 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ser. NAACL-HLT '12, Montre ́al, Canada, 2012.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02, Morristown, NJ, USA, 2002, pp. 311– 318.
- H. Sajjad, A. Fraser, and H. Schmid, "A statistical model for unsupervised and semi-supervised transliteration mining," in *Proceedings of the Association for Computational Linguistics (ACL'12)*, Jeju, Korea, 2012.

References

- N. Durrani, H. Sajjad, H. Hoang, and P. Koehn, “Integrating an unsupervised transliteration model into statistical machine translation,” in *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2014)*, Gothenburg, Sweden, April 2014.
- R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016, pp. 1715–1725. [Online]. Available: http://www.aclweb.org/anthology/P16-1162