AFRL-RH-WP-TR-2018-0105

THE AFRL IWSLT 2018 SYSTEMS: WHAT WORKED, WHAT DIDN'T

Brian Ore
Eric Hansen
Grant Erdmann
Jeremy Gwinnup

Human Trust and Interaction Branch

Katherine Young

N-Space Analysis

October 2018
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2018-0105 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//	//signature//
ERIC HANSEN, Work Unit Manager	RICHARD D. SIMPSON, DR-IV, DAF
Human Trust and Interaction Branch	Chief, Human-Centered ISR Division
Airman Systems Directroate	Airman Systems Directorate
711th Human Performance Wing	711th Human Performance Wing
Air Force Research Laboratory	Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE ***(DD-MM-YY)***
04-10-18

2. REPORT TYPE
Interim

3. DATES COVERED ***(From - To)***
01-10-17 – 01-09-18

4. TITLE AND SUBTITLE
The AFRL IWSLT 2018 Systems: What Worked, What Didn't

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Brian Ore, Eric Hansen, Katherine Young*, Grant Erdmann and Jeremy Gwinnup

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H07P

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
*N-Space Analysis, LLC

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Human-Centered ISR Division
Human Trust and Interaction Branch
Wright-Patterson AFB, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHXS

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2018-0105

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release.

13. SUPPLEMENTARY NOTES
88ABW-2018-4946, 4 October 2018

14. ABSTRACT
This report summarizes the Air Force Research Laboratory (AFRL) machine translation (MT) and automatic speech recognition (ASR) systems submitted to the spoken language translation (SLT) and low-resource MT tasks as part of the IWSLT18 evaluation campaign

15. SUBJECT TERMS
machine translation, automatic speech recognition, spoken language translation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Eric Hansen
Unclassified	Unclassified	Unclassified	SAR	10	**19b. TELEPHONE NUMBER** *(Include Area Code)* N/A

The AFRL IWSLT 2018 Systems: What Worked, What Didn't

Brian Ore, Eric Hansen, Katherine Young, Grant Erdmann and Jeremy Gwinnup

Air Force Research Laboratory

{brian.ore.1, eric.hansen.5, katherine.young.1.ctr, grant.erdmann, jeremy.gwinnup.1}@us.af.mil

Abstract

This report summarizes the Air Force Research Laboratory (AFRL) machine translation (MT) and automatic speech recognition (ASR) systems submitted to the spoken language translation (SLT) and low-resource MT tasks as part of the IWSLT18 evaluation campaign.

1. Introduction

As part of the evaluation campaign for the 2018 International Workshop on Spoken Language Translation (IWSLT18) [1], the AFRL Human Language Technology team applied and improved techniques from previous workshops [2] and Conference on Machine Translation efforts [3] to the Spoken Language Translation and Low-Resource Machine Translation tasks.

2. Spoken Language Translation

2.1. Automatic Speech Recognition

This section describes the ASR systems that were developed for the baseline condition of the Speech Translation task. We trained two different English systems and performed system combination to obtain the final hypothesis for translation. Section 2.1.1 describes that language models (LMs) that were used for decoding and rescoring. Section 2.1.2 discusses the Kaldi ASR system, and Section 2.1.3 describes the Hidden Markov Model ToolKit (HTK) Tensorflow system. Finally, Section 2.1.4 describes how we segmented the test data and performed system combination.

2.1.1. Language Models

LMs were estimated on the provided TED data and subsets of News Crawl 2007-2017 and News Discussions versions 1-3. The subset of each news corpus was selected using cross-entropy difference scoring [4] with TED as the in-domain text, and selection thresholds were chosen to use 1/8 of each corpus to train N-gram LMs, and 1/16 of each corpus to train a recurrent neural network (RNN) LM. Interpolated bigram, trigram, and 4-gram LMs were estimated using the SRILM Toolkit,[1] and a RNN maximum entropy LM was trained using the RNNLM Toolkit.[2] The RNN included 160 hidden units, 300 classes in the output layer, 4-gram features for the direct connections, and a hash size of 10^9. The LM vocabulary included 100,000 words that were chosen using the select-vocab tool from SRILM.

Table 1: Kaldi WER. Decoding was performed using a trigram LM trained on TED.

Acoustic Training Data	dev2010	tst2010	tst2013
Speech-Translation TED	19.8	19.6	30.5
TEDLIUM	16.9	14.8	22.3
Combined	16.6	15.1	23.6

2.1.2. Kaldi System

The acoustic training data available for this year's evaluation included the Speech-Translation TED corpus and the TEDLIUM corpus. Based on a preliminary analysis of the Speech-Translation TED corpus, we removed all segments longer than 15 seconds from this corpus. The devtest and off-limit talks were sequestered from TEDLIUM, and a third data set was created by searching the Speech-Translation TED and TEDLIUM corpora for non-overlapping time segments. Next, an initial set of ASR systems were trained on each data set using the Kaldi open source speech recognition toolkit [5]. All Kaldi models discussed in this paper are based on the chain time delay neural network (TDNN)-rectified linear unit (ReLU) setup using i-vectors.[3] Standard data augmentation methods were applied during the Mel frequency cepstral coefficient (MFCC) feature generation stage, such as speech and volume perturbation. Each system was decoded using the same trigram LM, which was estimated from the provided TED data using the SRILM toolkit. Table 1 shows the word error rate (WER) obtained on dev2010, tst2010, and tst2013.

Based on the results in Table 1, a Kaldi ASR system was trained on TEDLIUM using the interpolated bigram LM described in Section 2.1.1. This model was then used to decode all of the audio from the Speech-Translation TED corpus (including segments longer than 15 seconds), and the ASR derived transcripts were folded in with the TEDLIUM data, as in a semi-supervised training scenario, to build the final Kaldi ASR system. This data set is referred to as TEDLIUM+ASR

[1] http://www.speech.sri.com/projects/srilm

[2] http://www.fit.vutbr.cz/$\sim$imikolov/rnnlm

[3] http://github.com/kaldi-asr/kaldi/tree/master/egs/swbd/s5c/local/chain

Table 2: Kaldi WER. Decoding was performed using an interpolated bigram LM, and rescoring was applied using an interpolated 4-gram and RNN LM.

ASR System	dev2010	tst2010	tst2013
Kaldi TEDLIUM	14.0	11.9	17.7
Kaldi TEDLIUM+ASR	13.5	11.4	17.0

in the remainder of this paper.

The test data was decoded as follows. First, the recognition lattices from the Kaldi bigram system were rescored with the 4-gram LM. Next, 1000-best lists were extracted from each lattice and rescored with the RNN LM. The final LM scores were obtained by linearly interpolating the log probabilities from the 4-gram and RNN LM. Interpolation weights of 0.25 for the 4-gram and 0.75 for the RNN were chosen based on results from previous experiments. Table 2 shows the final WER obtained with each system. Based on these results, we used the TEDLIUM+ASR system in all remaining experiments.

2.1.3. HTK-Tensorflow System

A hybrid neural network hidden Markov model (HMM) speech recognition system was developed using Tensorflow [6] and a version of HTK[4] that we modified according to the method of [7]. First, a Gaussian mixture model (GMM)-HMM system was trained on TEDLIUM. Phonemes were modeled using word-position-dependent state-clustered across-word triphones, and the final HMM set included 6000 shared states with an average of 28 mixtures per state. The feature set consisted of 12 perceptual linear prediction (PLP) coefficients, plus the zeroth coefficient, with mean and variance normalization applied on a per talk basis. Delta, acceleration, and third differential coefficents were appended to form a 52 dimensional vector, and heteroscedastic linear discriminant analysis (HLDA) was used to reduce the feature dimension to 39. Speaker adaptive training (SAT) was applied using constrained maximum likelihood linear regression (CMLLR) transforms, and the models were discriminatively trained using the minimum phone error (MPE) criterion.

A residual network (ResNet) was trained on the TEDLIUM+ASR data set described in Section 2.1.2. This network is based on the 18-layer network described in [8], with the batchnorm and ReLU activations moved to utilize full pre-activation residual units described in [9], and an additional fully connected layer for i-vector input. Figure 1 shows the ResNet structure. A context window of 17 was applied to the feature input, which included 40 log filterbank outputs normalized to zero mean and unit variance on a per talk basis. The 100 dimensional i-vectors were extracted on a per-talk basis with an i-vector extractor that was trained on

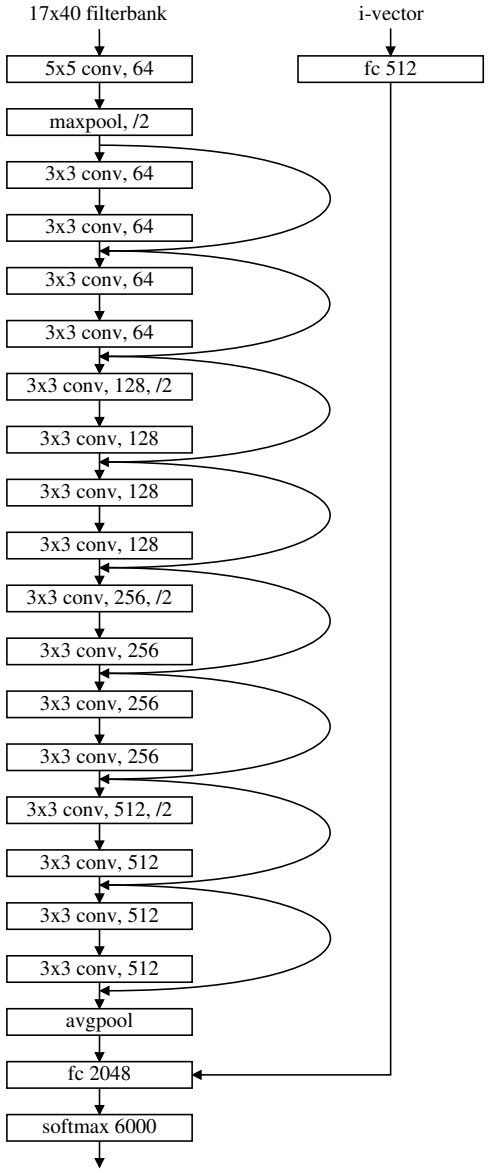

Figure 1: ResNet architecture based on [8, 9] with convolutional (conv), max pooling (maxpool), average pooling (avgpool), and fully connected (fc) layers. $H \times W$ is the filter size and $/2$ indicates that a stride of 2 was applied.

TEDLIUM using the same procedure as our IWSLT 2015 system [10]. Cross entropy training was performed using a mini-batch size of 512 and an initial learning rate of 0.0005 that was adjusted according to the QuickNet newbob algorithm.[5]

Recognition lattices were produced using HDecode with the interpolated trigram LM described in Section 2.1.1, and then rescored with the 4-gram and RNN LM using the same procedure as the Kaldi system. Next, confidence scores were estimated at the acoustic frame level by aligning the 20-best hypotheses for each utterance and counting the number of matching HMM states. An adapted ResNet was estimated

[4] `http://htk.eng.cam.ac.uk`

[5] `http://www.icsi.berkeley.edu/Speech/faq/nn-train.html`

Table 3: HTK-Tensorflow WER. Decoding was performed using an interpolated trigram LM, and rescoring was applied using an interpolated 4-gram and RNN LM.

ASR System	dev2010	tst2010	tst2013
ResNet	15.4	12.9	17.5
ResNet-Adapted	15.3	12.6	16.1

for each talk using frames that had a confidence score of 0.9 or higher and a single epoch of cross-entropy training with a learning rate of 0.0000625. Finally, the test set was decoded a second time and LM rescoring was reapplied. Table 3 shows the WER on dev2010, tst2010, and tst2013.

2.1.4. Test Segmentation and System Combination

The WER results reported in the previous sections were obtained by evaluating each ASR system on the automatically derived segments from the baseline implementation.[6] It was discovered that the segment boundaries did not always align with non-speech; therefore, we decided to use an alternative segmentation method.

A neural network based speech activity detector (SAD) was developed using Tensorflow. The SAD was trained on 40 hours from the TEDLIUM corpus using the automatically generated phoneme alignments from the HTK GMM-HMM system to define the speech/non-speech boundaries. The network included a context window of 41 frames on the input, a hidden layer of 1024 neurons with rectified linear activation functions, and 2 output units corresponding to speech and non-speech. The feature set consisted of 40 log filterbank outputs that were normalized to zero mean and unit variance. Automatic segmentation of the test data was performed by evaluating the SAD, applying a dynamic programming algorithm to choose the best sequence of states, and defining utterance boundaries at the midpoint of each non-speech segments longer than 0.5 seconds. Lastly, non-speech segments longer than 1.0 second were trimmed from each utterance.

The final hypothesis was selected by applying N-best recognizer output voting error reduction (ROVER) to the output from the Kaldi TEDLIUM+ASR and HTK-Tensorflow ResNet-Adapted system. Table 4 shows the WER obtain using the updated segmentation. Comparing Table 4 with the results in Table 2 and 3, we can see that the updated segmentation method provided a substantial improvement in WER.

2.2. ASR Postprocessing

We employed the provided SLT.KIT punctuator component to re-punctuate our ASR output before applying a truecaser model to induce the most common case for an English word before translating with the Marian section described in the next section.

[6] `http://github.com/isl-mt/SLT.KIT`

Table 4: WER using the updated test segmentation method. The final ASR hypothesis was obtained using N-best ROVER.

ASR System	dev2010	tst2010	tst2013
Kaldi TEDLIUM+ASR	9.5	7.7	12.8
ResNet-Adapted	11.2	8.6	11.2
N-best ROVER	9.5	6.9	9.8

Table 5: English-German cased BLEU scores for the SLT task. For comparison purposes, this table includes the scores obtained with the reference English source text.

English Transcripts	dev2010	tst2010	tst2013
Reference	27.10	27.40	28.83
ASR	18.48	17.20	18.40

2.3. Machine Translation

Lastly, a Marian [11] neural machine translation system was employed to translate the repunctuated text from English into German. This system was trained on the 41 million lines of preprocessed data provided by the WMT18 organizers for the news-translation shared task[12]. The data was truecased for uniformity, then a byte-pair encoding (BPE) [13] model was trained jointly on the source and target data with 90k merge operations.

As described in our WMT18 news-task efforts[3], we used the same parameters in training our Marian transformer [14] model:

- We used an encoding depth of 6 layers and a decoding depth of 6 layers.
- We used 8 transformer heads.
- We held the vocabulary size constant during training to 90k entries each for source and target.
- We held the word embedding dimensionality to 512 for all models.
- We used 1024 units in the hidden layer (where appropriate).
- We exclusively used the WMT `newstest2014` test set for validation.

2.4. Results

Results of scoring our repunctuated, translated ASR output and various references are shown in Table 5.

3. Low-Resource Machine Translation

For the low-resource translation task, we tried a variety of approaches with Marian [11], and Moses[15] toolkits. We

Table 6: Corpus size for each language pair in training corpus

Lang. Pair	Lines
Basque–English	5,623
French–English	288,366
Spanish–English	278,297
Total corpus	572,286

tried additional approaches with stemming and morphological processing, but systems trained with data processed in this manner were not ready in time for evaluation submission.

3.1. Common Training Corpus

For many of the experiments across different toolkits and systems, we constructed a common training corpus with uniform preprocessing in order to reduce variables when comparing different conditions.

Using the provided parallel Basque–English, French–English, and Spanish–English TED corpora [16], we construct a training corpus containing all three language pairs. Sizes of each portion of the training corpus are listed in Table 6. A joint BPE model was trained with 89,500 merge operations on the combination of all languages in the training data, then applied to the unified training corpus.

A similar corpus for use in backtranslation was constructed from the provided English–Basque, French–Basque, and Spanish–Basque corpora. Due to the small size of each of these component corpora, we also add the Basque–English portion of the OpenSubititles Corpus[7]. Sizes of each portion of this backtranslation training corpus are listed in Table 7. The BPE model from the 'forward' was used to segment the source and target data.

For some Marian experiments, we also constructed monolingual Basque and English corpora for use in constructing pretrained word embeddings. We use 50 million lines from the English monolingual CommonCrawl corpus selected for use in backtranslation from our WMT17 news-task efforts [17]. Additional monolingual Basque data was taken from the Commoncrawl website[8] and language-filtered using a modified C implementation[9] of the algorithm outlined in [18], yielding a Basque monolingual corpus of 38 million lines. We then apply BPE to each of these corpora with the same model as above and use word2vec[19] to generate 512-dimension word embeddings compatible with our settings in Marian.

3.2. Marian Systems

We spent the bulk of our efforts building systems with the Marian toolkit, experimenting with a variety of settings along two major categories: Sentence-weighting and backtranslated systems.

[7] `http://www.opensubtitles.org`
[8] `http://www.commoncrawl.org`
[9] `https://github.com/saffsd/langid.c`

Table 7: Corpus size for each language pair in backtranslation training corpus

Lang. Pair	Lines
English–Basque	5,623
French–Basque	6,948
Spanish–Basque	6,668
Basque–English OpenSubtitles	458,380
Total corpus	477,619

3.2.1. Sentence-Weighted training

We used the "forward" corpus outlined in Section 3.1 to train Marian systems with the same network parameters as outlined in the SLT translation system in Section 2.3. In Table 8, we note our baseline system (#1) score 11.11 cased BLEU on dev2018. Next, we utilize the sentence-weighting feature of Marian that allows each sentence to be assigned a "weight" to determine how much of an effect each will have during training. A score of 1.0 is assigned to sentences from the Basque–English portion of the training corpus, French–English and Spanish–English sentences are assigned a score of 0.5. The system trained with these weights (#2) shows a +2.41 increase in BLEU.

Using the same data as system #2, we train a system that uses BEER[22] as the validation metric. While we have seen performance gains using this tactic in other work, here the resulting system(#3) performs -0.75 BLEU worse than the previous system.

Next, we consider averaging and ensembling of models. We take the 4-best model checkpoints from system #2 and average them into a single model, resulting in system #4's +0.87 BLEU gain over system #2.

Lastly, we decode with an ensemble of system #4 and a model averaged from the four best checkpoints of system #3, resulting in a BLEU score of 15.45. This system (#5) was then submitted as our entry to the low-resource MT task.

3.2.2. Backtranslated training corpus

As a contrast, we use the "backtranslation" corpus to train a shallow "s2s" Marian system that translates English, Spanish, and French into Basque. We then translate a 2 million line portion of the English monolingual corpus described in 3.1 into something resembling Basque and then use the combination of the two in conjunction with the small amount of provided Basque–English data to train two Marian "bi-deep" [20, 21] systems, both using BEER [22] as the training validation metric. These systems are listed as #6 (without pretrained word embeddings) and #7 (with pretrained word embeddings) in Table 8. We note that pretrained word embedding system scores -1.46 cased BLEU lower than the equivalent system without the pretrained embeddings, counter to

Table 8: Results for various MT systems decoding Basque–English dev2018 measured in cased BLEU. Our submission system is highlighted in bold text.

#	System	BLEU
1.	marian-eseufr-trans	11.11
2.	marian-eseufr-trans-weight	13.52
3.	marian-eseufr-trans-weight-beervalid	12.77
4.	marian-eseufr-trans-weight-avg4	14.39
5.	**marian-eseufr-trans-weight-avg4X2**	**15.45**
6.	marian-bt-bideep-beervalid	11.23
7.	marian-bt-bideep-preembed-beervalid	9.77
8.	moses-bt-bpe	14.06

our experience with our WMT18 systems.

3.3. Moses System

Using both the provided Basque–English data and the back-translated corpus outlined in Section 3.2.2 we train a Moses system in a similar vein to the one employed in our WMT18 submission: This system employed a hierarchical reordering model [23] and 5-gram operation sequence model [24]. The 5-gram English language model was trained with KenLM on the constrained monolingual corpus from our WMT15 [25] efforts. Our uniform BPE model used was applied to the parallel training data, but the language modelling corpus used the Russian–English joint BPE model from our WMT18 submission, possibly degrading performance due to this BPE mismatch. System weights were tuned with the Drem [26] optimizer using the "Expected Corpus BLEU" (ECB) metric.

This system, listed as #8 in Table 8 performs better than the two other Marian-based backtranslation systems (#6 and #7).

3.4. Results

Results of various systems described in the above sections are listed in Table 8. Our final submission system (#5) is highlighted in bold text.

4. Conclusions

Our experimentation this year show positive results in spoken language translation, especially our ASR component. However, for the low-resource MT task, we note that various approaches we have previously employed with great effect in high-resource conditions need further adaptation and refinement when scaling down to extremely low-resource conditions.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 4 Oct 2018. Originator Reference Number: RH-18-118975 Case Number: 88ABW-2018-4946.

5. References

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2018 Evaluation Campaign," in *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, Bruges, Belgium, October 2018.

[2] M. Kazi, , E. Salesky, B. Thompson, J. Taylor, J. Gwinnup, T. Anderson, G. Erdmann, E. Hansen, B. Ore, K. Young, and M. Hutt, "The MITLL-AFRL IWSLT-2016 systems," in *Proc. of the 13th International Workshop on Spoken Language Translation (IWSLT'16)*, Seattle, Washington, December 2016.

[3] J. Gwinnup, T. Anderson, G. Erdmann, and K. Young, "The afrl wmt18 systems: Ensembling, continuation, combination," in *Proceedings of the Third Conference on Machine Translation*. Brussels, Beligum: Association for Computational Linguistics, October 2018.

[4] R. Moore and W. Lewis, "Intelligent selection of language model training data," in *Association Computational Linguists 2010 Conference Short Papers*, Uppsala, Sweden, July 2016.

[5] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2011.

[6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42, January 2012.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, June 2016.

[9] ——, "Identity mappings in deep residual networks," in *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, October 2016.

[10] M. Kazi, B. Thompson, E. Salesky, T. Anderson, G. Erdmann, E. Hansen, B. Ore, K. Young, J. Gwinnup, M. Hutt, and C. May, "The MITLL-AFRL IWSLT 2015 systems," in *Proc. of the 12th International Workshop on Spoken Language Translation (IWSLT'15)*, Da Nang, Vietnam, December 2015.

[11] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in c++," in *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, 2018, pp. 116–121. [Online]. Available: http://aclweb.org/anthology/P18-4020

[12] O. Bojar, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, and C. Monz, "Findings of the 2018 conference on machine translation (WMT18)," in *Proceedings of the Third Conference on Machine Translation*. Brussels, Belgium: Association for Computational Linguistics, October 2018.

[13] 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)*. Association for Computational Linguistics, 2016, pp. 1715–1725.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[15] 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 *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07, 2007, pp. 177–180.

[16] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[17] J. Gwinnup, T. Anderson, G. Erdmann, K. Young, M. Kazi, E. Salesky, B. Thompson, and J. Taylor, "The afrl-mitll wmt17 systems: Old, new, borrowed, bleu," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 303–309. [Online]. Available: http://www.aclweb.org/anthology/W17-4728

[18] M. Lui and T. Baldwin, "Cross-domain feature selection for language identification," in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 2011, pp. 553–561. [Online]. Available: http://www.aclweb.org/anthology/I11-1062

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *International Conference on Learning Representations (ICLR) Workshop*, 2013.

[20] A. V. Miceli Barone, J. Helcl, R. Sennrich, B. Haddow, and A. Birch, "Deep architectures for neural machine translation," in *Proceedings of the Second Conference on Machine Translation*. Association for Computational Linguistics, 2017, pp. 99–107. [Online]. Available: http://aclweb.org/anthology/W17-4710

[21] R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. Miceli Barone, and P. Williams, "The university of edinburgh's neural mt systems for wmt17," in *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 389–399. [Online]. Available: http://www.aclweb.org/anthology/W17-4739

[22] M. Stanojević and K. Sima'an, "Fitting sentence level translation evaluation with many dense features," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 202–206. [Online]. Available: http://www.aclweb.org/anthology/D14-1025

[23] M. Galley and C. D. Manning, "A simple and effective hierarchical phrase reordering model," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08, 2008, pp. 848–856.

[24] N. Durrani, H. Schmid, and A. Fraser, "A joint sequence translation model with integrated reordering," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon, June 2011, pp. 1045–1054.

[25] J. Gwinnup, T. Anderson, G. Erdmann, K. Young, C. May, M. Kazi, E. Salesky, and B. Thompson, "The afrl-mitll wmt15 system: There's more than one way to decode it!" in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, September 2015, pp. 112–119. [Online]. Available: http://aclweb.org/anthology/W15-3011

[26] G. Erdmann and J. Gwinnup, "Drem: The AFRL submission to the WMT15 tuning task," in *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September 2015, pp. 422–427.