Supervised Adaptation of Sequence-to-Sequence Speech Recognition Systems using Batch-Weighting

Christian Huber, Juan Hussain, Tuan-Nam Nguyen, † Kaihang Song, Sebastian Stüker
KIT - Karlsruhe Institute of Technology
firstname.lastname@kit.edu
† uoell@student.kit.edu

Alexander Waibel
Carnegie Mellon University
alexander.waibel@cmu.edu

Abstract

When training speech recognition systems, one often faces the situation that sufficient amounts of training data for the language in question are available but only small amounts of data for the domain in question. This problem is even bigger for end-to-end speech recognition systems that only accept transcribed speech as training data, which is harder and more expensive to obtain than text data.

In this paper we present experiments in adapting end-to-end speech recognition systems by a method which is called batch-weighting and which we contrast against regular fine-tuning, i.e., to continue to train existing neural speech recognition models on adaptation data. We perform experiments using theses techniques in adapting to topic, accent and vocabulary, showing that batch-weighting consistently outperforms fine-tuning.

In order to show the generalization capabilities of batch-weighting we perform experiments in several languages, i.e., Arabic, English and German. Due to its relatively small computational requirements batch-weighting is a suitable technique for supervised life-long learning during the life-time of a speech recognition system, e.g., from user corrections.

1 Introduction

When building an automatic speech recognition (ASR) system for a specific domain, one is often faced with the fact that only very limited amounts of training data for the target domain are available. This problem has become bigger with the advent of end-to-end speech recognition systems. The old ASR systems used the Bayes theorem to solve the speech recognition with the help of an acoustic model (AM) and a language model (LM). The acoustic model needed to be trained on transcribed speech recordings, the language model was trained on text only. Therefore, topic adaptation, could be done with the help of textual training data only, by adapting or training a language model on topic specific data. As text only data is easier to come by than transcribed speech, topic adaptation was often feasible. For end-to-end speech recognition systems this option is no longer available, as they only accept transcribed speech as training data. However, transcribed speech for a specific domain is often more difficult to find than text data, thus making it more difficult or expensive to find or create fitting adaptation data.

For HMM based ASR systems that use Gaussian mixture models (GMMs) for estimating the emission probabilities of the HMM, several techniques for adapting to speakers or channels were available (Gales et al., 1996; Gales, 1998; Puming Zhan and Westphal, 1997). These techniques often could also be applied in an unsupervised or semi-supervised manner during inference.

For end-to-end ASR systems such techniques need to be newly created. In this paper we are examining the use of fine-tuning for end-to-end ASR for adapting them to different domains. We thereby examine different dimensions of domain adaptation, such as adapting to topics, accents and vocabulary. In the experiments we compare fine-tuning, i.e., continuing to train an end-to-end system on adaptation data, to a technique called batch-weighting, in which we mix adaptation data with the data for training the background model in a certain ratio at mini-batch level. Batch-weighting was thereby inspired by a technique from machine translation (Wang et al., 2017) and is explained in detail in section 5. We adapt this technique for automatic speech recognition.

Further, for the different dimensions of domain adaptation, we examine the fine-tuning of different parts of the end-to-end ASR systems, e.g., only the encoder or only the decoder, in order to test the hypothesis that encoder layers are mainly con-

Proceedings of the Second Workshop on Life-long Learning for Spoken Language Systems, pages 9–17
Online, December 7, 2020. ©2020 Association for Computational Linguistics

cerned with learning features and acoustic properties, while the decoder models the linguistic properties of the recognizer's domain.

We performed our experiments in several languages — Arabic, English and German — and in different domain scenarios in order to show batch-weighting's generalization capabilities to new adaptation scenarios. Our experiments thereby show that batch-weighting consistently outperforms simple fine-tuning.

We also report the computational time needed for performing batch-weighting in the different scenarios. The low computation times (less than six hours in all cases) makes this technique suitable for life-long learning, when small amounts of supervised adaptation data can be collected during the life of a system, e.g., by user corrections.

2 Related Work

There are a several studies about adaptation of Neural Machine Translation (NMT) systems. Chu and Wang (2018) divided NMT domain adaptation methods into four categories: Data centric, training objective centric, architecture centric and decoding centric:

- **`Data centric`**: (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013) selected the sentences that are similar to in-domain data from out-of-domain data.

- **`Training objective centric`**: Chen et al. (2017) used sentence weighting for adaptation of part-of-speech (POS) tagging, a named entity (NE) recognition task. Wang et al. (2017) used sentence weighting, domain weighting and batch weighting for NMT, and Yan et al. (2019) used word weighting. (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Neubig and Hu, 2018) used fine-tuning, and Chu and Dabre (2019) used mixed fine-tuning, while Kobus et al. (2016); Chu et al. (2017) combined mixed fine-tuning and adding domain tag.

- **`Architecture centric`**: Baniata et al. (2018) used shared decoder and domain specific encoders to adapt NMT for new language. (Gu et al., 2019; Britz et al., 2017) trained Domain Discriminator and NMT model with some part shared in parallel.

- **`Decoding centric`**: Gulcehre et al. (2015) used shallow fusion, whose outputs are generated by the weighted sum of the NMT and RNNLM probabilities. Dou et al. (2019) combined shallow fusion, deep fusion and domain differential.

In the area of end-to-end speech recognition Nguyen et al. (2019) trained a speech recognition system on a multi-domain corpus. By using a domain identification (DI) vector derived from the activation of a bottle-neck layer in a domain classifying network they prime their speech recognition system to different domains present in the training data. The paper then shows improvements in Word Error Rate (WER) when adapting the speech recognition system in an unsupervised manner using the DI vector to a domain for which no training or adaptation data is available. In contrast in our experiments we work with small amounts of adaptation data that are not sufficient for training a complete system, but can be exploited for adapting an existing system.

3 Dimensions of Adaption

In our work we examine the adaptation of our system to different dimensions of variability. Sometimes these different dimensions are subsumed under the term *domain* (Nguyen et al., 2019). However, on other occasions *domain* is used to describe the topic of speech data only. We will follow the latter use of *domain* in this paper and will explicitly address the different dimensions of the speech data for which we examined suitable adaptation techniques: Topic adaptation, accent adaptation and vocabulary adaptation.

3.1 Topic Adaptation

In some situations, we need a model that works well in a specific domain, but the target domain data-set (also known as in-domain data-set) is too small to train a meaningful ASR model alone. In order to obtain a high-performance model in the low resource target domain, we adapt a well trained general seq2seq model to the target domain.

3.2 Accent Adaptation

In some situations, it may be difficult to recognize the audio of non-native speakers correctly. The speakers often have a significant non-native accent which does not match the training data from native speakers or even other non-native speakers. Our

in-domain training data consists just of a few hours audio of non-native speakers of a specific native language. We adapt the seq2seq model on both specific accent domain and multi-accent domain with our own data-sets.

3.3 Vocabulary Adaptation

In some situations, it may be crucial to recognize specific topic words correctly. The Word Error Rate (WER) usually does not reflect the performance on these specific words, therefore we evaluate if these words are recognized correctly via another metric. We recorded data-sets containing certain words of the new domains which the baseline systems don't recognize correctly. To measure how well our systems recognize the new words, we calculate an word accuracy (WA) where a new word is counted as recognized if and only if it is contained in the hypothesis.

4 Data

We conduct experiments on the languages English, German and Arabic, in order to make sure that batch-weighting generalizes across languages. Tables 1, 2 and 3 contain a summary of the speech data-sets that we used as general, out-of-domain training data-sets (out), and of the in-domain data-sets (in) that match our target domain.

4.1 Out-of-domain Data-Sets

4.1.1 English

The baseline system in English has been trained on the `TED-LIUM` (Rousseau et al., 2012) and `How2` (Sanabria et al., 2018) corpus. We divided 789 hours of speech as the training set, 18.3 hours as the validation set and 2.6 hours as the test set.

4.1.2 German

The German baseline system has been trained on 433 hours of speech data consisting of speech from the European Parliament, radio news and lectures. A test set of 50 minutes was randomly selected from this domain and excluded from the training data.

4.1.3 Arabic

The baseline system in Arabic has been trained on `Alj.1200h`. It consists of 1200 hours of broadcast videos recorded during 2005–2015 from the Aljazeera Arabic TV channel as described in Ali et al. (2016). As reported, 70% of this set is in Modern Standard Arabic (MSA) and the rest is Dialectal Arabic (DA), such as Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR). The categories of the speech range from conversation (63%), interview (19%), to report (18%). The used test set `Alj-MSA+dialect.10h` of 10 hours is described in Ali et al. (2016) as well. It includes non-overlapped speech from Aljazeera, which was prepared according to Ali et al. (2016) for an Arabic multi-dialect broadcast media recognition challenge. For our task, we normalized Hamza and Alif. The test set `Alj-MSA.2h` is a subset from `Alj-MSA+dialect.10h` where we cut only MSA utterances free from dialects from the beginning of the set until we reached the duration of 2 hours.

Corpus	Speech data	Utterances
A: Training Data		
How2+Ted (out)	789 h	473K
How2+Ted (out) validation set	18.3 h	11K
Atis (in)	3.6 h	1800
Japanese accent data-set (in)	4.7 h	2227
Multi accent data-set (in)	8.7 h	8986
B: Test Data		
How2+Ted (out) test set	2.6 h	1155
Atis (in) test set	34.4 min	355
Japanese accent data-set (in) test set	1.2 h	496
Multi-accent data-set (in) test set	2.1 h	2135

Table 1: Summary of the English speech data-sets

Corpus	Speech data	Utterances
A: Training Data		
Training set (out)	433 h	321K
MINI-Questions training set (in)	3.7 h	1441
MINI-Questions validation set (in)	23 min	173
B: Test Data		
Test set (out)	50 min	605
MINI-Questions test set (in)	18 min	189
New words test set (in)	32 min	325

Table 2: Summary of the German speech data-sets

4.2 Topic Adaptation Data-Sets

4.2.1 English

For topic adaptation in English we use the `ATIS` data-set. `ATIS` (Air Travel Information Services) contains speech about various hypothetical travel planning scenarios from 36 speakers. There are many American city names, airport names and abbreviations, that makes the general model perform

Corpus	Speech data	Utterances
A: Training Data		
Alj.1200h (out)	1200 h	374K
MINI-Quest-Ans.3.34h (in)	3.34 h	915
B: Test Data		
Alj-MSA.2h (out)	2 h	1188
MINI-ANS.42m (in)	42 min	224
MINI-Ques.50m (in)	50 min	225

Table 3: Summary of the Arabic speech data-sets

poorly. We used 3.6 hours of speech as training set and 34.4 minutes of speech as test set.

4.2.2 German

For topic adaptation in German we recorded 3.7 hours of data from six speakers (called `MINI-Questions training set`) containing questions of a psychological interview (Sheehan et al., 1998). For the `MINI-Questions validation` and `test set` we recorded 23 minutes and 18 minutes of speech, respectively, from one speaker. The speakers of the `MINI-Questions training` and `test set` are disjoint. For the recording we employed our online application TEQST[1] which allows the user to read texts and record with their own mobile devices.

4.2.3 Arabic

The data-set `MINI-Quest-Ans.3.34h` for topic adaptation in Arabic consists of 915 utterances and 3.34 hours of reading MINI (Sheehan et al., 1998) questions and free answers from two speakers. We transcribed the answers with our ASR system and then corrected them manually with the application DaCToR (Hussain et al., 2020) which we used to record a part of data and to correct the automatic transcription. There are two corresponding test sets. The other part is recorded with TEQST (section 4.2.2). `MINI-ANS.42m` has been processed similarly to `MINI-Quest-Ans.3.34h` and consists of 224 free answers on M.I.N.I questions with a duration of 38 minutes by one speaker and `MINI-Ques.50m` which contains 231 M.I.N.I questions by the same speaker as from `MINI-ANS.42m` with a duration of 43 minutes.

4.3 Accent Adaptation Data-Set

For accent adaptation we conducted experiments with the two following English data-sets:

[1] https://github.com/TEQST/TEQST

The first accent domain that we adapted the seq2seq model on is a Japanese accented English one. A four-hour audio, which was compiled by a working student in our Institute in 2010 was used as the training data for this domain. On the other hand, the test data was actually a recording of a one-hour-and-a-half English lecture delivered by the Japanese Professor Nakamura in April 2020. This audio copy was provided by the University of Tokyo, one of our partners who was in need of an English ASR system for transcribing their lectures. Those training and test data-sets were collectively called the `Japanese accent data-set`.

For the second accent domain, the condition was considerably different in which the data-set contained numerous audios of speeches in some international scientific conferences. It can be seen that those people were from different countries and they spoke English with dissimilar accents, consequently making it harder to adapt the model effectively. To be more specific, the training data was made up of 39 recordings of 39 presentations in the `EUROSPEECH 1993` Conference. Similarly, the test data was collected by obtaining speech recordings in the `InterACT25 2006` workshop. In the end, we had a two-hour test data-set of 24 speeches with a wide range of accents. These data-sets were called the `multi-accent data-set`.

4.4 Vocabulary Adaptation Data-Set

For vocabulary adaptation we recorded a German `new words test set` containing 32 minutes of speech. In this test set, words of the German `MINI-Questions training set` which are not recognized correctly with our baseline system (e.g. substance names) have been taken, put in other context and have been recorded. For recording of this new words test set the application TEQST (see section 4.2.2) was used.

5 Experiments and Results

For all experiments, we first trained a general ASR seq2seq model on the out-of-domain data-set, then adapted the model with the in-domain data-set. For the experiments in English topic adaptation and in German we used a Transformer based seq2seq model (Vaswani et al., 2017; Pham et al., 2019), for Arabic and English accent adaptation we used an encoder-decoder plus attention based system (Nguyen et al., 2020).

For the adaptation we compare fine-tuning and

batch-weighting (Wang et al., 2017). In the batch-weighting strategy the training data-set is a combination of in- and out-of-domain data. To describe how the data of both data-sets is combined we report the out-of-domain ratio, i.e., the number of tokens in a mini batch from the out-of-domain data-set divided by the total number of tokens in the mini batch. A ratio of 0 is equivalent to conventional fine-tuning using only in-domain data and 1 is equivalent to training without in-domain data. Furthermore, we combined these methods with freezing layers. We froze the encoder and all layers except the softmax-layer.

For all the experiments we report the time for the adaptation. Note that it is possible for methods with frozen layers to take longer compared to their counterparts with no frozen layers since different methods can require different amounts of update steps to obtain the best performance.

The tables 4, 5, 6 and 7 contain a summary of the experiments.

English For English the baseline model achieved a Word Error Rate (WER) of 11.0% on the out-of-domain test set, i.e., TED talks. On the out-of-domain test set, ATIS, the baseline system achieves a WER of 43.1%.

German The German baseline model yields WERs of 15.8% and 32.6% on the out-of-domain test set and on the in-domain test set, respectively. The baseline systems achieves an accuracy of 32.0% on the new words test set.

Arabic The Arabic baseline system achieves a WER of 12.6% on the out-of-domain data, Aljazeera shows, and 40.0% on MINI answers and 30.4% on MINI questions, our in-domain data.

Non-Native English The baseline system for our non-native tests achieved a WER of 7.8% on the out-of-domain test data, and 23.5% and 21.6% on the two accented in-domain data-sets.

5.1 Topic Adaptation

5.1.1 English

The results for the topic adaptation experiments on English are summarized in Table 4. The batch-weighting method with out-of-domain ration 0.3 and no frozen layers obtained 4.4% WER on the in-domain, which is a 12% relative improvement compared to the fine-tuning approach, and is only 0.3% worse on the out-of-domain data-set than the baseline. Compared with the 3.1% reduction of WER of the fine-tuning, the results show that adding an appropriate amount of out-of-domain data to the training data-set during adaptation can effectively reduce forgetting on the out-of-domain. Freezing the encoder or all layers except the softmax-layer performed worse on the in-domain data-set than without freezing layers.

5.1.2 German

For the language German we found that for both fine-tuning and batch-weighting the WERs on the out-of-domain test set decreased ((A-C) in table 5). This suggests that there may be a better point of stopping the baseline training. Batch-weighting was performed with an initial ratio of 0.5 and the distance of this ratio to zero and one was then split in half multiple times to obtain the other ratios used. The best full model, model with frozen encoder and model with all layers except the softmax-layer frozen achieved 21.8%, 25.6% and 27.8% WER on the out-of-domain data-set outperforming fine-tuning by 1.1%, 5.3% and 5.7%, respectively. As for the English system adapting the full model performed best and batch-weighting worked better than fine-tuning.

5.1.3 Arabic

For adapting the Arabic system to M.I.N.I questions and answers, we employed batch-weighting by increasing the out-of-domain ratio from 0.05 to 0.95 with a step of 0.05. The validation set is a mixed set from in- and out-domain data. As shown in table 6, the system succeeds to adapt to the target domain without forgetting by training the full model, by freezing the encoder with the ratios 0.2 and 0.4 respectively. The model suffers from a slight forgetting (0.8%) when training by freezing all layers except the softmax-layer. As we notice from table 6 the best adapting results (11.4% for `MINI-ANS.42m` and 3.8% for `MINI-Ques.50m`) are reached when training the whole model inclusive the encoder. The reason could be referred to the channel difference of the recording with mobile platforms from the out-domain training data (see sections 4.1.3 and 4.2.3).

Description	Out-of-domain ratio	Ted (out)	Atis (in)	Time
Baseline	-	**11.0**	43.1	-
A: Full model				
Fine-tuning	0.0	14.1	5.0	12 min
Batch-weighting	0.3	11.3	**4.4**	63 min
B: Frozen encoder				
Fine-tuning	0.0	14.0	4.8	**12 min**
Batch-weighting	0.5	11.4	4.8	123 min
Batch-weighting	0.3	11.5	4.9	44 min
C: All layers except softmax-layer frozen				
Fine-tuning	0.0	20.5	5.3	38 min
Batch-weighting	0.5	11.4	5.5	53 min
Batch-weighting	0.3	11.5	5.5	65 min

Table 4: Summary of the experiments for the English ASR-System domain adaptation (values are the WER ↓)

Description	Out-of-domain ratio	Test set	MINI-Q. test set	New words test set acc. (↑)	Time
Baseline	-	15.8	32.6	32.0	-
A: Full model					
Fine-tuning	0.00	15.8	22.0	52.3	50 min
Batch-weighting	0.88	**14.9**	22.2	54.3	29 min
Batch-weighting	0.97	15.4	**21.8**	58.4	40 min
B: Frozen encoder					
Fine-tuning	0.00	15.1	27.0	42.1	35 min
Batch-weighting	0.75	15.1	26.4	43.2	107 min
Batch-weighting	0.94	15.4	25.6	49.2	34 min
C: All layers except softmax-layer frozen					
Fine-tuning	0.00	15.1	29.5	36.6	59 min
Batch-weighting	0.94	15.0	28.0	53.0	142 min
Batch-weighting	0.97	15.0	27.8	53.0	95 min
D: All layers except softmax-layer frozen with feature caching					
Batch-weighting	0.88	15.2	28.5	57.4	**10 s**
E: Feature caching + 1-layer LM on top of the decoder					
Batch-weighting	0.50	16.9	27.2	**70.1**	86 min

Table 5: Summary of the experiments for the German ASR-System (values are the WER ↓)

Description	Out-of-domain ratio	Alj-MSA.2h	MINI-ANS.42m	MINI-Ques.50m	Time
Baseline	-	12.6	40.0	30.4	-
A: Full model					
fine-tuning	0	17.3	14.1	5.3	7 min
Batch-weighting	0.2	**12.6**	**11.4**	**3.8**	48 min
B: Frozen encoder					
fine-tuning	0	21.2	25.3	8.6	19 min
Batch-weighting	0.4	12.7	25.8	6.1	27 min
C: All layers except softmax-layer frozen					
fine-tuning	0	13.9	30.0	19.6	18 min
Batch-weighting	0.2	13.4	26.6	11.0	68 min

Table 6: Summary of the experiments for the Arabic ASR-System (values are the WER ↓)

Description	Out-of-domain ratio	Test set	Japanese accent test set	Multi-accent test set	Time
Baseline	-	7.8	23.5	21.6	-
A: Full model					
Fine-tuning	0.0	7.3	**18.8**	22.5	183 min
Batch-weighting	0.5	7.2	18.9	**20.2**	351 min
B: Frozen encoder					
Fine-tuning	0.0	8.1	24.6	-	-
C: Frozen decoder					
Fine-tuning	0.0	7.3	**18.8**	22.7	149 min
Batch-weighting	0.5	7.2	**18.8**	**20.1**	291 min

Table 7: Summary of the experiments for the non-native English ASR-System adaptation (values are the WER ↓)

5.2 Accent Adaptation

In addition to the adaptation strategies described at the start of the chapter, we inspected the efficiency of the fine-tuning process with frozen decoder. We expected this method to be significantly better than fine-tuning with frozen encoder, and as effective as fine-tuning the whole model due to the adaptation on the acoustic domain.

For the `Japanese accent test set`, the batch-weighting method with frozen decoder produced the best WER 18.8%. As can be seen from Table 7, the frozen decoder had an equally effective performance in comparison with fine-tuning the whole model and worked a lot better than fine-tuning with frozen encoder. Moreover, the result of frozen encoder was proved to be worse than the baseline model. Interestingly, the result on the out-of-domain test set was even better after applying fine-tuning on the in-domain data, which exceeded our original expectation. Therefore, we assumed that fine-tuning on the harder acoustic domain could improve the general performance of the encoder component. On the other hand, the result could not be noticeably improved with the batch-weighting.

For the `multi-accent test set`, the results did not show that the normal fine-tuning could work as well as the Japanese accent one. However, it can be observed that batch-weighting of the whole model could improve WER from 21.6 to 20.2.

Finally, batch-weighting with frozen decoder produced the best results on both Japanese accent and multi-accent domains.

5.3 Vocabulary Adaptation

For vocabulary adaption we did not only measure WER but also the word accuracy (WA) on the new words as described in Section 3.3. The baseline model achieved a WA of 32.0% on the new words test set. This is rather high for words the baseline model did not recognize correctly in other context since in the MINI-Questions training set the are a lot of enumerations, e.g., of substance names. Putting these new words in separate sentences makes it easier for the model to recognize them. The best full model, model with frozen encoder and model with all layers except the softmax-layer frozen achieved 58.4%, 49.2% and 52.8% accuracy (A-C in table 5), i.e., significantly better than the baseline.

In a scenario where the adaptation has to be done within a very short time, e.g., during a lecture where the system should adapt to human corrections within a few seconds, it is possible to use the approach of freezing all layers except the softmax-layer. This allows to speed up the adaptation process by caching the output of the decoder before the softmax-layer. These features can then be used to train the softmax-layer. This is faster because it is only required to process the speech and text once by the encoder and decoder, respectively, and this can be done in a precomputation step.

We tried to cache the features after the decoder with the model in training and inference mode, respectively. The second one performed better and also better than when training without feature caching ((D) in table 5). Since the validation loss increased constantly during training when using inference mode features we chose the validation accuracy (which increased up to some point) to determine the point to stop the training. Using this technique reduces the time requirements significantly.

We also tried to extend the model by adding a language model on top of the decoder. We tested language models with one and two layers. The 1-layer language model outperformed all other approaches tested with 70.1% accuracy on the new words test set and is reported in table 5 (E).

6 Conclusion

In this paper we examined the supervised adaptation of end-to-end speech recognition systems on small amounts of adaptation data when large amounts of general, out-of-domain training data are available. We used a technique called batch-weighting and contrasted it against regular fine-tuning, showing that batch-weighting delivers consistently better performance.

For this we performed experiments on several dimensions of domain adaptation: Topic, accent and vocabulary. We also performed experiments on three languages — Arabic, English and German — to show that batch-weighting generalizes across different languages and scenarios. For a rule of thumb to choose a good mixing ratio further experiments have to be conducted.

Due to its comparatively short run-time and computational resources necessary batch-weighting is suitable for life-long learning of an ASR systems during deployment, e.g., from user corrections.

7 Acknowledgement

The projects on which this paper is based were funded by the Federal Ministry of Education and Research (BMBF) of Germany under the numbers 01IS18040A and 01EF1803B. The authors are responsible for the content of this publication. We would also like to express our sincere gratitude to Professor Yoshi Nakamura from the University of Tokyo, who had delivered his lecture and made the Japanese accent test data-set available for our research.

References

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 279–284. IEEE.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018.

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 40–46.

Chenhui Chu and Raj Dabre. 2019. Multilingual multi-domain adaptation approaches for neural machine translation. *arXiv preprint arXiv:1906.07978*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *arXiv preprint arXiv:1701.03214*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Zi-Yi Dou, Xinyi Wang, Junjie Hu, and Graham Neubig. 2019. Domain differential adaptation for neural machine translation. *arXiv preprint arXiv:1910.02555*.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

M. J. F. Gales, D. Pye, and P. C. Woodland. 1996. Variance compensation within the mllr framework for robust speech recognition and speaker adaptation. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1832–1835 vol.3.

M.J.F. Gales. 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12(2):75–98.

Shuhao Gu, Yang Feng, and Qun Liu. 2019. Improving domain adaptation translation with domain invariant and specific information. *arXiv preprint arXiv:1904.03879*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Juan Hussain, Oussama Zenkri, Sebastian Stüker, and Alex Waibel. 2020. Dactor: A data collection tool for the relater project. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6627–6632.

Catherine Kobus, Josep Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79.

Robert C. Moore and Will Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

Thai-Son Nguyen, Sebastian Stüker, Jan Niehues, and Alex Waibel. 2020. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7689–7693. IEEE.

Thai-Son Nguyen, Sebastian Stüker, and Alex Waibel. 2019. Toward cross-domain speech recognition with end-to-end models. In *In Proceedings of the Life Long Learning for Spoken Language Systems Workshop colocated with ASRU 2019*, Singapore.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *Proc. Interspeech 2019*, pages 66–70.

Puming Zhan and M. Westphal. 1997. Speaker normalization based on frequency warping. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042 vol.2.

Anthony Rousseau, Paul Deleglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 125–128.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347.

David V. Sheehan, Yves Lecrubier, K. Harnett Sheehan, Patricia Amorim, Juris Janavs, Emmanuelle Weiller, Thierry Hergueta, Roxy Baker, and Geoffrey C. Dunbar. 1998. The mini-international neuropsychiatric interview (m.i.n.i.): The development and validation of a structured diagnostic psychiatric interview for dsm-iv and icd-10. *The Journal of Clinical Psychiatry*, 59(20):22–33.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Shen Yan, Leonard Dahlmann, Pavel Petrushkov, Sanjika Hewavitharana, and Shahram Khadivi. 2019. Word-based domain adaptation for neural machine translation. *arXiv preprint arXiv:1906.03129*.