

Data balancing for boosting performance of low-frequency classes in Spoken Language Understanding

Judith Gaspers*, Quynh Do*, Fabian Triefenbach

Amazon Alexa, Aachen, Germany

{gaspers, doquynh, triefen}@amazon.de

Abstract

Despite the fact that data imbalance is becoming more and more common in real-world Spoken Language Understanding (SLU) applications, it has not been studied extensively in the literature. To the best of our knowledge, this paper presents the first systematic study on handling data imbalance for SLU. In particular, we discuss the application of existing data balancing techniques for SLU and propose a multi-task SLU model for intent classification and slot filling. Aiming to avoid over-fitting, in our model methods for data balancing are leveraged indirectly via an auxiliary task which makes use of a class-balanced batch generator and (possibly) synthetic data. Our results on a real-world dataset indicate that i) our proposed model can boost performance on low frequency intents significantly while avoiding a potential performance decrease on the head intents, ii) synthetic data are beneficial for bootstrapping new intents when realistic data are not available, but iii) once a certain amount of realistic data becomes available, using synthetic data in the auxiliary task only yields better performance than adding them to the primary task training data, and iv) in a joint training scenario, balancing the intent distribution individually improves not only intent classification but also slot filling performance.

1. Introduction

Spoken Language Understanding (SLU) models are essential components in voice-controlled devices, such as Amazon Alexa, Siri and Google Assistant. Typically, SLU addresses the two sub-tasks of intent classification (IC) and slot filling (SF). While the former identifies a speaker’s intent, the latter extracts semantic constituents from the natural language query. For example, given the user utterance “play music by volbeat”, the IC subtask should identify *PlayMusic* as intent, while the SF subtask should detect “volbeat” as *Artist*. Recently, mostly DNNs are explored for SLU, which model IC and SF jointly to leverage the interaction between the two subtasks (e.g. [1], [2], [3]).

Academic research on SLU has mostly focused on improving overall accuracy on datasets with a static distribution and a fixed set of intents. By contrast, in real-world SLU applications, new intents are continuously added over time, yielding dynamically changing and highly imbalanced data distributions. This issue, which is known as *data imbalance* in machine learning [4], usually leads to poor predictive performance on minority classes. However, for an SLU application, it is important to support all intents, because the set of supported functionalities is announced to the customers, and there is no direct correspondence between the frequency with which a certain functionality is used and its importance. For example, an utterance “where is my phone?” is likely rarely used, but may be still important to the customer.

Notably though, in Academic SLU research the problem of low performance for low-frequency intents is typically hidden, as overall IC accuracy is measured, which is governed by performance of the head intents, as the percentage of test samples for tail classes is low.

In industry research, recently the data sparseness problem for new classes has been explored. Corresponding approaches investigated in particular data augmentation with synthetic data, such as machine translated (MT) data [5] or data collected via paraphrase generation techniques [6]. However, typically, the focus has been on how synthetic data can be generated, and sparse classes were simply balanced by adding synthetic data without taking overall data distributions into account.

To the best of our knowledge, this paper provides the *first* systematic study on handling data imbalance for SLU. In particular, we study the impact of data imbalance and evaluate different data balancing approaches for SLU. Our task has the following *specific challenges*: i) the class distribution used by customers during application is unknown during model training time, because it is evolving over time as new features are added and user behaviour changes, and ii) our aim is to boost performance of low-frequency intents, while not decreasing accuracy for the head intents. The latter is important, as the head intents are frequently used by customers, and thus a decrease in accuracy could risk a drop in customer satisfaction with the device.

While there is a large body of work addressing the data imbalance problem, it has not been studied sufficiently yet for modern DNN-based models and several findings are not consistent across tasks [4], indicating that different methods may be beneficial for different tasks. In addition, common data balancing methods like random over-sampling or weighted loss functions (e.g. [7], [8], [4]) may cause over-fitting, potentially leading to a performance drop of the head intents on unseen data during application.

To overcome the over-fitting problem, inspired by Zhang et al. (2019) [9], we use a multi-task framework including a primary task and an auxiliary task sharing a common feature extraction component. Although the two tasks are trained alternately in a multi-task setting, only the primary decoders are used during the inference phase. Thus, if we add extra information to the auxiliary task, it will be indirectly injected to the primary task through the feature extraction. Intuitively, this can help us to prevent the primary model from over-fitting. In this work, we apply two data imbalance handling techniques, i.e. training with a class-balanced batch generator and data augmentation with synthetic data to the auxiliary task.

We present empirical results on a real-world imbalanced SLU dataset. In particular, we compare our proposed approach with applying standard techniques including over-sampling, data augmentation with synthetic data, a re-weighting scheme with a DNN loss, and a class-balanced batch generator, directly on the primary task. Our results indicate that:

* The first two authors contributed equally to this paper.

1. our proposed model can boost performance on low frequency intents significantly while avoiding a performance decrease on the head intents which is a potential issue faced by common data re-sampling and re-weighting methods,
2. synthetic data are useful for intent bootstrapping, but
3. once a certain amount of realistic intent data becomes available, using synthetic data in the auxiliary task only yields better performance than adding them to the primary task training data, and
4. in a joint training scenario, balancing the intent distribution individually improves both IC and SF performance.

2. Related Work

We are not aware of previous work addressing the data imbalance problem in a systematic manner for SLU. However, there is a large body of work addressing this topic in other fields, particularly in image processing (e.g. [8], [4]). The most common approaches to data balancing include different re-sampling and re-weighting strategies. In random over-sampling, samples for the minority classes are duplicated [8]. While this has been shown to be quite effective, it increases model training times and has been found to cause over-fitting [10]. In addition, synthetic over-sampling of minority classes (SMOTE [7]) as well as random under-sampling of majority classes [11] has been explored. Weighted loss functions can be applied to change the impact of (certain) classes during optimization. A common strategy for re-weighting a loss function is using the (smoothed) inverse of class frequencies (e.g. [12], [13]). Moreover, leveraging information about the class distribution on a reference dataset has been explored, for instance, dynamic sampling [14] and a label-distribution-aware margin loss [15] have been explored. However, a challenge in our task is that the class distribution during application is unknown during model training and hence cannot be leveraged.

Another related line of research explores boosting performance of low-frequency classes via knowledge transfer from head to tail classes. (e.g. [16], [17]).

Zhang et al. (2019) explore data balancing for image processing tasks by making use of an auxiliary task which combines a class-balanced and a random batch generator [9]. We have adapted this idea for our SLU task, e.g. towards handling joint tasks, to use different batch generators for primary and auxiliary tasks, and we additionally study the integration of synthetic data.

Different methods have been proposed for generating synthetic data to boost performance of new languages or features in SLU via data augmentation (e.g. [5], [6]). While we include synthetic data into our study, our focus is not on data generation.

3. Method

For the SLU task, we use a joint model for IC and SF based on BERT [18]. We develop a multi-task framework consisting of two of these SLU tasks i.e. a primary and an auxiliary SLU task sharing a common feature extraction. While the two tasks are trained jointly, only the primary task decoders are used for the inference phase. The *auxiliary task* is trained using a *class-balanced batch generator* (CBG) and (optionally) *data augmentation with synthetic data*.

3.1. SLU Model

Figure 1 shows our multi-task model for handling data imbalance in SLU. Our model deals with two tasks:

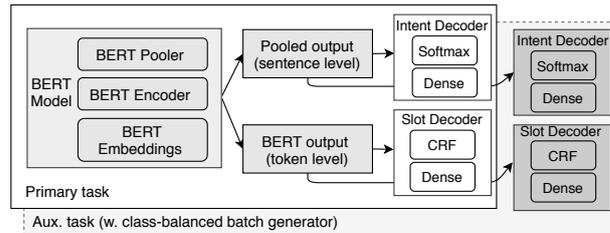


Figure 1: A multi-task model for handling data imbalance with a class balanced batch generator.

- Primary task: Standard SLU, which is a joint model of IC and SF using a regular random batch generator.
- Auxiliary task: SLU using a class balanced batch generator, which has the same model architecture as the primary task, but is trained with a special batch generator to assure class balance in each training batch. Moreover, synthetic data may be used during training to balance low-frequency classes with synthetic samples.

The SLU model consists of a BERT [18] encoder, an intent decoder and a slot decoder. As shown in Figure 1, the BERT encoder’s outputs at sentence level and token level are used as inputs for the intent and slot decoders, respectively. The intent decoder is a standard feed-forward network including two standard dense layers and a softmax layer on top. The slot decoder uses a CRF layer on top of two dense layers to leverage the sequential information of slot labels. The two losses of IC and SF are optimized jointly with balanced weights (1.0:1.0).

To perform multi-task training, we alternate through tasks using a ratio of 1.0:1.0 between primary task batches and auxiliary task batches, respectively, during the training process. The training process is completed when both of the tasks are done.

3.2. Class-Balanced Batch Generator For A Joint Model

Algorithm 1: CBG

Input: \mathcal{D} - training data set, \mathcal{C} - set of classes (labels),
and b - batch size

```

size=0;
size_per_class=b/|C|;
while size < |D| do
  new_batch = ∅
  for c in C do
    S = { size_per_class instances of class c selected
          randomly from D }
    new_batch = S ∪ new_batch
  end
  size=size+|new_batch|
  train model using new_batch
end

```

Algorithm 1 shows our class-balanced batch generator. Each batch contains the same number of instances per class. To avoid training time explosion on large-scale datasets, we stop an epoch when the total number of the generated instances in the epoch exceeds the original training size.

We have to deal with two data distributions due to having two sub-tasks. Since slot labels are often shared across several intents, we assume that there is a strong correlation between the distributions so that when the IC distribution is balanced, the

Dataset	Snooze	Browse	Cancel	Head I.
Train	40	16	–	166,493
Val.	10	4	–	20,846
Test	449	634	623	20,710
NMT data	20,285	7,855	2,657	–

Table 1: Number of samples for three simulated low-frequency intents and the remaining seven intents (head intents).

slot distribution is balanced accordingly. Therefore, our class-balanced batch generator is performed *on the IC distribution only*.

3.3. Data Augmentation With Synthetic Data

Due to language expansion of voice-controlled devices, a large amount of annotated SLU data may be available in different languages, in particular in English. These data can be leveraged for SLU model development in another language (e.g. [5], [2], [19]). In this work, following Gaspers et al. (2018) [5], we automatically translate English source data into our target language, and use the translated data as extra training data in the auxiliary task. Thus, in contrast to previous work, our approach supports utilizing synthetic data for re-balancing rather than just directly for model training. This can help us to avoid the negative impact of low quality synthetic data on the primary task.

4. Experiments

In this section, we describe our experiments comparing different data balancing techniques on a real-world German SLU dataset.

4.1. Datasets

Realistic SLU data We extracted a random data sample from a commercial German SLU system; the data are representative of requests to voice-controlled devices and were manually annotated with intent and slot labels (see Table 1). The sample is from the *Notifications* domain and comprises 351,066 samples. We split the dataset into 80% train, 10% validation and 10% test data. The data are highly imbalanced for intent classes with a long-tail distribution, where the smallest classes comprise a single sample only. Due to the small amount of test instances for low-frequency intents, we cannot reliably evaluate performance for them. Therefore, we created an artificial set up to study intent bootstrapping in a systematic way. In particular, we first removed data of all intents having fewer than 1,000 samples, leaving 10 classes. This ensures that at least 100 test samples are available per intent, which are needed to measure performance reliably in a large-scale set up. We selected the three intents with the lowest frequency from the remaining intents to simulate intent bootstrapping and filtered out all of their samples from the train and validation datasets. To simulate the growing data amounts per class in a developing system, where new features are added over time, we randomly sampled a different amount for each of the three intents and re-added 80% and 20% to the filtered train and validation datasets, respectively. We added 0, 20 and 50 samples for *CancelReminderIntent*, *BrowseReminderIntent* and *SnoozeNotificationIntent*, respectively.

Machine-translated data We translated data from an English NLU system into German using a transformer-based neural machine translation (NMT) system trained with Sockeye [20]. The NMT system was trained on 4,000 segments of internal data as well as 28,733,606 segments of publicly available data; slot labels were projected from the English source utterances to the

German translations using the alignment model *fast_align* [21]. The number of NMT-generated samples is 2,657, 7,855 and 20,285 for *CancelReminderIntent*, *BrowseReminderIntent* and *SnoozeNotificationIntent*, respectively. This data set is used for *data augmentation* of low-frequency classes in our experiments.

4.2. Models

We train and evaluate the following models on our SLU data:

- *Baseline*: The standard model without any data imbalance handling technique; this baseline is obtained by training the primary task individually on SLU training data.
- *Over-sampling*: Common random over-sampling method; for each intent class, SLU training data are up-sampled to the number of samples in the head class. The primary task is then trained individually on the up-sampled data.
- *Balanced-loss*: Common re-weighting method with DNN loss; The cross entropy loss of IC is balanced by using class frequencies. The model is obtained by training the primary task individually on SLU training data.
- *CBG*: The primary task is trained individually on SLU training data using the class-balanced batch generator.
- *Mul.-CBG*: Proposed method *without* data augmentation; an auxiliary task with class-balanced batch generator (without using machine translated data) is applied.
- *Data-aug.*: Common training data augmentation method; SLU training data are augmented with the synthetic NMT data for low-frequency classes. The primary task model is then trained individually on the augmented data.
- *Data-aug.+Over-sampl.*: Combination of the above *data-aug.* and *over-sampling* approaches. SLU training data are first augmented by adding all of the available synthetic NMT data for low-frequency classes and subsequently up-sampled as in *over-sampling*. The model is then obtained by training the primary task individually on the up-sampled and augmented training set.
- *Data-aug.+Balanced-loss*: Combination of the above *data-augmentation* and *balanced-loss* approaches. SLU training data is augmented by adding the available NMT data for low-frequency classes. The model is then obtained by training the primary task individually on augmented training data with balanced loss.
- *Data-aug.+CBG*: Combination of the above *data-augmentation* and *CBG* approaches. SLU training data is augmented with the available NMT data. The primary task model is then trained individually on the augmented SLU training set using a class-balanced batch generator.
- *Mul.-CBG+Data-aug.*: Proposed method *with* data augmentation; the auxiliary task with class-balanced batch generator is applied on the NMT-augmented dataset.

4.3. Settings

In our experiments, we use pre-trained multilingual BERT [18] (size 768), and max-pooling for sentence representation. Each of our decoders has 2 dense layers of size 768 with gelu activation. The dropout values used in IC and SF decoders are 0.5 and 0.2, respectively. For optimization, we use Adam optimizer with learning rate 0.1 and a Noam learning rate scheduler. We trained our model with batch size of 64.

Following Gaspers et al. (2018) [5], we use a semantic error rate, which measures IC and SF jointly and is defined as follows:

$$SemER = \frac{\#(\text{slot+intent errors})}{\#\text{slots in reference} + 1} \quad (1)$$

Method	Snooze	Browse	Cancel	Head I.
<i>Over-sampling</i>	-30.63	-56.03	-16.77	+7.03
<i>Balanced-loss</i>	-32.45	-40.37	-16.77	+11.43
<i>CBG</i>	-40.56	-56.77	-8.10	+8.57
<i>Mul.-CBG (ours)</i>	-3.59	-45.77	-9.93	-1.10
<i>Data-aug.</i>	+236.97	-32.66	-43.83	-0.22
<i>Data-aug.+Over-sampl.</i>	-33.33	-37.16	-44.98	+3.08
<i>Data-aug.+Balanced-loss</i>	-57.67	-32.98	-49.42	+3.96
<i>Data-aug.+CBG</i>	-64.85	-37.25	-45.09	+19.56
<i>Mul.-CBG+Data-aug. (ours)</i>	-29.74	-34.21	-2.63	-0.44

Table 2: Relative change in semantic error rate (%) on three low-frequency intents and on the head intents for data balancing methods on real-world SLU data compared to the baseline. A decrease in *SemER* implies better performance.

5. Results

Table 2 shows the performances of our experimental models on the head intents, and on each of the three low-frequency intents.

5.1. Performance without data augmentation

Performance for all of the low-frequency intents can be improved just by balancing the regular training data. However, for *CBG*, random *over-sampling* and *balanced-loss* methods, the boost in performance comes at the cost of a relative increase in *SemER* for the head intents of at least 7.03%. As head intents are much more frequently used by customers, and the customers are used to their comparatively high performance already, such a decrease in performance is usually not acceptable. By contrast, our approach with the class-balanced batch generator boosts performance for all low-frequency features, yielding up to 45.77% relative reduction in *SemER* per intent, without over-fitting on the low-frequency classes. Thus, our results suggest that leveraging data balancing techniques in the auxiliary rather than the main task is beneficial. This may be the case, because the model keeps the access to important information about the intent distribution, which is lost when data balancing is applied in the primary task.

5.2. Performance with data augmentation

As expected, without using any other techniques, performance is improved when simply adding NMT data (*data-aug.*) to *CancelReminderIntent*, which didn't have any intent samples beforehand. However, for *SnoozeNotificationIntent*, which already had 50 intent samples, simply adding NMT data (*data-aug.*) yields 236.97% increase in *SemER*. This highlights the fact that in an evolving system one needs to be careful with maintaining synthetic data. They can be very useful to bootstrap new features when no or few realistic feature data are available. However, as they are typically of comparatively low quality, they can decrease performance when they are kept while more and more realistic data become available. While random *over-sampling*, *CBG* and *balanced-loss* can mitigate this problem and improve performance on low-frequency classes, this again comes at the cost of a decrease in performance for the head intents (cf. *data-aug.+over-sampl.*, *data-aug.+CBG* and *data-aug.+balanced-loss*). *Multi-task-CBG+Data-aug.* indicates performance when NMT data are not added to the primary task training data, but used in the auxiliary task class-balanced batch generator during training only. For *SnoozeNotificationIntent* and *BrowseReminderIntent*, which already have a small number of samples in the realistic training data, performance is improved compared to adding NMT data to the primary task training data without overfitting on the head classes. For *SnoozeNotificationIntent* performance for *Mul.-CBG+Data-aug.* is better than both *Mul.-CBG* without

using data augmentation and *data-aug.* This suggests that once a certain amount of realistic intent data becomes available, it's beneficial to leverage NMT data only with *Mul.-CBG+Data-aug.* However, for intents without realistic training data, standard NMT data augmentation is preferable, as the intent cannot be recognized otherwise. Note that this is the case for *CancelReminderIntent* in this experiment.

5.3. Does balancing the intent distribution help slot filling?

Interestingly, the intent-based data balancing methods also improve SF performance, as indicated by the results for *CancelReminderIntent* by several methods without including data augmentation, in particular *Mul.-CBG*. Recall that there are no samples for this intent in the regular training data. Hence, the intent cannot be recognized, adding one error to *SemER* for each test utterance evaluation independent of the method (not applying data augmentation). In fact, the improvements for *CancelReminderIntent* are solely resulting from improved SF performance. We attribute this improvement to the fact that slots are shared across several intents and may benefit from intent-based data balancing, as we are using a multi-task model for IC and SF.

5.4. Performance on benchmark data

Unlike previous approaches to SLU, our method is focused on improving low-frequency classes which have typically very few samples in small-scale benchmark datasets and thus improvements in the overall standard SLU error metrics IC accuracy and slot F1 might not be expected. However, to investigate whether using an auxiliary task with a class-balanced batch generator improves upon a SOTA SLU model on a benchmark dataset we evaluated our approach on the small-scale ATIS dataset [22] using the version provided by [23]. In particular, we trained the baseline and *Mul.-CBG* models, and we found that *Mul.-CBG* improves over the baseline on IC accuracy and slot F1 from 97.4% to 97.5% and from 95.7% to 96.2%, respectively. This indicates that our method can even increase overall performance on small-scale benchmark data, and that our model yields comparable performance to state-of-the-art SLU systems [3, 2, 23]. In addition, the results provide further evidence that balancing the intent distribution also improves SF, i.e. from 95.7% to 96.2% in F1, where the latter even slightly outperforms the previously best reported results of 96.1% for SF on ATIS [3].

6. Conclusion

We presented a study comparing different techniques for handling data imbalance for DNN-based SLU. Aiming to boost performance for low-frequency intents, we proposed a multi-task model for SLU in which we make use of an auxiliary task to deal with data imbalance. Our results on a real-world SLU dataset indicate that: i) in contrast to common data re-sampling and re-weighting methods, our method can boost performance on low frequency intents significantly without decreasing performance of the head intents, ii) synthetic data are beneficial for bootstrapping new intents when realistic intent data are not available, but iii) once a certain amount of realistic intent data becomes available, using synthetic data in the auxiliary task only yields better performance than adding them to the primary task, and iv) in a joint training scenario, balancing intent distribution individually improves not only intent classification but also slot filling performance. Overall, our method achieved relative error rate reductions of up to 45.77% for low-frequency intents.

7. References

- [1] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proceedings of INTERSPEECH*, 2016.
- [2] Q. N. T. Do and J. Gaspers, "Cross-lingual transfer learning for spoken language understanding," *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019. [Online]. Available: <http://arxiv.org/abs/1904.01825>
- [3] Q. Chen, Z. Zhuo, and W. Wang, "Bert for joint intent classification and slot filling," *arXiv:1902.10909*, 2019.
- [4] J. Leevy, T. Khoshgofaar, and R. Bauder, "A survey on addressing high-class imbalance in big data," *J Big Dat*, vol. 5 (42), 2018. [Online]. Available: <http://arxiv.org/abs/1907.11692>
- [5] J. Gaspers, P. Karanasou, and R. Chatterjee, "Selecting machine-translated data for quick bootstrapping of a natural language understanding system," *Proceedings of NAACL-HLT*, 2018.
- [6] E. Cho, H. Xie, and W. M. Campbell, "Paraphrase generation for semi-supervised learning in NLU," in *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 45–54. [Online]. Available: <https://www.aclweb.org/anthology/W19-2306>
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, 2002. [Online]. Available: <http://dx.doi.org/10.1613/jair.953>
- [8] H. Guo, Y. Li, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017.
- [9] J. Zhang, L. Liu, P. Wang, and C. Shen, "To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions," 2019.
- [10] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 1–6, Jun. 2004. [Online]. Available: <https://doi.org/10.1145/1007730.1007733>
- [11] M. A. Tahir, J. Kittler, K. Mikolajczyk, and F. Yan, "A multiple expert approach to the class imbalance problem using inverse random under sampling," in *Multiple Classifier Systems*, J. A. Benediktsson, J. Kittler, and F. Roli, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 82–91.
- [12] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [13] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," 2018, cite arxiv:1805.00932Comment: Technical report. [Online]. Available: <http://arxiv.org/abs/1805.00932>
- [14] S. Pouyanfar, Y. Tao, A. Mohan, H. Tian, A. S. Kaseb, K. Gauen, R. Dailey, S. Aghajanzadeh, Y. Lu, S. Chen, and M. Shyu, "Dynamic sampling in convolutional neural networks for imbalanced data classification," in *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 112–117.
- [15] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 1567–1578.
- [16] W. Ouyang, X. Wang, C. Zhang, and X. Yang, "Factors in finetuning deep model for object detection with long-tail distribution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2016.100>
- [17] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," 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. 7029–7039. [Online]. Available: <http://papers.nips.cc/paper/7278-learning-to-model-the-tail.pdf>
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [19] A. Johnson, P. Karanasou, J. Gaspers, and D. Klakow, "Cross-lingual transfer learning for Japanese named entity recognition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 182–189. [Online]. Available: <https://www.aclweb.org/anthology/N19-2023>
- [20] F. Hieber, T. Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post, "Sockeye: A Toolkit for Neural Machine Translation," *arXiv preprint arXiv:1712.05690*, Dec. 2017. [Online]. Available: <https://arxiv.org/abs/1712.05690>
- [21] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, Jun. 2013, pp. 644–648. [Online]. Available: <https://www.aclweb.org/anthology/N13-1073>
- [22] G. Tür, D. Hakkani-Tür, and L. P. Heck, "What is left to be understood in atis?" in *SLT*, D. Hakkani-Tür and M. Ostendorf, Eds. IEEE, 2010, pp. 19–24. [Online]. Available: <http://dblp.uni-trier.de/db/conf/slt/slt2010.html#TurHH10>
- [23] C.-W. Goo, G. Gao, Y.-K. Hsu, C.-L. Huo, T.-C. Chen, K.-W. Hsu, and Y.-N. Chen, "Slot-gated modeling for joint slot filling and intent prediction," in *NAACL-HLT*. Association for Computational Linguistics, 2018, pp. 753–757. [Online]. Available: <http://aclweb.org/anthology/N18-2118>