CSC 585: Homework 3

Andrew Zupon
Department of Linguistics
University of Arizona
zupon@email.arizona.edu

Abstract

In this paper, I present an method for using the Mean Teacher neural network architecture (Tarvainen and Valpola, 2017) to improve part-of-speech tagging in Scottish Gaelic, a low-resource language. I introduce noise during training in the form of tagged Irish data, helping the tagger generalize better over the limited Scottish Gaelic data.

1 Introduction

Part-of-speech tagging is often considered a solved task for well-resourced languages (Manning, 2011). The same cannot be said for low-resource languages. Low-resource languages often do not have any corpora available, and what is available is typically smaller than comparable English corpora. Part-of-speech taggers rely on large amounts of training data, which makes developing resources for low-resource languages difficult.

Scottish Gaelic is an endangered Celtic language spoken primarily in the Highlands and islands of northwestern Scotland. In addition to being endangered, Scottish Gaelic is morphologically complex[1]. The limited data and it's morphological complexity make Scottish Gaelic a particular challenge for natural language processing.

Automatic part-of-speech taggers for Scottish Gaelic have been developed by Lamb and Danso (2014) and Zupon (2017). Lamb and Danso (2014)'s best model uses a Brill bigram tagger, and Zupon (2017) develops an LSTM neural network tagger. In both cases, the paucity of Scottish Gaelic data, in conjunction with the language's complex morphology, lead to results still well below well-resourced languages.

To try to address the problem of limited data, this paper develops a part-of-speech tagger using the Mean Teacher architecture (Tarvainen and Valpola, 2017). The Mean Teacher model introduces noise into a student model and a teacher model, with the teacher model updating its model weights based on the exponential moving average model weights of the student model. The benefits of this architecture include being able to generalize better, as the added noise "push[es] decision boundaries away from the labeled data points" (Tarvainen and Valpola, 2017, p. 1). One limitation of the original Mean Teacher model is that the authors only test their architecture on image recognition and not on language data.

The key difference between my model and the model presented in Tarvainen and Valpola (2017) is in how noise is added to the model. In their image recognition system, noise involved translations or flips of the input images, Gaussian noise on the input layer, and dropout within the network. For my part-of-speech tagger, the noise is tagged Irish data, added to the original Scottish Gaelic training data.

The model presented in this paper provides two novel contributions to part-of-speech tagging for low-resource languages. First is the adoption of Tarvainen and Valpola (2017)'s Mean Teacher method for part-of-speech tagging. As discussed above, Tarvainen and Valpola (2017) only implement their Mean Teacher method for image recognition. The second contribution is the use of a second language as noise added to training. Previous work on machine translation of low-resource languages has used a third language as an intermediate "pivot" between source language and target language (Firat et al., 2016; Chen et al., 2017), but in this paper, the additional language is incorporated directly into the original training data.

One downside of the Mean Teacher model pre-

[1]The ARCOSG tagset for Scottish Gaelic (Lamb and Danso, 2014) includes 242 unique tags. The Penn Treebank tagset commonly used for English has only 36 unique tags.

sented here is that it relies on having sufficient data in a closely-related second language (Irish) to add to the model—a situation that many low-resource languages are not in. Despite this, the model presented here is applicable to Scottish Gaelic and various other low-resource languages and language pairs, even if it does not generalize to every language.

2 Related Work

In this section I discuss previous work on part-of-speech tagging for Scottish Gaelic, along with previous computational work on low-resource languages more generally.

Lamb and Danso (2014) develop the first automatic part-of-speech tagger for Scottish Gaelic. They achieve an accuracy 76.6% using a Brill bigram tagger. One major shortcoming of their work—and in fact all work on low-resource languages—is the sparsity of tagged data. Another issue was because their result is based on only a small piece of their Scottish Gaelic Corpus. The present paper tries to augment the limited Scottish Gaelic data by introducing tagged Irish data into training.

Zupon (2017) builds on Lamb and Danso (2014) by developing an LSTM neural network part-of-speech tagger for Scottish Gaelic, achieving an accuracy of 80.33% after 30 epochs. Despite this improvement, there are many shortcomings of Zupon's paper. In particular, the split of data between training and testing partitions was done semi-arbitrarily, without using k-fold cross-validation. In addition, no statistical significance is shown. Even with these flaws, this forms the baseline for the present paper.

Much work has been done on neural machine translation of low-resource languages. While the present paper focuses on part-of-speech tagging within one language, recent work treats part-of-speech tagging as a kind of translation (Moeller and Hulden, 2018). Instead of translating from source to target language, we "translate" from source tokens to part-of-speech tags.

Artetxe et al. (2017b) describe a system of neural machine translation that solely requires monolingual corpora. They use a shared encoder for both source and target language that uses unsupervised cross-lingual word embeddings (Artetxe et al., 2017a; Zhang et al., 2017). This shared encoder can then reconstruct the original input, allowing translation between the source and target language without the need for parallel corpora—a great help for low-resource languages where such corpora are few and far between.

Firat et al. (2016) and Chen et al. (2017) describe different approaches to "zero-resource" neural machine translation, where parallel corpora between source and target language are nonexistent. The former uses a many-to-one translation model, where the source language is first translated into an intermediate pivot language, and then both the original source text and the pivot translation are used to translate into the target language. The latter uses a student-teacher architecture for translating from source to target (the "student" model) with help from a pivot-to-target model (the "teacher" model).

The benefits of Artetxe et al. (2017b)'s approach is that it does not require an intermediary pivot language, unlike Firat et al. (2016)'s or Chen et al.'s approaches. The "zero-resource" models still require parallel corpora for the pivot-to-target translations. For the present paper, I adopt a compromising position. Adding Irish data introduces an additional language to the model, but in the implementation presented here, the added data is treated as a part of the Scottish Gaelic source corpus, not as a separate, intermediate corpus.

Currey et al. (2017) give a method for improving neural machine translation for low-resource languages by augmenting the training data of parallel corpora with a monolingual bitext from the target language. The bitext copies the text of the target language, so that each source sentence in the bitext is the same as the target sentence. This data is then added to the original parallel training data, with no modifications to the system otherwise. The authors show that simply adding this copied monolingual data for English↔Turkish and English↔Romanian language pairs yields improvements in accuracy. However, they also show that adding this copied data does not improve English↔German translation, a language pair with much more parallel data available than the other language pairs. While their work focused on translation as opposed to part-of-speech tagging, it does support the idea that data augmentation is a viable approach when working with low-resource languages.

Pivoting back to part-of-speech tagging, Buys and Botha (2016) describe a method of morpho-

logical tagging for low-resource languages similar to the machine translation methods described above. Their method involves projecting source-language part-of-speech tags to an aligned target-language text. This allows morphological tagging of a low-resource text in the absence of a part-of-speech tagger for that language, but it does rely on having sufficient parallel data to adequately align the source and target languages. As with the "zero resource" machine translation methods above, this has the shortcoming of relying on an intermediate pivot language.

My proposal in this paper will improve upon these previous works. Unlike the "zero-resource" machine translation models (Firat et al., 2016; Chen et al., 2017), my model does not require an additional translation step to go from source to target. Instead, the helper language (Irish) is implemented right into the source. My proposal also improves upon Artetxe et al. (2017b)'s unsupervised approach. Their model relies on having aligned bilingual word embeddings as a starting point. While their approach does appear very promising, the benefit of my proposal is that it does not require any pretrained embeddings; the Irish data is put straight into the Scottish Gaelic training data as is.

3 Discussion of Previous Baseline

This section discusses the baseline I compare against in this paper—Zupon (2017)'s LSTM part-of-speech tagger for Scottish Gaelic—and performs a modest error analysis of that work's shortcomings.

Zupon (2017)'s best model is an LSTM tagger run for 30 epochs. The LSTM was built using DyNet (Neubig et al., 2017), a neural network toolkit. The LSTM has 1 hidden layer with a size of 200, and an embedding dimension of 50. There were no pretrained word embeddings for Scottish Gaelic, so they were initialized randomly and updated during training. During training, the training data is shuffled and the sentences are batched into minibatches of size 256. The learning rate for training is 0.01. The results from rerunning Zupon (2017)'s code on `dev` and `test` partitions, along with the `test` results from the original paper, are given in Table 1.

Evaluation Partition	Overall Accuracy	Unknown Word Accuracy
`dev`	80.96	23.72
`test`	80.84	23.99
Zupon (2017)	80.33	22.84

Table 1: Results from rerunning Zupon (2017)'s code on both the `dev` and `test` data, along with results from the original paper on `test`. The original paper does not give results for `dev`.

While Zupon (2017)'s model does improve over Lamb and Danso (2014)'s work, it still lags behind part-of-speech taggers for other, well-resourced languages. A likely cause for this discrepancy is the morphological complexity of the language, the complexity of the tagset, and the limited availability of training data. These factors result in errors in three main areas.

First is the fact that individual words can appear with many different tags. *A* and *an*, the two most frequently incorrectly tagged words, appear in the training corpus (correctly) with a variety of different tags depending on the context. *A* alone shows up with `Ug`, `Dp3sm`, `Qq`, `Sp`, and various others.

Second is the fact that many of the tags in the tagset are extremely similar, often making it difficult to differentiate them. Most of the top ten labels that were not predicted by the model are variants of `Nc`, the tag for common nouns. The difference between many of the `Nc` tags comes down to gender marking (`m/f`) and case marking (`n/d/g`). The similarity of these tags makes it more likely that the system will make a small mistake, such as switching the gender or case marking, for Scottish Gaelic than it would for a language like English that does not mark gender or case on nouns.

Third is the size of the tagset in general. When looking at the top 10 tags predicted that were incorrect, one stands out over all the rest: `Sp`, the generic adposition/preposition tag, was predicted incorrectly 2,304 times, almost twice as many times as the next most common ncorrect tag. Considering the tagset, this is not terribly surprising. The ARCOSG tagset has 14 different tags for prepositions alone. Again, a language like English, with only one tag for prepositions (`IN`), or two if you count infinitival 'to' (`TO`), does not have this problem.

These three issues clearly relate to the complexity of the tagset and the limited available data for Scottish Gaelic. In the present paper, I aim to address these issues by adding Irish data. With the

Irish data added in training, the model may be able to generalize beyond the limited Scottish Gaelic training data and learn the more general patterns of the language.

4 Approach

My approach in this paper attempts to boost performance on a Scottish Gaelic part-of-speech tagger by augmenting the training data with data from Irish, a closely related language.

4.1 Changes from Previous Baseline

This approach makes some changes from the previous baseline as discused in the last section. While my previous baseline used the full ARCOSG tagset, here I use a shortened set of tags. Each tag is truncated to the first letter, which corresponds to the general category of the word (e.g. noun, verb), but loses the more specific information of the full tags (such as gender, number, and case information). This was done partly to alleviate the issues described above due to the size of the ARCOSG tagset and partly due to time constraints[2]. Another modification in my approach is simply the number of epochs used for training. The previous baseline used 30 epochs, but in rerunning the baseline model, accuracy levels off fairly quickly. Training to 30 epochs does not achieve much higher results than 20 or even 10, and so for that reason (and also time) my approach here is only trained for 10 epochs.

Another reason for these changes is due to the nature of my approach. Ultimately, I would like to compare my new model with the previous baseline, full tagset and all. However, this paper presents a test case for adding Irish data to the Scottish Gaelic training data. In this, the interesting result is how the models with Irish compare to models without Irish, rather than how they compare with the previous baseline. If adding Irish does show a positive effect, then it will be worthwhile to perform a more complete comparison with the previous baseline.

4.2 Description of Model Types

To compare how influential the added Irish data is, I trained four different model types with varying amounts of Irish and Scottish Gaelic data. The new baseline model includes only the original Scottish Gaelic training data, with no added Irish data. The Scottish Gaelic training data includes 46,794 words. The smallest Irish model adds 500 sentences from the New Corpus for Ireland[3] (Kilgarriff et al., 2006), which includes 12,000 words. This is equivalent to adding an additional 25.6% of the original training data. The medium Irish model adds 600 sentences from the Irish corpus, which includes 26,400 words. This is equivalent to adding 56.5% of the original training data. The largest Irish model adds 1000 sentences from the Irish corpus, which includes 44,000 words. This is equivalent to adding 94.0% of the original training data.

Finally, for each of these four model types I compare a student model and two teacher models. The student model uses weights from training w. The first teacher model simply uses an average of the last five student model weights as its weights $\tilde{w}$. The second teacher model uses an Exponential Moving Average (EMA) of the student model weights to generate its weights $\tilde{w}$. This is essentially the same as the consistency cost discussed by Tarvainen and Valpola (2017). With the EMA, the teacher weight at timestep t ($\tilde{w}_t$) comes from both the previous teacher weight ($\tilde{w}_{t-1}$) and the current student weight (w_t). α is a hyperparameter; in this paper $\alpha = 0.1$. The equation is given below in 1:

$$\tilde{w}_t = \alpha * \tilde{w}_{t-1} + (1 - \alpha) * w_t \quad (1)$$

4.3 Results

For each model type and each sub-model, I compare the Accuracy, Precision (macro), Recall (macro), and F1 (macro) for the student and two teacher models. The results are given in Table 2.

From Table 2, we see a couple of patterns. First, there generally seems to be little variation between the student model and the two teacher models. Overall and on unknown words, the model-type with no added Irish achieves the best Recall scores, 76.81% and 12.97% with the student model and weighted teacher model, respectively. The model-type with 600 Irish sentences perform the best on Accuracy, Precision, and F1, both overall and on unknown words. The student model scores highest on overall F1 (77.15%). The weighted teacher model scores highest on overall Accuracy (91.36%) and overall Precision

[2]Training with the full tagset takes six times longer than with the shortened tagset.

[3]http://corpas.focloir.ie/

MODEL		RESULTS							
		Overall				Unknown			
Model	# Irish	Acc	P	R	F1	Acc	P	R	F1
Student	0	90.24	81.36	**76.81**	77.06	62.33	9.70	12.96	9.45
Teacher-W	0	90.20	81.30	76.73	76.97	62.48	9.71	**12.97**	9.47
Teacher-EMA	0	90.26	77.91	75.89	76.37	64.10	9.97	12.87	9.71
Student	500	90.38	83.18	75.48	76.15	65.91	14.63	12.39	9.82
Teacher-W	500	90.41	83.22	75.39	76.10	66.18	14.65	12.37	9.84
Teacher-EMA	500	90.52	78.29	74.29	75.44	69.06	15.12	12.07	9.94
Student	600	91.34	83.90	75.79	**77.15**	76.71	15.38	9.44	10.00
Teacher-W	600	**91.36**	**84.02**	75.74	77.14	76.74	15.28	9.39	9.93
Teacher-EMA	600	91.32	78.83	75.00	76.53	**76.82**	**16.44**	10.08	**10.29**
Student	1000	89.98	76.77	75.83	75.97	66.60	12.06	11.92	9.34
Teacher-W	1000	90.17	77.11	75.77	76.15	67.73	12.10	11.90	9.42
Teacher-EMA	1000	90.33	77.64	75.25	76.18	69.57	11.68	11.11	9.06

Table 2: Results for each model type and sub-model, comparing student and two teacher models (weighted and EMA). # Irish is the number of Irish sentences added to the Scottish Gaelic training data. Results are given for Accuracy (Acc), Precision (P), Recall (R), and F1 overall and for unknown words. The best score for each evaluation metric is in bold.

(84.02%). The EMA teacher model scores highest on unknown word Accuracy (76.82%), unknown word Precision (16.44%), and unknown word F1 (10.29%).

A quick glance at the results in Table 2 shows that adding 500 sentences of Irish to the training data does improve results very slightly (except Recall), and adding 600 Irish sentences improves results even more (again except for Recall). However, the last three rows of the table show that adding 1000 Irish sentences generally appears to lower performance. The two apparent exceptions to this are unknown word Accuracy and Precision, which still get a boost from the added data.

At this point, no tests for statistical significance have been performed on the results, but this is a necessary next step for this project.

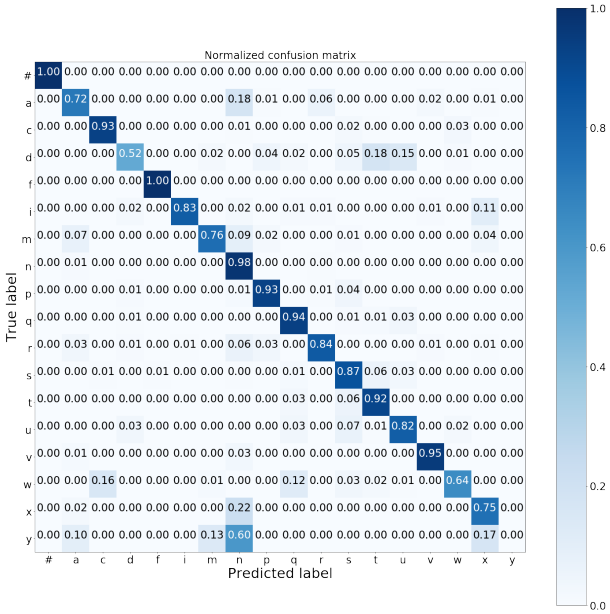

Figure 1: Normalized confusion matrix for weighted teacher model with 600 Irish sentences.

4.4 Error Analysis

As the results in Table 2 show, adding Irish data shows mixed results. Adding 500 sentences does not change results much, adding 600 sentences shows slight gains, but adding 1000 sentences yields a performance drop.

In this section, I will perform an error analysis of my approach. Figure 1 shows a normalized confusion matrix for the weighted teacher model using 600 Irish sentences, tested on the development set.

This matrix shows that for most tags, the system does a reasonable job. However, the performance on some tags is not so good. For example, the tags `a` (adjectives) and `d` (demonstrative and possessive determiners) are below 80% and 60%, respectively, and `y` (abbreviations) is at a whopping 0%. For `y`, this is not a mystery; it only shows up in training 188 times, making it by far the least common tag.

For `a` and `d` (and others), the reason for the

lower performance could have to do with the forward-only training. In Scottish Gaelic, like in English, determiners precede the noun. Unlike English, in Scottish Gaelic adjectives typically (but not always) follow the noun. These orders may contribute to the lower performance.

In order to address this problem, I test a second approach. In my second approach, I train the model both forwards and backwards.

By training in both directions, we can try to capture context on both sides of a word, rather than just the previous context.

5 Approach 2

To attempt to address some of the shortcomings shown in the confusion matrix in Figure 1, I test a second approach. This second approach uses the same basic architecture and settings as my original model, except now the model trains both forwards and backwards. This bidirectional training will ideally capture context on both sides of a word, rather than only remembering previous context.

5.1 Results

As the confusion matrix in Figure 2 shows, my attempt to train both forwards and backwards does not seem to improve the model.

Interestingly, results for `a` and `d` do go up, but results for many other parts of speech go down. Another interesting finding is that results for the tag `v` (verbs) goes up from 95% to 97%. This could be partly due to Scottish Gaelic word order. Unlike English, Scottish Gaelic is verb initial. In a forward-only model, there will be little previous context to help classify verbs. When backwards training is added, however, there will be forwards context to help with classification.

Accuracy, Precision, Recall, and F1 scores for my second approach are given in Table 3. As this table shows, bidirectional training in this case occasionally beats the forwards-only training, but not across the board.

6 Conclusions

This paper provides a test case for introducing data from another language into training. In particular, I looked at introducing Irish language data into training for a Scottish Gaelic part of speech tagger. Both languages are very closely related, so the hypothesis was that the Irish could augment the Scottish Gaelic data enough to make the system more robust.

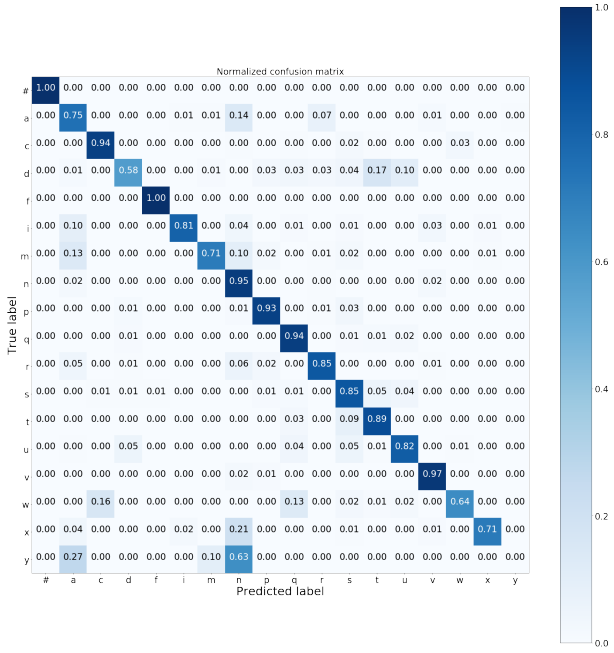

Figure 2: Normalized confusion matrix for weighted teacher model with 600 Irish sentences. Forwards and backwards training.

As my initial results show, adding Irish to training has varying effects. Adding only 500 sentences results in a small decrease in performance. Adding 600 sentences yields a small performance boost. Adding 1000 sentences brings performance back down. Of course, none of these patterns were shown to be statistically significant, so their value remains unclear.

To address some errors found in the original model, I also implemented a version of bidirectional training. The goal was to give context on both sides of a target word, rather than just the previous context. As the results in Table 3 showed, even though the classification of some tags improved (as shown in the confusion matrices in Figure 1 and Figure 2), overall the bidirectional training did not improve the system.

The approaches described here are small steps, and there are many clear areas for future work on this topic.

One next idea is to implement the other types of noise used in Tarvainen and Valpola (2017)'s Mean Teacher model. Their model, which is used for image classification, uses both dropout and Gaussian noise. The model presented in the current paper only uses the introduction of Irish data as noise.

MODEL		RESULTS							
		Overall				Unknown			
Model	# Irish	Acc	P	R	F1	Acc	P	R	F1
Teacher-W	600	**91.36**	**84.02**	75.74	77.14	**76.74**	**15.28**	9.39	**9.93**
Teacher-W-bi	600	91.16	82.61	**77.70**	**79.59**	73.00	12.77	**10.27**	9.74

Table 3: Results for forwards-only and bidirectional training with 600 Irish sentences. Weighted teacher models only. The bidirectional model beats the forwards-only model for overall Recall and F1 and for unknown word Recall. The forwards-only model beats the bidirectional model elsewhere.

Another area for future work would be to use a character-based LSTM instead of or in conjunction with the word-based LSTM used here. Because of the incredibly small amount of tagged Scottish Gaelic data to work with, and the lack of pretrained word embeddings, a character-level model might alleviate some performance issues due to unknown words encountered during testing.

In conclusion, adding data from another language may seem like a worthwhile endeavor, but depending on the nature of the two languages and the data available, it might be a better use of time to simply tag more data in the target language.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017a. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 451–462.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017b. Unsupervised neural machine translation. *CoRR*, abs/1710.11041.

Jan Buys and Jan A Botha. 2016. Cross-lingual morphological tagging for low-resource languages. *arXiv preprint arXiv:1606.04279*.

Yun Chen, Yang Liu, Yong Cheng, and Victor O. K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. *CoRR*, abs/1705.00753.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. *CoRR*, abs/1606.04164.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient corpus development for lexicography: building the New Corpus for Ireland. *Language resources and evaluation*, 40(2):127–152.

William Lamb and Samuel Danso. 2014. Developing an automatic part-of-speech tagger for Scottish Gaelic. In *Proceedings of the First Celtic Language Technology Workshop*, pages 1–5.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1195–1204. Curran Associates, Inc.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.

Andrew Zupon. 2017. Tagging Scottish Gaelic. Unpublished University of Arizona Manuscript.