Porjazovski, Dejan; Leinonen, Juho; Kurimo, Mikko

Attention-Based End-To-End Named Entity Recognition From Speech

Published in:
Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Proceedings

DOI:
10.1007/978-3-030-83527-9_40

Published: 01/01/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Porjazovski, D., Leinonen, J., & Kurimo, M. (2021). Attention-Based End-To-End Named Entity Recognition From Speech. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Proceedings* (pp. 469 - 480). (Lecture Notes in Computer Science; Vol. 12848). Springer. https://doi.org/10.1007/978-3-030-83527-9_40

Attention-Based End-To-End Named Entity Recognition From Speech

Dejan Porjazovski, Juho Leinonen, and Mikko Kurimo

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
{dejan.porjazovski,juho.leinonen,mikko.kurimo}@aalto.fi

Abstract. Named entities are heavily used in the field of spoken language understanding, which uses speech as an input. The standard way of doing named entity recognition from speech involves a pipeline of two systems, where first the automatic speech recognition system generates the transcripts, and then the named entity recognition system produces the named entity tags from the transcripts. In such cases, automatic speech recognition and named entity recognition systems are trained independently, resulting in the automatic speech recognition branch not being optimized for named entity recognition and vice versa. In this paper, we propose two attention-based approaches for extracting named entities from speech in an end-to-end manner, that show promising results. We compare both attention-based approaches on Finnish, Swedish, and English data sets, underlining their strengths and weaknesses.

Keywords: Named entity recognition, Automatic speech recognition, End-to-end, Encoder-decoder

1 Introduction

Named entity recognition (NER) is one of the main natural language processing (NLP) tasks. The goal of this task is to find entities and classify them into predefined categories. These categories can vary depending on the application area, but the most common ones include person, location, organization, and date.

Named entities are heavily used in spoken language understanding (SLU) [4] [16] [10], where the goal is to understand what has been spoken. For example, SLU is an essential part of personal assistants in home automation and smartphone devices. These personal assistants usually take speech as input, in which case the named entities need to be recognized from spoken data.

Doing NER from speech imposes several challenges for the system. There are far fewer annotated training data for spoken language than for textual data. The speech can be informal, not following the conventional syntax of the language, which can cause difficulties in detecting the entities. The generated transcripts from an automatic speech recognition (ASR) system usually do not contain capitalization and punctuation, which can cause the system to miss the entities.

The most common approach for doing named entity recognition from speech is through a pipeline approach. In this approach, the ASR system generates transcripts, and the NER system detects the entities in those transcripts. The output of the ASR

system is usually lower-cased and noisy, in the sense that the word order can be mixed, words might be missing or misspelled, etc. When developing a NER system for speech data, these factors need to be taken into account.

It is possible to try to restore the capitalization and the punctuation from the transcribed speech as explored in [7]. A maximum entropy model was used for NER on transcripts generated by a speech recognition system for Chinese, utilizing n-best lists [23]. These approaches improve the performance of the system on noisy speech data but they are still sensitive to the speech recognition output and error propagation. To deal with that, an end-to-end (E2E) approach was proposed that directly extracts named entities from French speech [6]. The authors used an architecture similar to the Deep Speech 2 [1], which was trained using the CTC algorithm [8]. A similar approach of E2E named entity recognition using the Deep Speech 2 architecture for the English language was explored in [22]. This is different from our proposed models, which use either attention-based encoder-decoder (AED) or a hybrid CTC/AED architecture.

In this paper, we propose two approaches for doing E2E NER from speech. To the best of our knowledge, this is the first attempt at NER using AED architecture in an E2E manner. The first approach is called augmented labels (AL) and it is either a standard AED or a hybrid CTC/AED architecture, where the transcripts are augmented with named entity tags during training. The second is a multi-task (MT) approach, where there are two decoder branches. One branch for doing automatic speech recognition and another one for doing named entity recognition.

2 Data

For the Finnish experiments, we used the Finnish parliament data set [15], consisting of about 1500 hours of recordings from the Finnish parliament. Since we do not have true named entity labels for this data set, we used a separate NER system to annotate it. The NER system is a bidirectional LSTM (BLSTM) neural network [9] with a Conditional random field (CRF) [12] layer on top, that utilizes morph, character and word embeddings. The architecture is explained in more detail in [18]. The number of tokens and named entity tags in the data set are presented in Table 1.

Table 1. Data distribution for the Finnish parliament data set.

Parameters	Count
Audio length	1500 h
Total tokens	7.3 M
Unique tokens	337423
PER tags	44984
LOC tags	73860
ORG tags	65463

For the Swedish experiments, we used the Sprakbanken corpus, which is a public domain corpus hosted by the National Library of Norway. It consists of 259 hours of

recordings. Since the corpus does not contain ground truth named entities, we used the Swedish BERT model [14] to obtain the annotations. The number of tokens and named entity tags are presented in Table 2.

Table 2. Data distribution for the Swedish data set.

Parameters	Count
Audio length	259 h
Total tokens	1.4 M
Unique tokens	69310
PER tags	23258
LOC tags	7585
ORG tags	2231

Even though the goal of this paper is mainly focused on low-resource languages like Finnish and Swedish, we additionally wanted to verify the performance of the models on a well-known language, like English.

For the English experiments, we used the whole LibriSpeech data set [17], consisting of about 1000 hours of recordings. The named entities for this data set were obtained using the large uncased BERT model [5], fine-tuned on the CoNLL 2003 data set [19], which we lower-cased before training. For testing the model with gold-standard named entity tags, we used a data set which is a subset of a combination of multiple speech recognition data sets, such as CommonVoice, LibriSpeech, and Voxforge. We will call this data set English-Gold. The data set is annotated and provided by [22]. The number of tokens and named entity tags in the English data sets are presented in Table 3.

Table 3. Data distribution for the English LibriSpeech and English-Gold data sets.

Parameters	LibriSpeech	English-Gold
Audio length	1000 h	148 h
Total tokens	9.6 M	1.3 M
Unique tokens	87600	41379
PER tags	194172	50552
LOC tags	66618	23976
ORG tags	11415	5025

3 Methods

To do E2E named entity recognition from spoken data, we will explore two approaches. In the first approach, we will build an attention-based encoder-decoder model for ASR by augmenting the labels with NER tags. In the second approach, we will explore multi-task learning where the model simultaneously learns to transcribe speech and annotate

it with named entity tags. Additionally, for the English and Swedish experiments, we utilize the CTC loss, as explored in [21].

Generally, the E2E ASR models can benefit from an external language model [20] but in our experiments we exclude it. The reason for that is because the augmented labels approach produces an output where each word is followed by a named entity tag. In such a case, adding an external language model trained on text will not benefit us. On the other hand, the baseline ASR models can benefit from an external language model but the goal of this paper is to explore an alternative way of doing named entity recognition from speech, as opposed to the standard pipeline approach.

3.1 Pipeline NER Systems

To see how our proposed models perform in comparison to the pipeline approach, where an ASR system generates the transcripts and then a NER system annotates them, we trained BLSTM-CRF models for each of the data sets. The architecture of these models is identical to the NER branch in the multi-task approach, described later in the paper. The models are trained on the original transcripts for each of the data sets. Since the English-Gold data set is small, we used the LibriSpeech model to initialize the weights and then fine-tune it on that particular data.

3.2 Baseline ASR System

The baseline ASR architecture is the same as the augmented labels approach, which is explained later in the paper. The only difference is that for the training of the baseline models, we used the original transcripts, whereas for the augmented labels approach we used the original transcripts augmented with named entity tags. We choose the architectures to be identical so that we can give a fair comparison between them.

3.3 Augmented Labels Approach

For this approach, we developed an attention-based encoder-decoder architecture that takes audio features as input and produces transcripts with named entity tags. Let $X = (x_1, x_2, ..., x_T)$ be the audio features, where each feature is represented as x_i and i is the order of the feature. Additionally, we define the output character set $Y = (y_1, y_2, ..., y_T)$, where y consists of all the characters plus the special tokens: <UNK>, <sos>, <eos>, O, PER, LOC, and ORG . The goal is to model the conditional probability:

$$P(Y|X) = \prod_i P(y_i|Y_{<i}, X) \tag{1}$$

In simpler terms, it predicts the i-th output character, given the previous characters and the input features X. It does this using an encoder and a decoder.

The encoder is a BLSTM neural network, that uses audio features as input and compresses them in a single hidden representation. This hidden representation is used to initialize the decoder.

The decoder is an LSTM neural network that takes the hidden vector, produced by the encoder and generates the transcripts using an attention mechanism. As an attention mechanism, we used Luong attention [13]. The scoring function for the attention is hybrid + location-aware, as described in [3]. It is defined as:

$$score(h_{enc}, h_{dec}) = v * tanh(W^e * h_{enc} + W^d * h_{dec} + W^c * conv + b) \quad (2)$$

where, h_{enc} and h_{dec} are the hidden states of the encoder and the decoder, $tanh$ is a hyperbolic tangent non-linearity, v and b are learnable weights, together with the W matrices. The location-aware element $conv$ is a convolution defined as:

$$conv = F * \alpha_t \quad (3)$$

where, F is a learnable matrix and α_t is the alignment vector.

For the experiments where we additionally used the CTC loss, the final ASR loss is calculated as:

$$L_{asr} = \lambda L_{ctc} + (1 - \lambda) L_{aed} \quad (4)$$

where, L_{ctc} is the CTC loss, L_{aed} is the decoder loss and λ is the weighting factor that determines the contribution of the separate loss functions to the final loss.

As true labels, we used the transcripts, augmented with named entity tags, in a way that each word is followed by its tag. This way, the model will jointly produce ASR transcripts and NER tags.

3.4 Multi-Task Approach

The multi-task approach is an attention-based encoder-decoder architecture, similar to the augmented labels approach. The difference between them is that this approach has two separate decoder branches. The first branch does the automatic speech recognition and is like the one in the augmented labels. The second one does the named entity tagging and it consists of BLSTM with a CRF layer on top. This approach uses hard parameter sharing, where the encoder is shared between both branches. Since it is a multi-task learning approach, we have two separate loss functions that need to be jointly optimized. The final loss function is calculated as:

$$L = \beta L_{asr} + (1 - \beta) L_{ner} \quad (5)$$

where L_{asr} is the loss from the ASR decoder, L_{ner} is the loss from the NER decoder, and β is a weighting factor that determines the contribution of both loss functions.

Similar to the augmented labels approach, in the experiments where we utilized the CTC loss, the ASR loss L_{asr} is calculated as in Equation 4.

4 Experiments

In all the experiments, we used logarithmic filter banks with 40 filters and Adam optimizer [11]. For the multi-task approach, after the models converged, we additionally

froze the encoder and the ASR decoder and trained only the NER branch, which improved the multi-task NER results on most of the data sets. We will refer to this model as MT*. The code was developed using Pytorch and is publicly available. [1]

Speech features consist of a large number of timesteps, so processing them using a standard BLSTM network is computationally expensive. To deal with that we used a pyramidal BLSTM network. The pyramidal structure reduces the computational time by concatenating every two consecutive timesteps in each layer.

In the Finnish and English experiments, the encoder consists of 5 pyramidal BLSTM layers, whereas in the Swedish experiments we used 3 normal and 2 pyramidal BLSTM layers. The reason for that is because the Swedish data set consists of short utterances, so there are not many timesteps to be processed. The hidden size of the BLSTM networks is 450 in all the experiments, except for the Finnish, where we used a hidden size of 300. After the last BLSTM layer, a dropout of 0.1 is applied.

In the augmented labels approach, the decoder consists of a character embedding layer with a size of 150 and a single layer LSTM network. For the English and Swedish experiments, the LSTM has a size of 450, whereas for the Finnish experiments, it has a size of 300. The location-aware element in the attention has 150 filters for the English and Swedish, and 100 filters for the Finnish experiments. A dropout of 0.1 is applied after the attention mechanism.

In the multi-task approach, the ASR decoder is identical to the one in the augmented labels, for all the experiments. The NER decoder uses pre-trained 300 dimensional fastText word embeddings [2] as an input to the one-layer BLSTM. The size of the BLSTM layer is 450 for the English and Swedish experiments, and 300 for the Finnish ones. The BLSTM is followed by a fully connected layer with the same size and a dropout layer with a probability of 0.1. In the end, the output is passed through a CRF layer that produces the tag probabilities.

Since the English-Gold data is relatively small with only 148 hours, we used the LibriSpeech data to pre-train the model and then fine-tune it on the English-Gold data set.

In all the experiments, we allocated data for testing, which was not used during training. As a loss function, we used the negative log-likelihood. For combining the ASR and NER losses, as in Equation 5, we used β weighting factor of 0.8. For the Swedish and English experiments, we additionally utilized the CTC loss, together with negative log-likelihood, like in the Equation 4, with a λ weighting factor of 0.2.

5 Results

In this section we present the results obtained on Finnish, Swedish, and English data sets, comparing both the augmented labels and multi-task approaches. For the evaluation of the ASR results, we used the word error rate (WER) metric, and for the evaluation of the named entity recognition results, we used the micro average F1 score.

[1] https://github.com/Tetrix/E2E-NER-for-spoken-Finnish

5.1 Finnish Results

In Table 4, we can see how both the augmented labels and multi-task approaches compare against the baseline ASR model in terms of WER when evaluated on the Finnish parliament data. From the results, we can notice that both approaches perform in pair with the baseline ASR model, falling slightly behind. We can also see that the multi-task approach performs slightly better than the augmented labels approach in terms of WER. In Table 5, we can see how both approaches perform in terms of precision, recall, and F1 score. Additionally, we evaluated our models on the original transcripts and on the transcripts that were generated by the models. We used the multi-task and the fine-tuned multi-task models to do the evaluation on the original transcripts. From the results, we can see that the fine-tuned multi-task model performs slightly better than the standard multi-task model. On the transcripts generated by the model, which is a harder task, we compared both multi-task approaches, along with the augmented labels and the pipeline approach. The ASR transcripts for the pipeline approach were generated using the multi-task model, for all the data sets. From the results, we can see that the fine-tuned multi-task approach achieved the best F1 score. We can also notice that both multi-task approaches perform better than the pipeline approach, whereas the augmented labels approach falls behind.

Table 4. WER on the Finnish test set.

Model	WER
Baseline ASR	**34.95**
AL	36.06
MT	35.80

Table 5. Precision, recall and F1 score for the Finnish test set.

Transcripts	**Model**	**Prec**	**Rec**	**F1**
Original	MT	93.70	92.88	93.29
	MT*	93.75	93.69	**93.72**
Generated	Pipeline	93.63	85.64	89.46
	AL	92.65	81.61	86.78
	MT	93.35	87.80	90.49
	MT*	93.17	88.80	**90.93**

5.2 Swedish Results

Next, we present the Swedish results. In Table 6, we can see how both approaches perform in terms of WER, in comparison to the baseline model. Similar to the Finnish experiments, we can see that both models fall slightly behind the baseline ASR model. Additionally, we can observe that the augmented labels approach performs better than

the multi-task approach. From Table 7, we can see how our models perform on the NER task when evaluated on the original and the generated transcripts. When evaluated on the original transcripts, the fine-tuned multi-task model performs better than the standard multi-task model, similar to the Finnish experiments. On the transcripts generated by the models, we can observe that the augmented labels approach achieves the highest F1 score. We can also observe that both the augmented labels and the fine-tuned multi-task approaches outperform the pipeline approach.

Table 6. WER on the Swedish test set.

Model	WER
Baseline ASR	**33.44**
AL	33.82
MT	34.58

Table 7. Precision, recall and F1 score for the Swedish test set.

Transcripts	**Model**	**Prec**	**Rec**	**F1**
Original	MT	97.76	91.27	94.40
	MT*	98.32	93.48	**95.84**
Generated	Pipeline	69.35	79.37	74.02
	AL	74.96	78.13	**76.51**
	MT	70.14	77.94	73.83
	MT*	74.19	76.67	75.41

5.3 English Results

Next, we present the results obtained on the English data sets. In Table 8, we can see how our models perform in terms of WER when evaluated on the LibriSpeech and the English-Gold test sets. From the table, we can see that both approaches perform slightly better than the baseline ASR model trained on the LibriSpeech data. On the English-Gold, on the other hand, the multi-task model performs slightly better than the baseline, whereas the augmented labels yields worse results. On the Libri clean test set, both approaches perform really close, whereas on the Libri other test set, the multi-task approach performs slightly better. Additionally, the multi-task approach performs better than the augmented labels on the English-Gold test set as well.

On the NER task, presented in Table 9, when evaluated on the original transcripts, the fine-tuned multi-task approach outperforms the normal multi-task approach on all the English data sets. On the transcripts generated by the models, we can see that the pipeline approach is better than our proposed E2E models on the LibriSpeech test sets. On the manually annotated English Gold test set, on the other hand, the multi-task approach achieves the best F1 score. Additionally, both the multi-task and the augmented labels approaches perform better than the pipeline approach.

Table 8. WER on the LibriSpeech and English-Gold test sets.

Model	Libri clean	Libri other	English-Gold
Baseline ASR	12.74	31.61	23.26
AL	**12.34**	30.88	23.51
MT	12.35	**30.56**	**23.07**

Table 9. Precision, recall and F1 score for the English test sets.

		Libri clean			Libri other			English Gold		
Transcripts	**Model**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
Original	MT	87.82	86.01	86.90	86.95	86.23	86.59	64.44	77.09	70.20
	MT*	88.41	86.46	**87.43**	87.55	86.13	**86.83**	81.86	68.02	**74.30**
Generated	Pipeline	76.43	79.09	**77.74**	64.07	74.40	**68.85**	79.24	71.28	75.05
	AL	79.77	63.47	70.69	70.21	52.15	59.85	82.60	69.30	75.21
	MT	74.63	76.77	75.68	60.90	73.44	66.59	77.04	84.89	**80.78**
	MT*	76.33	77.10	76.72	63.33	71.75	67.29	81.86	68.02	74.30

6 Analysis of the Results

To further investigate the NER performance of the models, we plotted confusion matrices. In Figure 1, we can see how the augmented labels and fine-tuned multi-task approaches perform on individual named entity classes on the Finnish data set. We can notice from the confusion matrices that both approaches are doing a pretty good job at detecting the entities, especially the location. On the other hand, they sometimes confuse non-entities with entities. This is especially visible in the person and organization classes, where some non-entities are tagged with either of them.

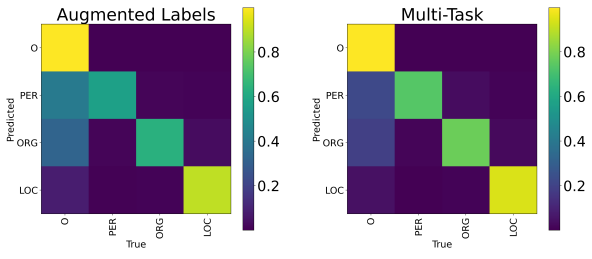

Fig. 1. Confusion matrices for the AL and MT* models, evaluated on the transcripts generated by the models, using the Finnish parliament test set.

Similar to the Finnish results, in Figure 2, we can observe that on the Swedish data set, the models do not have difficulties recognizing the entities. Furthermore, we can see that in a small number of cases, the models confuse the person entity with a location. Additionally, we can see that most of the mistakes that the models make are by confusing non-entities with entities, just like in the Finnish results.

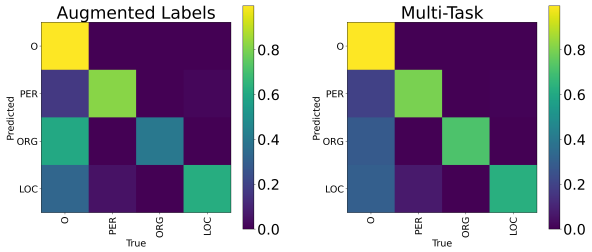

Fig. 2. Confusion matrices for the AL and MT* models, evaluated on the transcripts generated by the models, using the Swedish test set.

On the English-Gold test set, as shown in Figure 3, we can observe that the models make more mistakes than on the other data sets. That is especially the case with the organization entity. The reason for that could be because there are far fewer organization entities in the LibriSpeech and English-Gold data sets, in comparison to the other entities. To ensure that the bad recognition score for the organization entity is expected, we additionally compared the score to the one obtained by the pipeline model. When evaluated on the test data, the pipeline approach also got a low score for the organization entity. Generally, since the English-Gold data set is a combination of many different data sets, it is expected that the domain mismatch negatively impacts the NER.

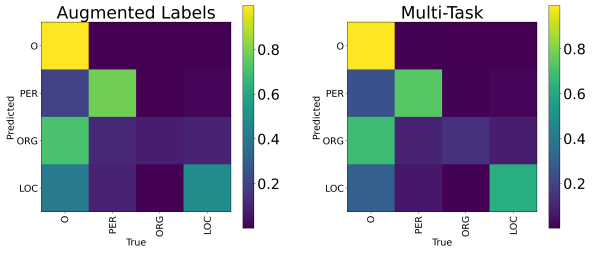

Fig. 3. Confusion matrices for the AL and MT* models, evaluated on the transcripts generated by the models, using the English-Gold test set.

7 Conclusion

In this paper, we presented two approaches for end-to-end named entity recognition and evaluated them on Finnish, Swedish, and English data sets. We showed that both approaches perform similarly in terms of WER, against the baseline models. Even though the WER results are not in pair with the current state of the art, the goal of this paper is to show that named entities can be learned in an E2E manner, without sacrificing too much of the ASR performance. This allows the ASR part to be optimized for the NER task and vice versa. In terms of the F1 score, both approaches achieve promising results. When comparing both systems, the multi-task approach outperforms the augmented labels approach on the NER task by a significant margin, in all the experiments, except the Swedish, when evaluated on the transcripts generated by the models. When compared against the standard pipeline approach, our proposed models achieve better results on most of the experiments. Generally, we can say that the multi-task approach is more flexible, allowing us to additionally fine-tune the NER branch, which gives an improvement in almost all the experiments. In the future, we plan to replace the models with a Transformer architecture and see how it performs in comparison to the BLSTMs.

8 Acknowledgment

This work was supported by the Kone Foundation. This work was supported by the Academy of Finland (grant 329267) and EU's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069). The computational resources were provided by Aalto ScienceIT.

References

1. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182 (2016)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
3. Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y.: Attention-based models for speech recognition. In: Advances in neural information processing systems. pp. 577–585 (2015)
4. Deoras, A., Sarikaya, R.: Deep belief network based semantic taggers for spoken language understanding. In: Interspeech. pp. 2713–2717 (2013)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Ghannay, S., Caubrière, A., Estève, Y., Camelin, N., Simonnet, E., Laurent, A., Morin, E.: End-to-end named entity and semantic concept extraction from speech. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 692–699. IEEE (2018)
7. Gravano, A., Jansche, M., Bacchiani, M.: Restoring punctuation and capitalization in transcribed speech. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4741–4744. IEEE (2009)

8. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376 (2006)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Jeong, M., Lee, G.G.: Jointly predicting dialog act and named entity for spoken language understanding. In: 2006 IEEE Spoken Language Technology Workshop. pp. 66–69. IEEE (2006)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). pp. 282–289 (2001)
13. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
14. Malmsten, M., Börjeson, L., Haffenden, C.: Playing with words at the national library of sweden – making a swedish bert (2020)
15. Mansikkaniemi, A., Smit, P., Kurimo, M., et al.: Automatic construction of the finnish parliament speech corpus. In: INTERSPEECH. vol. 8, pp. 3762–3766 (2017)
16. Mesnil, G., He, X., Deng, L., Bengio, Y.: Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In: Interspeech. pp. 3771–3775 (2013)
17. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210. IEEE (2015)
18. Porjazovski, D., Leinonen, J., Kurimo, M.: Named entity recognition for spoken finnish. In: Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery. pp. 25–29 (2020)
19. Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)
20. Toshniwal, S., Kannan, A., Chiu, C.C., Wu, Y., Sainath, T.N., Livescu, K.: A comparison of techniques for language model integration in encoder-decoder speech recognition. In: 2018 IEEE spoken language technology workshop (SLT). pp. 369–375. IEEE (2018)
21. Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T.: Hybrid ctc/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing **11**(8), 1240–1253 (2017)
22. Yadav, H., Ghosh, S., Yu, Y., Shah, R.R.: End-to-end named entity recognition from english speech. arXiv preprint arXiv:2005.11184 (2020)
23. Zhai, L., Fung, P., Schwartz, R., Carpuat, M., Wu, D.: Using n-best lists for named entity recognition from chinese speech. In: Proceedings of HLT-NAACL 2004: Short Papers. pp. 37–40 (2004)