Transfer Learning for British Sign Language Modelling

Abstract

Automatic speech recognition and spoken dialogue systems have made great advances through the use of deep machine learning methods. This is partly due to greater computing power but also through the large amount of data available in common languages, such as English. Conversely, research in minority languages, including sign language, is hampered by the severe lack of data. This has led to work on transfer learning methods, whereby a model developed for one language is reused as the starting point for a model on a second language, which is less resourced. In this paper, we examine two transfer learning techniques of finetuning and layer substitution for language modelling of British Sign language. Our results show improvement in perplexity when using transfer learning with standard stacked LSTM models, trained initially using a large corpus for standard English from the Penn Treebank corpus.

1 Introduction

Spoken dialogue systems and voice assistants have been developed to facilitate natural conversation between machines and humans. They provide services through devices such as Amazon Echo Show and smartphones to help the user do tasks (McTear, 2004) and, more recently, for open domain chitchat (Serban et al., 2016), all through voice. Recent advances have been facilitated by the huge amounts of data collected through such devices and have resulted in the recent success of deep machine methods, providing significant improvements in performance. However, not all languages are able to benefit from these advances, particularly those that are under-resourced. These include sign language and it means that those with hearing disabilities are not able to leverage such interactive systems nor the benefits that automatic transcription and translating of signing would afford.

Here, we advance the state of the art with respect to transcribing the British Sign Language (BSL), which is the official language for the deaf people in UK. BSL enables communication of meaning through parameters such as hand shape, position, hand orientation, motion, and non-manual signals (Stokoe, 1960). BSL can be perceived only visually as it has no standard notation for writing the signs, as with letters and words in English. Analogous to the International Phonetic Alphabet (IPA), highly detailed mapping of visual indicators to written form are available, such as HamNoSys (Hanke, 2004) or Sign-Writing (Sutton, 1995) among others. These, however, are not standardised and only a handful of experts know how to use them. Recent methods for automatic speech recognition (ASR) use deep neural models to bypass the need for phoneme dictionaries (Hannun et al., 2014), which are then combined with language models. Previous work (Mocialov et al., 2016; Mocialov et al., 2017), have shown that we can use visual features to automatically predict individual signs. This work follows on in that these individual signs are to be used with a language model to take into account context and therefore increase accuracy of the transcriber, which outputs a string of word-like tokens. These tokens are called glosses (Sutton-Spence and Woll, 1999; Cormier et al., 2015). All the BSL glosses in the corpus can be found in the BSL SignBank.

This paper focuses on language modelling, a common technique in the field of ASR and Natural Language Processing to model the likelihood of certain words following each other in a sequence. We

1http://bslsignbank.ucl.ac.uk/
propose to use transfer learning approaches, such as finetuning and layer substitution. The use of transfer learning technique can overcome the data sparsity issue in statistical modelling for scarce resource languages by using similar resources that can be found in large quantities and then further training the models on a specific low resource data.

Our contribution is that we show that a stacked LSTM model, trained on the preprocessed Penn Treebank (PTB) dataset\(^2\) that contains English language monolingual data and then finetuned on the preprocessed BSL monolingual corpus\(^3\), can achieve superior model quality than if training the same architecture directly on the BSL dataset. This is not necessarily an unsurprising result as there are marked differences between the two languages, particularly with the respect to the syntax (Sutton-Spence and Woll, 1999).

The paper begins with presenting methods for modelling languages and how they can be utilised in the BSL modelling. Section 2.2 gives an overview of how transfer learning can be achieved as well as the use of transfer learning in sign languages. Section 3 gives an overview of the datasets that are used in this paper, their statistics, and pre-processing steps to create two monolingual corpora for statistical model training. Section 4 describes in detail the setup for the experiments in this paper. Section 5 presents the results of the models employed for this research and discusses these results and the limitations of the approach taken in terms of the data used in Section 5.3. The paper is concluded and future work is proposed.

2 Related Work

2.1 Sign Language Modelling

Despite the availability of many alternatives for language modelling, such as count-based n-grams and their variations (Chen and Goodman, 1999; Rosenfeld, 2000; MacCartney, 2005; Bulyko et al., 2007; Guthrie et al., 2006), hidden Markov models (Dreuw and Ney, 2008; Dreuw et al., 2008), decision trees and decision forests (Filimonov, 2011), and neural networks (Deena et al., 2016; Mikolov et al., 2010), research in sign language modelling predominantly employs simple n-gram models, such as in (Cate and Hussain, 2017; Forster et al., 2012; Massó and Badia, 2010).

The reason for the wide-spread use of n-grams in sign language modelling is the simplicity of the method. There is an obvious disconnect between n-grams and the sign language in that sign language is perceived visually, while the n-grams are commonly applied to text sequence modelling. For this reason, the authors in (Stein et al., 2007; Zhao et al., 2000; Dreuw et al., 2008; Massó and Badia, 2010; Forster et al., 2013) model glosses, such as the ones shown on Figure 2, which are obtained from the transcribed sign languages.

Glosses model the meaning of a sign in a written language, but not the execution. Therefore, the true meaning of what was signed may get lost when working with the higher-level glosses. To overcome this issue and to incorporate valuable information into sign language modelling, additional features are added, such as non-manual features (e.g. facial expressions) (San-Segundo et al., 2009; Massó and Badia, 2010; Zhao et al., 2000; Stein et al., 2007).

2.2 Transfer Learning

While transfer learning is a more general machine learning term, cross-domain adaptation of language models is used in the language modelling literature (Deena et al., 2016; Ma et al., 2017). Models are usually trained on some specific domain that consists of a specific topic, genre, and similar features that can be identified by an expert. For example, a restaurant domain when a new type of a restaurant is created then the system needs to be able to adapt and be able to understand and discuss this new type of the restaurant. Unfortunately, it is nearly impossible to train a model for all possible configuration of current or future features. Commonly, a set of features are extracted from the raw data. When features change, re-training is required. Useful features can also be extracted without expert knowledge with such technique as Latent Dirichlet Allocation (LDA). These features usually take the form of words that

\(^2\)https://catalog.ldc.upenn.edu/ldc99t42

\(^3\)http://www.bslcorpusproject.org/
represent topics in the data (Deena et al., 2016). Best practice tries to avoid re-training the models every time one of the features changes as the domain changes due to the overhead involved.

Model-based adaptation to the new domains, on the other hand, is achieved by either finetuning or the introduction of adaptation layer(s) (Yosinski et al., 2014). Finetuning involves further training the already pre-trained model using the data from the new domain. The intuition behind the finetuning is that it is much quicker to learn new information with related knowledge. The adaptation layer approach incorporates new knowledge by re-training only the adaptation layer, whereas the rest of the model remains exactly the same as if it was used in the original domain and acts as a feature extractor for the new domain (Deena et al., 2016).

2.2.1 Transfer Learning in Sign Languages

Transfer learning has been applied to sign languages in computing for various purposes to demonstrate that the method is suitable for the task due to the lack of substantial domain-specific sign language data. Transfer learning has been successfully applied to static pose estimation, transferring the knowledge from pose estimation to the sign language pose estimation (Gattupalli et al., 2016) and classification of finger-spelled letters in American Sign Language (Garcia and Viesca, 2016; Karthick Arya, 2017; Chaudhary, 2017; Muskan Dhiman, 2017). In particular, most of the transfer learning in sign language has been applied to the static image recognition to recognise the hand shape in an image using convolutional neural networks.

3 Corpora

The BSL Corpus and the preprocessed Penn Treebank (PTB) corpus were chosen for this research. The monolingual PTB dataset consists of telephone speech, newswire, microphone speech, and transcribed speech. The dataset is preprocessed to eliminate letters, numbers, or punctuation and was used by Mikolov (2010). The BSL corpus contains video conversations among deaf native, near-native and fluent signers across the United Kingdom. Almost all of the approximately one hundred recorded conversations are annotated for thirty seconds each at the gloss level using ELAN annotation tool (Schembri et al., 2013).
All the recording of the signers were made using up to four standard video cameras with a plain backdrop to provide full body view of the individuals, as well as, views from above of their use of signing space. The conversations between the signers included signing personal experience anecdotes, spontaneous conversations (Schembri et al., 2013).

The BSL data that we focused on was narratives between two participants, where one person had to think of a topic to sign about to another participant during the elicitation.

The corpus is annotated with glosses, taken from the BSL SignBank in ELAN as shown in Figure 2a. Figure 2b shows all the glosses of the first two sentences. As mentioned above, gloss is an identifier of a unique sign, written in English and should represent its phonological and morphological meaning (Schembri et al., 2013). In the corpus, the glosses are identified throughout the videos for both left and right hands as sometimes different signs can be signed with different hands at the same time. Apart from the glossing, the annotations include the corresponding free English written translation of the meaning of the signing split into sentences (see the Free Translation in the Figure 2). Figure 2c shows which glosses are considered for the BSL modelling and which are ignored. This is done to match the vocabulary of the PTB corpus for the transfer learning purposes.

3.1 Data Pre-processing

For the BSL corpus, we ignore the free translation and extract English text from the glosses, preserving the order of the signs executed. For example, in Figure 2, right-hand glosses identify the following order of the signs: good, explain, about, puppy, etc. excluding glosses, such as PT:PRO for pointing signs or PT:POSS for possessives and others (Figure 2c), which are explained in more detail in Fenlon et al. (2014). Since the gloss annotation does not include explicit punctuation, it is impossible to tell where a signed sentence begins and where it stops. To overcome this limitation of the gloss annotation, we use the Free Translation annotation, which gives the boundaries of sentences in videos. Later, we split the extracted glosses into sentences using these sentence boundaries. By the end of the pre-processing stage, we have glosses (excluding special glosses for pointing signs, possessives or other non-lexical glosses) in the order that the corresponding signs were executed in the video, split into sentences. As a result, we extracted 810 sentences from the BSL corpus with an average length of the sentence being 4.31 words, minimum and maximum lengths of 1 and 13 words respectively. A monolingual dataset has been created with the extracted sentences. As obtained from the PTB dataset (Merity et al., 2017), the English language corpus has 23.09 words on average per sentence with minimum being 3 and maximum 84 words per sentence. The pre-processed BSL corpus has a vocabulary of 666 words, while the PTB dataset has a vocabulary of 10,002 words. From this point on in this paper, we will use the term ‘words’ to refer to both glosses in the BSL and words in the PTB datasets because we aim to use a common vocabulary for training our models.

---

Figure 2: a) The BSL Corpus Annotation in ELAN; b) Table shows full text of the annotated glosses for the two first sentences from the ELAN annotation; c) Glosses that are used for the BSL modelling

---

3 http://www.bslcorpusproject.org/cava/
Both monolingual datasets were split into training, validation, and testing sets as required for training and evaluation of the statistical models. Both datasets were split using ratio 85:15. The smallest part, in turn, was split 50:50 for validation and testing for the two datasets.

4 Language Modelling Methodology

4.1 Statistical Language Models

Perplexity measure has been used for evaluation and comparison purposes of different models. We used the following formula to calculate the perplexity values: $e^{\text{Cross-Entropy}}$ as used in (Bengio et al., 2003), which approximates geometric average of the predicted words probabilities on the test set. We have explicitly modelled out-of-vocabulary (OOV), such as `<unk>` placeholder in all the experiments.

4.1.1 Neural Models

For comparison, we use two methods: 1) stacked LSTM and 2) Feed-Forward (FFNN) architectures to create the BSL language models. All models are implemented in PyTorch\textsuperscript{6} with weight-drop recurrent regularisation scheme for the LSTMs, which is important for overcoming commonly known LSTM model generalisation issues (Merity et al., 2017; Merity et al., 2018). The feed-forward model, on the other hand, had no regularisations as it is less susceptible to overfitting due to the much smaller number of parameters.

The parameters that were modified to achieve the lowest perplexity were input size of the overall input sequence for the recurrent neural network (back-propagation through time, BPTT), batch size, and learning rate. The parameters were selected using the grid search approach using perplexity metric. All the neural models were trained for 100 epochs.

In the case of the neural networks, the sequences of words were tokenised (i.e. turned into integers) and the tokenisation was stored to ensure the same tokenisation during the transfer learning phase. The input, therefore, consisted of a set of tokens, while the outputs (i.e. predicted words) were turned into a one-hot vectors.

![Diagram of neural models](http://pytorch.org/)

Figure 3: The two types of neural models used to test transfer methods for sign language modelling

---

\textsuperscript{6}http://pytorch.org/
4.1.1.1 Stacked LSTMs

Figure 3a shows the architecture of the stacked LSTM model. The model consists of an embedding layer of 400 nodes, which, together with the tokenisation, turns string of words into a vector of real numbers. Secondly, three LSTM layers with 1150 nodes each are stacked vertically for deeper feature extraction. Thirdly, the linear layer downsizes the stacked LSTMs output to the vocabulary size and applies linear transformation with softmax normalisation. The weights of the embedding and the linear layers are tied. This means that the two layers share the same weights, which reduces the number of parameters of the network and makes the convergence during training faster. The same architecture was used in (Merity et al., 2017) to model PTB dataset, reporting 57.3 perplexity, utilising cache in the model from recent predictions.

4.1.1.2 FFNN

Figure 3b shows the Feed-forward model architecture. The model does not have the stacked LSTMs layers. Instead, the stacked LSTMs are substituted with one hidden fully-connected rectifier layer, which is known to overcome the vanishing gradient problem. The weights of the embedding and the outputs layers are not tied together. Similar architectures have been used for language modelling in (Le et al., 2013; Mikolov et al., 2009; de Brébisson et al., 2015) with the hidden layer having different activation functions with the PTB dataset being used in (Audhkhasi et al., 2014), reporting 137.32 perplexity.

4.1.2 Training the Models

Transfer learning was achieved with both finetunning and substitution. Both FFNN and LSTM were trained on the PTB dataset and then either finetuned or the last layer was substituted with the new adaptation layer, freezing the rest of the weights, and further training on the BSL dataset.

To achieve finetunning, first the best model is saved after the training of both the FFNN and the stacked LSTMs on the PTB dataset. Then the training is restarted on the BSL corpus, having initialised the model with the weights, trained on the PTB dataset.

To perform layer substitution as a transfer learning approach, the same first step as with the finetunning is repeated and the model, trained on the PTB, is saved. When the training is restarted on the BSL dataset, the saved model is loaded and the last linear layer is substituted with a layer that has as many nodes as the BSL vocabulary. Later, all the weights of the network are locked and will not be modified during the optimisation. Only the weights of the last substituted layer will be modified. This method uses the pretrained network as a feature extractor and only modifies the last layer weights to train the model for the BSL dataset.

5 Results

This section is split into two subsections. We firstly present results without transfer learning, namely both the FFNN and the stacked LSTMs models trained and tested on the PTB dataset or trained and tested on the BSL. Later we present results with the transfer learning, where both FFNN and the stacked LSTMs models are trained on the PTB dataset and then finetuned and tested on the BSL.

To show that the two languages are different as discussed in Section 3.1, we applied model, trained on one language to the other language and vice versa. As a result, the model trained on English language and applied to the BSL scored 1051.91 in perplexity using SRILM toolkit (Stolcke, 2002). Conversely, a model trained on the BSL has been applied to the English language and scored 1447.23 in perplexity. As expected, the perplexity is high in both cases, which means that the probability distribution over the next word in one language is far from the true distribution of words in the other language.

5.1 Without Transfer Learning

Table 1 shows perplexities on the two datasets with two statistical models. From the table, we can infer that the trained models on the PTB dataset have lower perplexity than the same architectures trained on the BSL dataset. This can be explained by the fact that the PTB dataset has more data than the BSL dataset.
### Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Penn Treebank (PTB)</th>
<th>The BSL Corpus Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFNN 5-gram</td>
<td>190.46</td>
<td>258.1</td>
</tr>
<tr>
<td>Stacked LSTMs</td>
<td>65.91</td>
<td>274.03</td>
</tr>
<tr>
<td>OOV</td>
<td>6.09%</td>
<td>25.18%</td>
</tr>
</tbody>
</table>

Table 1: Perplexities on either the PTB or the BSL test sets using models trained and tested on the same corpus (i.e. PTB and BSL)

...dataset and, therefore, statistical models can generalise better. Furthermore, the amount of data is further reduced in the BSL case as the OOV covers a quarter of the overall dataset.

### 5.2 With Transfer Learning

Table 2 shows perplexities on the two datasets with two statistical models, applying transfer learning. From this table, it can be seen that the substitution approach gives very similar results independent of the whether FFNN or stacked LSTMs model is used (123.92 versus 125.32). The best result is achieved with the finetunning approach on the stacked LSTMs model, while the worst perplexity result is on the FFNN model with the finetunning approach. Similar results have been reported in (Irie et al., 2016), where finetuned GRU performed worse than finetuned LSTM model. In addition, the OOV count differs from that of the Table 1 due to the fact that a subset of the vocabulary, observed in the PTB dataset during training is then identified in the BSL dataset during testing.

<table>
<thead>
<tr>
<th>Method</th>
<th>Finetunning</th>
<th>Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFNN</td>
<td>179.3</td>
<td>123.92</td>
</tr>
<tr>
<td>Stacked LSTMs</td>
<td>121.46</td>
<td>125.32</td>
</tr>
<tr>
<td>OOV</td>
<td>12.71%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Perplexities on the BSL test set after applying the transfer learning on FFNN and LSTMs

### 5.3 Discussion

The salient idea of this paper is whether transfer learning is a legitimate method for modelling one language with the knowledge of another, assuming the languages are different, but share some common properties, such as vocabulary. This theory is intuitive and has been discussed in linguistics (Kaivapalu and Martin, 2007). In our case, PTB corpus covers most of the vocabulary found in the BSL corpus (12.71% OOV) by the virtue of the gloss annotation of the BSL corpus (Schembri et al., 2013). However, the languages are assumed to be different as they evolved independently of one another.

The results obtained are different from the ones reported in similar research. For example, for the FFNN model, Audhkhasi et al. (2014) report 137.32 versus our achieved 190.46 perplexity and for the stacked LSTMs model, Merity et al. (2017) report 57.3 versus our achieved 65.91 perplexity. This can be explained by the fact that not all the regularisation techniques had been used in this research as in the past research and the model training had been restricted to 100 epochs. Further training may further reduce the perplexity to that reported in Merity et al. (2017).

From the results, we can see that the transfer learning leads to superior models than the models trained on the BSL directly (258.1 and 274.03 against 123.92 and 125.32). Since the quality of the trained models...
using either of the approaches is similar in case of the stacked LSTMs model (121.46 and 125.32), the choice between the finetunning and substitution can be guided based on the convergence speed. During the substitution, only one layer of the network is replaced with a new one and the rest of the weights in the network are locked, therefore, one set of weights will be optimized. This is in contrast to the finetunning method, which optimizes all of the weights, which may, in turn, require more interactions, depending on how different the new data is.

6 Conclusion

This paper shows how transfer learning techniques can be used to improve language modelling for the BSL language at the gloss level. Statistical modelling techniques are used to generate language models and to evaluate them using a perplexity measure.

The choice of the transfer learning technique is guided by the scarcity of available resources of the BSL language and the availability of the English language dataset that shares similar vocabulary with the annotated BSL. Feed-forward and recurrent neural models has been used to evaluate and compare generated language models. The results show that the transfer learning can achieve superior quality of the generated language models. However, our pre-processed BSL corpus lacks constructs that are essential for a sign language, such as classifier signs, gestures, pointing signs, buoys and others. Nevertheless, transfer learning for modelling the BSL shows promising results and should be investigated further.

6.1 Future Work

Although this paper discusses the use of a model initially trained on English and presents promising preliminary results, the annotation of the BSL, used in this paper, is limited as this paper serves as a proof of concept. In particular, the annotation used is missing some of the grammatical aspects of the BSL, such as classifier signs, gestures, pointing signs, buoys and others. Inclusion of these modifiers into the BSL dataset for the BSL language modelling would increase the OOV count as the English language does not have equivalent language constructs. This raises a question whether a sign language can be modelled using other languages that may have these constructs, but would consists of a different vocabulary. More generally, is it possible to model a language with the transfer learning using other less-related languages? Similar questions have been partly answered for the written languages in the field of machine translation (Gu et al., 2018) by bringing words of different languages close to each other in the latent space. Unfortunately, nothing similar has been done for the sign languages.

From the methodological side of the modelling, additional advanced state of the art techniques should be experimented with to achieve greater quality of the generated models, such as attention mechanism for the recurrent neural networks. Finally, since this paper focuses on just one part of the use of the sign languages in computing, the approach described here should be employed as a part of a conversational system for the sign language users.

References


