Leveraging Pre-Trained Embeddings for Welsh Taggers

Ignatius Ezeani[1] **Scott Piao**[1] **Steven Neale**[2] **Paul Rayson**[1] **Dawn Knight**[2]
[1]School of Computing and Communications, Lancaster University, UK.
{i.ezeani, s.piao, p.rayson}@lancaster.ac.uk
[2]School of English, Communication and Philosophy, Cardiff University, UK.
{NealeS2, knightd5}@cardiff.ac.uk

Abstract

While the application of word embedding models to downstream Natural Language Processing (NLP) tasks has been shown to be successful, the benefits for low-resource languages is somewhat limited due to lack of adequate data for training the models. However, NLP research efforts for low-resource languages have focused on constantly seeking ways to harness pre-trained models to improve the performance of NLP systems built to process these languages without the need to re-invent the wheel. One such language is Welsh and therefore, in this paper, we present the results of our experiments on learning a simple multi-task neural network model for part-of-speech and semantic tagging for Welsh using a pre-trained embedding model from *FastText*. Our model's performance was compared with those of the existing rule-based stand-alone taggers for part-of-speech and semantic taggers. Despite its simplicity and capacity to perform both tasks simultaneously, our tagger compared very well with the existing taggers.

1 Introduction

The Welsh language can easily be classified as low resourced in the context of natural language processing because the lack of the commonly used resources in language research such as large annotated corpora as well as the standard computational tools and techniques for processing these resources.

There is still a long way to go for Welsh, but the situation is improving. For instance, Welsh is fortunate to have a fund that supports an on-going inter-disciplinary and multi-institutional project, the National Corpus of Contemporary Welsh (Corpws Cenedlaethol Cymraeg Cyfoes - *CorCenCC*)[1], which has been building a large-scale open-source language resource for contemporary Welsh language.

Existing Welsh part-of-speech (sections 2.1) and semantic (section 2.2) taggers produce good results, but their heavy dependence on hand-crafted rules and hard-coded resources may pose a maintenance challenge in future. Also, considering the speed with which languages evolve, especially on the internet, and the huge amount of unannotated corpora that can be collected from the web, we urgently need a system that is capable of learning from unstructured text in order to guarantee the generalisability and scalability of tagging tools.

Given the potential challenges with the existing approaches and considering the similarities between the tasks of part-of-speech (POS) and semantic (SEM) annotation, we propose to train a single neural network model that can jointly learn both of the tasks. We aim at requiring as little human annotation effort as possible and leveraging the linguistic patterns acquired from unsupervised language models like word embeddings. The main contributions of this research includes: (1) The first application of multi-task learning to POS and semantic tagging for any language that we know of, (2) The ability to improve OOV coverage for the Welsh language using pre-trained embeddings for semantic category extension, (3) Public release of two sets of manually checked gold-standard corpora for POS and semantic tagging of Welsh, (4) Inter-annotator agreement scores for Welsh semantic tagging, (5) Public release of the first Welsh semantic tagger (CySemTagger) (6) The first demonstration of multi-task learning to improve NLP task accuracy for Welsh, and (7) A demonstration of the usefulness of multi-task learning in a mono-lingual setting for a low re-

[1] http://www.corcencc.org/

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 270–280
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

source language.[2]

2 Background

POS tagging is a well studied NLP task. Much recent work on this task has moved away from English and European languages to other major languages such as Arabic (Aldarmaki and Diab, 2015), Chinese (Sun and Wan, 2016), dialects thereof (Darwish et al., 2018), and text types containing more noise such as historical (Yang and Eisenstein, 2016; Janssen et al., 2017), learner language (Nagata et al., 2018), code switching (Vyas et al., 2014) and social media varieties (Horsmann and Zesch, 2016; van der Goot et al., 2017). More recently, joint and multi-task learning approaches have been applied to link POS tagging and other tasks such as segmentation or tokenisation (Al-Gahtani and McNaught, 2015; Shao et al., 2017), dependency parsing (Nguyen and Verspoor, 2018) and lemmatisation (Arakelyan et al., 2018).

Besides being applied to other NLP applications and levels, multi-task learning has been applied with promising results to the semantic level in various scenarios, including cross-lingual sentiment analysis (Wang et al., 2018), opinion and semantic role labelling (Marasović and Frank, 2018), semantic parsing (Bordes et al., 2012), emotion prediction (Buechel and Hahn, 2018), irony detection (Wu et al., 2018) and rumour verification (Kochkina et al., 2018). However, there is very little research that applies multi-task learning to link Word Sense Disambiguation (WSD) or semantic tagging with another task. Here, we refer to the semantic tagging as coarse-grained word sense disambiguation based on an existing taxonomy of categories, e.g. in USAS (Rayson et al., 2004). Previously, semantic tagging in multiple languages has been shown to greatly benefit from POS tagging in the NLP pipeline, since it can help to filter out inapplicable semantic fields from the set of possible candidates (Piao et al., 2015).

Over the past few years, researchers started to port NLP tools and methods into low resource languages using a various approaches, such as porting lexicons from one language to another using bilingual dictionaries and parallel corpora (Piao et al., 2016) and cross-lingual word embeddings (Adams et al., 2017; Sharoff, 2018). Multi-task learning has also been proved useful in transferring the learning across languages in a multilingual setting where one of the languages has only sparse resources available (Junczys-Dowmunt et al., 2018; Lin et al., 2018; Choi et al., 2018), although less successful in named entity recognition settings (Enghoff et al., 2018). In our experiments, we focus on a low-resource mono-lingual setting with a small manually corrected corpus, and combine the Welsh POS and SEM annotation for the first time.

2.1 *CyTag*

The rule-based POS tagger under consideration in our work, *CyTag* (Neale et al., 2018), was built based on Constraint Grammar (CG) (Karlsson, 1990), in particular built around the latest version of the software, VISL CG-3[3]. The *CyTag* tagset[4] contains 145 fine-grained POS tags that can collapse into 13 *EAGLES*[5]-conformant broader categories.

CyTag utilises three steps to assign POS tags to tokens:

- A list of candidate POS tags is produced for each token.
- The list of candidate tags for each token is pruned to as few as possible (ideally one) using CG-formatted rules.
- The optimal tag for each token is selected, helped by some small additional processing steps for any cases that were still ambiguous after post-CG.

In the second step listed above, *CyTag* makes use of a CG-formatted 'grammar' file – currently containing 243 hand-crafted and hard-coded rules – to 'prune' the list of candidate tags to one for ambiguous tokens. The rules are formatted as follows:

action (reading) if (neighbour (features))

whereby *action* refers to the 'operation' to be performed on the *reading* e.g. ('selecting or 'removing'); *neighbour* is a nearby token of interest to the target token on whose *features* the *action* depends. *CyTag* was evaluated using a gold-standard annotated corpus containing 611 sentences (14,876 tokens), as will be described in subsection 3.1.

[2] Gold-standard corpora and tools are available on our GitHub account: https://github.com/CorCenCC

[3] http://visl.sdu.dk/cg3.html

[4] http://cytag.corcencc.org/tagset

[5] http://www.ilc.cnr.it/EAGLES/browse.html

Another recently-developed POS tagger for Welsh is the *WNLT-Tagger*, which forms part of the *Welsh Natural Language Toolkit (WNLT)*[6]. *WNLT-Tagger* is one of the four main modules in *WNLT*, which is itself built on the *GATE (General Architecture for Text Engineering)* framework (Cunningham, 2002).

2.2 *CySemTagger*: The Welsh Semantic tagger

CyTag is a precursor to *CySemTagger* (Piao et al., 2018) which is an automatic semantic annotation tool that depends on the POS tagged output to assign semantic tags to tokens in Welsh texts. *CySemTagger* employs the semantic tagset of Lancaster University's *UCREL Semantic Analysis System*, USAS[7]. The semantic tagset, which was originally derived from Tom McArthur's *Longman Lexicon of Contemporary English* (McArthur and McArthur, 1981), has 21 major discourse fields and 232 tags.

The *CySemTagger* is a knowledge-based and rule-based system with the following key components:

- lexicon look-up (both for single words and MWEs)
- part-of-speech tagging (*CyTag* and *WNLT-Tagger*)
- semantic category disambiguation
- output formatting and display

The *CySemTagger* tagger is designed to work with any POS-tagger but its performance was assessed so far only on the *coverage* of the Welsh text presented to it, i.e. the fraction of the tokens it is able to assign at least one of the valid semantic tags. The experiment presented in (Piao et al., 2018) indicates that, on the text coverage evaluation, the *CySemTagger* works better with *CyTag* than with *WNLT-Tagger*, as shown by the respective text coverage scores of 91.78% and 72.92% with both POS taggers.

3 Experiments

The *CyTag* and the *CySemTagger* are separate tools that use rule-based methods to achieve their results. The semantic tagger relies heavily on a part-of-speech tagger to function. The key aim of this paper is to implement a tagging system that:

- learns from unstructured data,
- leverages available embedding models,
- performs both tasks, POS and semantic tagging, simultaneously using a multi-task learning set up.

3.1 Experimental data

As mentioned earlier in section 2.1, the instances for training the POS and semantic taggers were extracted from the manually annotated gold standard evaluation corpus that has been constructed in the CorCenCC project, i.e. the data used for the *CyTag* and *CySemTagger* development. This training data comprises 611 tagged sentences (14,876 tokens) stored in eight input files that contain excerpts from a variety of existing Welsh corpora, including *Kynulliad314* (Welsh Assembly proceedings), *Meddalwedd15* (translations of software instructions), *Kwici16* (Welsh Wikipedia articles), *LERBIML17* (multi-domain spoken corpora) and some short abstracts of three additional Welsh Wikipedia articles. The fully manually checked version of the gold standard data, i.e. with the POS and SEM tags, will be released along with the multi-task model for parts-of-speech and semantic tagging.

The dataset used for training the multi-task model was built with the data instances extracted from the fully tagged version of the gold standard data. These data instances do not contain unambiguous tokens (e.g. punctuation and numbers) and those categorised as *unknown* are removed from the training data. The basic statistics from the data used in our experiment are shown in Table 1.

Although the data used in this experiment is comparatively smaller than what is often used by typical neural network projects, we assume it is sufficient for an exploratory research that aims to build a prototypical framework to support further developments for the Welsh language tools.

3.2 Embedding model

A key contribution of this work to Welsh NLP research is the application of pre-trained embeddings to build the model. Although most deep-learning frameworks provide an embedding layer

[6] https://sourceforge.net/projects/wnlt/

[7] http://ucrel.lancaster.ac.uk/usas/

Key item	Counts
sentences	611
tokens	14876
vocab length	3902
model vocab	3821
model vecsize	300
model oov	81
tagset size	392
punctuation	1667
unknown tags	44

Table 1: Basic statistics from the training data and embedding model used in this experiment.

that allows one to create embeddings from the training data, it is more beneficial to leverage existing models trained with much larger Welsh text data than to only rely on what is currently available. To that effect, we used the Welsh pre-trained embedding models built by the *FastText Project*[8] (Grave et al., 2018).

3.3 Design of experiment

The key input data to our pipeline consists of the 611 sentences that are jointly annotated with the POS and semantic tags. The combination of the annotation tags on the gold standard data makes it possible to extract the data in the different formats, as shown in Table 3. However, the format used for this experiment is the last one, *3-BOTH*, in which each token is tagged with a concatenation of the POS and semantic tags.

The extraction of the instance features for each token is carried out in two stage process which involves the *chunking* of the target word along with its three previous tokens (i.e. 4 words in total), as well as the *vectorisation* of the features. The chunking process proceeds with a sliding window along the sentence, with the target word being the rightmost in the chunk. The *vectorisation* then replaces each word in the chunk with its vector representation from a word-embedding model, forming a matrix of values that represent each training instance. The label for each instance is the *tag-ID* i.e. a unique integer number assigned to each of the tags.

3.4 Model architecture and training setup

The model we used is a simple neural network with only one hidden layer. Each instance is a con-

	Training		Evaluation	
	Accuracy	*Loss*	*Accuracy*	*Loss*
-Vector size				
10	72.44	1.160	70.26	3.517
50	99.09	0.036	94.51	5.313
100	99.04	0.032	94.76	4.775
200	99.09	0.027	95.23	4.650
300	99.05	0.030	95.38	4.994
-Mini-batch size				
8	99.08	0.032	95.29	4.450
16	99.10	0.030	95.55	4.552
32	99.04	0.034	95.03	4.758
64	99.13	0.030	94.97	4.905
-Dropout rates				
10	99.11	0.033	95.38	3.807
20	98.60	0.051	94.80	3.873
30	97.68	0.083	94.27	3.434
40	95.92	0.137	92.85	3.362
50	93.08	0.232	90.32	3.280

Table 2: Parameter optimisation: Training and Evaluation of scores on *Accuracy* and *Loss*. Parameter values in **bold** were chosen.

Tagtype	Example
0 - None	A fydd rhywfaint o ’r arian hwn yn cael ei ddefnyddio i sicrhau bod modd defnyddio tocynnau rhatach yn Lloegr yn ogystal ag yng Nghymru ?
1 - POS	A/Rha fydd/B rhywfaint/E o/Ar ’r/YFB arian/E hwn/Rha yn/U cael/B ei/Rha ddefnyddio/B i/Ar sicrhau/B ...
2 - SEM	A/Z5 fydd/A3 rhywfaint/N5 o/Z5 ’r/Z5 arian/I1 hwn/A3 yn/Z5 cael/A9 ei/Z8 ddefnyddio/A1 i/Z5 sicrhau/A7 ...
3 - BOTH	A/Rha/Z5 fydd/B/A3 rhywfaint/E/N5 o/Ar/Z5 ’r/YFB/Z5 arian/E/I1 hwn/Rha/A3 yn/U/Z5 cael/B/A9 ei/Rha/Z8 ddefnyddio/B/A1 i/Ar/Z5 sicrhau/B/A7 ...

Table 3: Different annotation formats for the experimental data. We used the *3-BOTH* format which combines the POS and semantic tags.

[8] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.cy.300.vec.gz

catenation of the embedding vectors of the target word and the three previous words. So the size of the input layer is the same as the length of the concatenated vectors. The key parameters required for the model training and evaluation are *vector size*, *mini-batch size* and *dropout rate*, and different values of each parameter are tested over runs of 50 epochs for each as shown in Table 2.

The output layer is the size of the tagset extracted from the training data. From the annotation format used, each token's tag is a combination of the POS and semantic tags and, as shown in Table 1, the total tagset size is 392. This is comparatively large but it will help facilitate the multi-task learning, which this work aims to achieve.

The model architecture is shallow, as only one hidden layer is used. Ideally, the size of the hidden layer should be somewhere between the sizes of both the input and the output layers (Reed and Marks, 1999). However, in order to reduce the number of parameters in this model, the size of 256 was arbitrarily chosen. For the hidden layer, the *Adam* optimiser (Kingma and Ba, 2014) was used with the *rectified linear unit* (*ReLu*) activation function (Nair and Hinton, 2010) as implemented in the integrated *TensorFlow-Keras* (Abadi et al., 2016), (Chollet et al., 2015) framework.

3.4.1 Vector size

Given the small size of the training data, and in order not to have too many parameters that can cause over-fitting, we tested the model with different vector sizes, (i.e. 10, 50, 100, 200, 300), averaged across a range of other parameters values for the *mini-batch* and *dropout*. The training and evaluation for parameter optimisation was performed over 50 epochs.

With regards to the evaluation accuracy, as shown in Figure 1, apart from $nvecs = 10$, all other vector sizes could converge within the first 30 to 40 epochs. However, the evaluation loss begins to rise within the first 10 epochs, with most $nvecs$ hitting nearly above 4.5 before reaching the 50th epoch. To balance this, a vector size of 100 was used, i.e. only the first 100 values were taken from each embedding vector to build the input layer, as suggested in (Brownlee, 2017). This produced an input layer size of 400.

3.4.2 Mini-batch size

The training set was chunked into *mini-batches* as described in (Ruder, 2016), with 8 instances per batch. The *mini-batch* values 8, 16, 32 and 64 were tested across other parameter values (see Figure 2). Their average performances indicate that, while there is only a small change in evaluation accuracies across the values, there is a slightly lower loss value with a mini-batch of 8 than the others.

3.4.3 Dropout rate

Given the small quantity of the training data, the architecture also implemented *dropout* regularisation (Srivastava et al., 2014) on the hidden layer to reduce the expected likelihood of *over-fitting*. Different dropout rates (10%, 20%, 30%, 40% and 50%) were tested as shown in Figure 3, and dropout rate of 30% was chosen to jointly mitigate the impact of on both the evaluation accuracy and the loss.

3.4.4 Batch Normalisation

Batch normalisation addresses the problem of *internal covariate shift* (Ioffe and Szegedy, 2015) by normalising the inputs to the model layers, thereby increasing the training speed. In some cases, it acts as a regulariser. Therefore, a version of the model architecture described above implements batch normalisation. This is because, during training, improvement rates in the model's evaluation accuracy slow down after the first 50 epochs while the loss continues to escalate. Techniques that speed up the learning were considered to investigate the combined impact of speed and regularisation on evaluation accuracy and loss.

3.4.5 Loss Function

As a multi-class classification task, the standard loss function is the *cross-entropy* with the $softmax$ logistic activation function, as described in equation 3.4.5 (Mannor et al., 2005).

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} log(p(y|X_i)_t)1[y_i = t] \qquad (1)$$

where N is the number of instances in the training batch, T is the number of unique tags while X_i, and y_i are a set of input values and the corresponding label respectively.

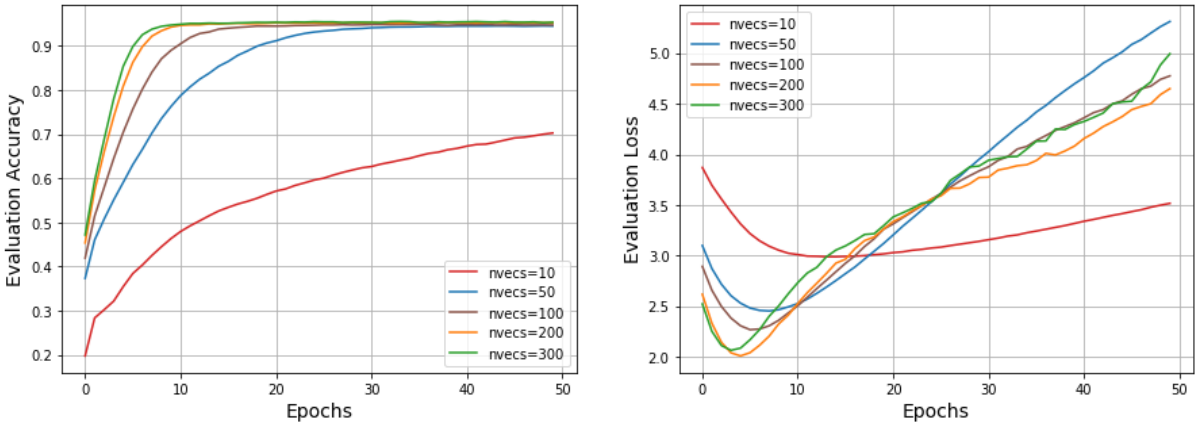

Figure 1: *Accuracy* vs *Loss* for different vector sizes

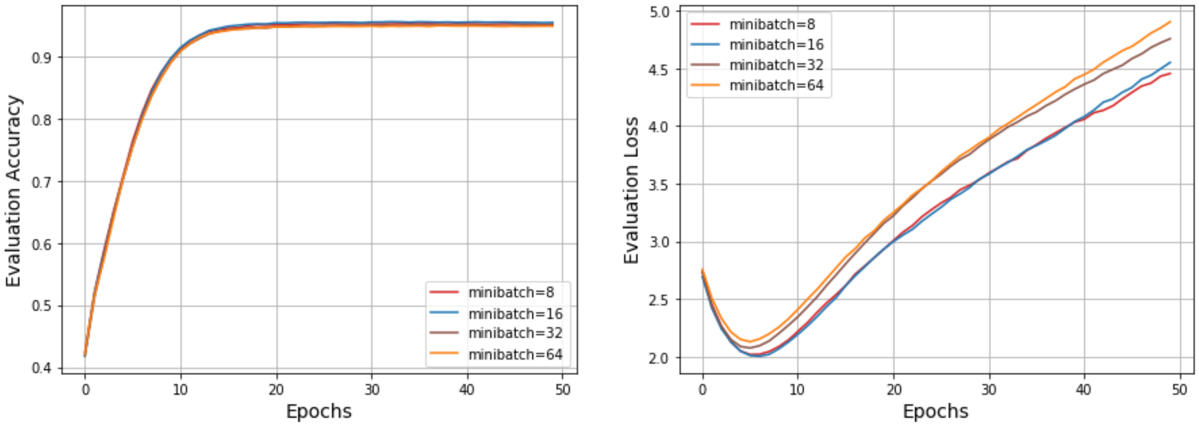

Figure 2: *Accuracy* vs *Loss* for different mini-batch sizes

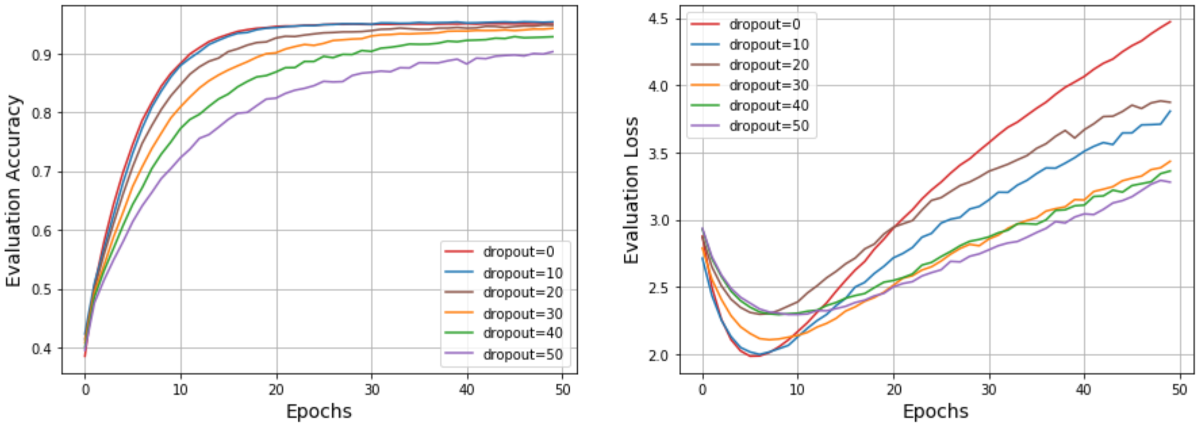

Figure 3: *Accuracy* vs *Loss* for different dropout rates

	Training		**Evaluation**	
	Accuracy	*Loss*	*Accuracy*	*Loss*
-dropout,-batchnorm	99.23	0.021	95.24	6.161
-dropout,+batchnorm	95.51	0.144	92.57	3.837
+dropout,-batchnorm	98.36	0.050	94.89	4.880
+dropout,+batchnorm	88.88	0.350	86.66	2.682

Table 4: Result summary for training and evaluation of accuracy and loss with or without dropout and batch normalisation

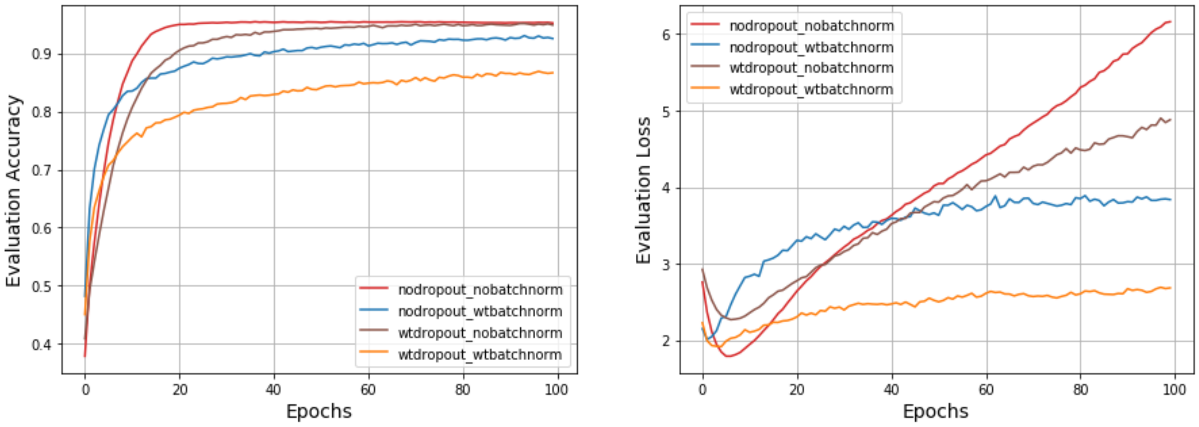

Figure 4: Evaluation graph for both accuracy and loss with and without dropout and/or batch normalisation.

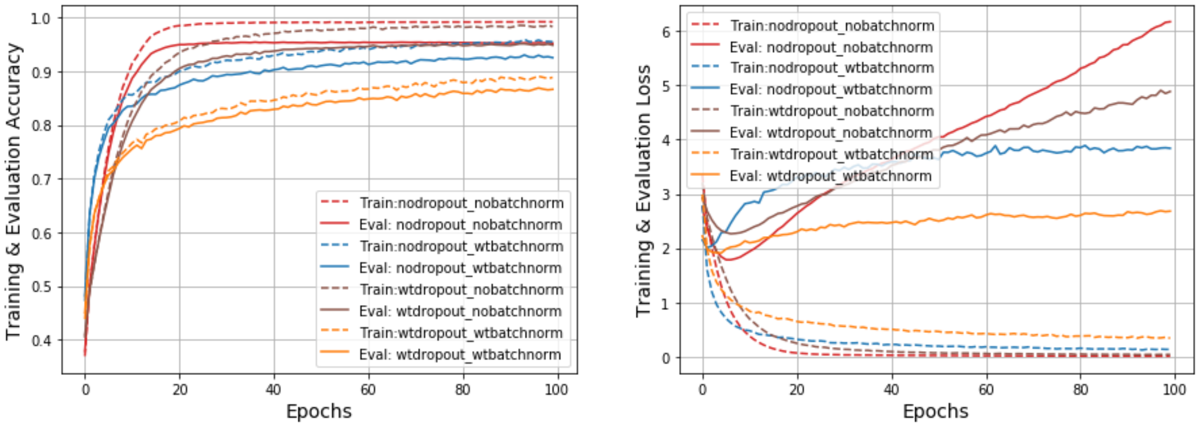

Figure 5: Training and evaluation graphs for accuracy and loss with and without dropout and/or batch normalisation.

4 Evaluation and discussion

With the *accuracy* of 93.64% and the *F1* of 95.06% reported previously for the *CyTag*, it represented the state-of-the-art in Welsh POS-tagging. Also, although the *CySemTagger* did not report those specific metrics, it is currently the only semantic tagger for Welsh language that we are aware of. Therefore, the evaluation results from the multi-tagger built in this experiment, which simultaneously performs both POS- and SEM-tagging, were compared against these tools.

The effects of dropout regularisation and batch normalisation were examined with the previously selected parameters for *vector size=100*, *mini-batches=8* and *dropout rate=30%*. As shown in Table 4, the results indicate that, at the detriment of accuracy, both dropout and the batch normalisation achieved significant reductions in evaluation loss. Without them, the training accuracy and loss scores for the multi-task tagger are 99.23% and 0.021 respectively while the evaluation scores are 95.24% and 6.161. However, with only dropout, training accuracy and loss scores are 98.36% and 0.050 while those of evaluation are 94.89% and 4.880.

Batch normalisation without dropout produced accuracy and loss scores of 95.51% and 0.144 respectively while those of evaluation produced 92.57% and 3.837 respectively. The combination of them achieved a significant reduction in evaluation loss (2.682), but with relatively poorer accuracy scores for training (88.88%) and evaluation (86.66%).

Figures 4 and 5 show that, as used in this experiment, the batch normalisation had a more regularising effect than the dropout, thereby slowing down convergence and avoiding over-fitting.

5 Conclusion

The main motivation for this work is to contribute a useful tool to the fledgling Welsh NLP research effort. There are two key objectives of this work: a) To build a multi-task classifier that can match the performance of the existing rule-based systems for Welsh POS and semantic taggers with as little human input as possible. b) To leverage existing language models such as word embedding created using unsupervised methods. Our work has demonstrated that these objectives can be achieved, although our results of a small-scale experiment can not be conclusive. The results obtained in this work compare favourably with those obtained from the existing rule-based models. We have also shown that, in a low resource setting, multi-task framework can also bring improvements to mono-lingual tasks, which is complementary to the previous findings from multilingual multi-task learning scenarios.

In our experiment, the neural network architecture was configured using pre-existing tools and frameworks, following suggestions from the literature. In future, we will focus on optimising the system parameters to improve the training efficiency and performance of the tagging models, as well as constructing larger training data.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA. USENIX Association.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947. Association for Computational Linguistics.

Hanan Aldarmaki and Mona Diab. 2015. Robust part-of-speech tagging of Arabic text. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 173–182. Association for Computational Linguistics.

Shabib AlGahtani and John McNaught. 2015. Joint Arabic segmentation and part-of-speech tagging. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 108–117. Association for Computational Linguistics.

Gor Arakelyan, Karen Hambardzumyan, and Hrant Khachatrian. 2018. Towards jointud: Part-of-speech tagging and lemmatization using recurrent neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 180–186. Association for Computational Linguistics.

Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2012. Joint learning of words and meaning representations for open-text semantic parsing. In *Artificial Intelligence and Statistics*, pages 127–135.

Jason Brownlee. 2017. How to use word embedding layers for deep learning with keras. https://machinelearningmastery.com/use-word-embedding-layers-deep-\learning-keras/.

Sven Buechel and Udo Hahn. 2018. Word emotion induction for multiple languages as a deep multi-task learning problem. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1907–1918. Association for Computational Linguistics.

Gyu Hyeon Choi, Jong Hun Shin, and Young Kil Kim. 2018. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

François Chollet et al. 2015. Keras. https://github.com/fchollet/keras.

Hamish Cunningham. 2002. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect Arabic POS Tagging: A CRF approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Jan Vium Enghoff, Søren Harrison, and Željko Agić. 2018. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201. Association for Computational Linguistics.

Rob van der Goot, Barbara Plank, and Malvina Nissim. 2017. To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 31–39. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Tobias Horsmann and Torsten Zesch. 2016. Ltl-ude $@$ empirist 2015: Tokenization and pos tagging of social media text. In *Proceedings of the 10th Web as Corpus Workshop*, pages 120–126. Association for Computational Linguistics.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Maarten Janssen, Josep Ausensi, and Josep Fontana. 2017. Improving POS Tagging in Old Spanish Using TEITOK. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 2–6. Linköping University Electronic Press.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 595–606. Association for Computational Linguistics.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Karlgren, Hans (ed.), Proceedings of 13th International Conference on Computational Linguistics*, volume 3, pages 168–173, Finland. Helsinki.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413. Association for Computational Linguistics.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809. Association for Computational Linguistics.

Shie Mannor, Dori Peleg, and Reuven Rubinstein. 2005. The cross entropy method for classification. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 561–568, New York, NY, USA. ACM.

Ana Marasović and Anette Frank. 2018. Srl4orl: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 583–594. Association for Computational Linguistics.

Tom McArthur and Thomas G McArthur. 1981. *Longman lexicon of contemporary English*. Longman London.

Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki, and Kotaro Funakoshi. 2018. A POS tagging model adapted to learner English. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 39–48. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Steve Neale, Kevin Donnelly, Gareth Watkins, and Dawn Knight. 2018. Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in Welsh. In *Proceedings of the 11th Edition of Language Resources and Evaluation Conference (LREC 2018) May 7-12, 2018.*, volume 3, pages 168–173, Japan. Miazaki.

Dat Quoc Nguyen and Karin Verspoor. 2018. An improved neural network model for joint POS tagging and dependency parsing. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 81–91. Association for Computational Linguistics.

Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'Egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1274. Association for Computational Linguistics.

Scott Piao, Paul Rayson, Dawn Archer, Francesca Bianchi, Carmen Dayrell, Mahmoud El-Haj, Ricardo-Mara Jimnez, Dawn Knight, Michal Ken, Laura Lfberg, Rao Muhammad Adeel Nawab, Jawad Shafi, Phoey Lee Teh, and Olga Mudraya. 2016. Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Scott Piao, Paul Rayson, Dawn Knight, and Gareth Watkins. 2018. Towards a Welsh Semantic Annotation System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, LREC2004*, pages 7–12.

Russell D. Reed and Robert J. Marks. 1999. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. MIT Press.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 173–183. Asian Federation of Natural Language Processing.

Serge Sharoff. 2018. Language adaptation experiments via cross-lingual embeddings for related languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resource Association.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Weiwei Sun and Xiaojun Wan. 2016. Towards accurate and efficient Chinese part-of-speech tagging. *Computational Linguistics*, 42(3):391–419.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979. Association for Computational Linguistics.

Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang. 2018. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 338–348. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Junxin Liu, Zhigang Yuan, and Yongfeng Huang. 2018. Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected lstm and multi-task learning. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 51–56. Association for Computational Linguistics.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical English. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328. Association for Computational Linguistics.

A Appendices

A.1 The Basic *CyTag* Tagset

The list of the coarse-grained (*basic*) *CyTag* part-of-speech categories used in this work is as shown below.

R han Ymadrodd	**CYTAG(ENG)**
Enw (Noun)	E (NN)
Y Fannod Benodol (Article)	YFB (ART)
Arddodiad (Preposition)	Ar PRE
Cysylltair (Conjunction)	Cys (CJN)
Rhifeiriau (Numeral)	Rhi (NUM)
Ansoddair (Adjective)	Ans (ADJ)
Adferf (Adverb)	Adf (ADV)
Berf (Verb)	B (VRB)
Rhagenw (Pronoun)	Rha (PRN)
Unigryw (Unique)	U UNI)
Ebychiad (Interjection)	Ebych (ITJ)
Gweddilliol (Others)	Gw (OTH)
Atalnodiad(Punctuation)	Atd (PUN)

A.2 The *USAS* Semantic Tagset

Below is a list and the descriptions of the USAS semantic top level categories:

Domain	Description
A	General and abstract terms
B	The body and the individual
C	Arts and crafts
E	Emotion
F	Food and farming
G	Government and public
H	Architecture, housing and the home
I	Money and commerce in industry
K	Entertainment, sports and games
L	Life and living things
M	Movement, location, travel and transport
N	Numbers and measurement
O	Substances, materials, objects and equipment
P	Education
Q	Language and communication
S	Social actions, states and processes
T	Time
W	World and environment
X	Psychological actions, states and processes
Y	Science and technology
Z	Names and grammar