# Whit’s the Richt Pairt o Speech: PoS tagging for Scots

**Harm F.K.M. Lameris and Sara Stymne**
Uppsala University
Uppsala, Sweden
harm.lameris.2403@student.uu.se, sara.stymne@lingfil.uu.se

## Abstract

In this paper we explore PoS tagging for the Scots language. Scots is spoken in Scotland and Northern Ireland, and is closely related to English. As no linguistically annotated Scots data were available, we manually PoS tagged a small set that is used for evaluation and training. We use English as a transfer language to examine zero-shot transfer and transfer learning methods. We find that training on a very small amount of Scots data was superior to zero-shot transfer from English. Combining the Scots and English data led to further improvements, with a concatenation method giving the best results. We also compared the use of two different English treebanks and found that a treebank containing web data was superior in the zero-shot setting, while it was outperformed by a treebank containing a mix of genres when combined with Scots data.

## 1 Introduction

Part-of-Speech (PoS) tagging is an essential Natural Language Processing (NLP) tool that is often seen as a first step in language analysis (Fang and Cohn, 2017) and can be useful for the analysis of large corpora, for instance in digital humanities (Hinrichs et al., 2019). For high resource languages, PoS tagging is occasionally considered to be a solved task, with recent state-of-the-art PoS taggers reaching high accuracies, such as Andor et al. (2016), who report an average accuracy of 97.5% across seven high-resource languages. However, these strong results are limited to languages and domains with large training data sets. State-of-the-art models commonly use Bi-directional Long Short Term Memory Recurrent Neural Network (Bi-LSTM RNN) architectures that rely on supervised learning methods. In order to obtain a high accuracy, large amounts of labelled data are required (Fang and Cohn, 2017). These resources are not available for low resource languages, and due to the high cost of annotation these are unfeasible to obtain (Garrette et al., 2013).

> Sae a thocht a’d gie ma pynt o view on aw ’is, an ’e story ahint whit caused iz tae stairt the editathon muivement an whit a hink anent the hale hing. On ’e 25t August 2020, a 4chan an Reddit post wis pit oot bi Ryan Dempsey, unner the uisername ’Ultach’, fae Ulster, anent an American Wikipedia admin, whase name a willnae mention tae jouk thaim gittin ony mair potential harassment (lat’s juist caw thaim Admin fae nou), an hou thay contrı́butit near hauf o aw the airticles on ’e Wikipedia, an hou the ither tap fower contrı́butors wis American an aw.

Figure 1: Scots excerpt, originally from Clark et al. (2020)

In this paper we focus on the Scots language, a language closely related to English, which is spoken in Scotland and Northern Ireland. Scots lacks linguistic resources and tools. In this paper we take a first step towards overcoming this, by examining PoS tagging for Scots. We investigate how well an English PoS tagger works on Scots, and try to improve on this by using a very small amount of Scots annotated data during training. As part of the study, a small sample of Scots texts were manually PoS tagged. An example of Scots is shown in Figure 1.

While statistics-based PoS taggers are occasionally used for truly low-resource languages (Agić et al., 2015), data scarcity can also successfully be alleviated using transfer learning (Lin et al., 2019). PoS tagging using cross-lingual transfer has proved effective on low-resource simulations of high-resource European languages, and even disparate non-Indo-European languages (Fang and Cohn, 2017). Yet, few attempts have been made to use cross-lingual models to tag European minority



*Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48
April 20, 2021 

languages such as Scots, despite the existence of closely related high-resource languages. One exception is Magistry et al. (2019) who focused on PoS tagging for three regional French languages.

We focus on investigating PoS tagging for Scots. First we want to investigate a zero-shot setting, to see how well a PoS tagger trained for English works on Scots, additionally comparing the performance of two different English data sets, the GUM treebank (Zeldes, 2017) and the EWT treebank (Silveira et al., 2014). Secondly, we annotate a small amount of data, 1040 tokens, in order to investigate how much could be gained from adding even a very small amount of data. We compared different methods for combining the Scots and English data, with zero-shot tagging and training only on a tiny amount of Scots data. The main contributions of this work are:

- The first attempt at creating a PoS-tagged Scots corpus. While the size of the corpus is small, it features multiple Scots dialects, and texts from multiple domains, and could therefore easily be expanded upon. This corpus is publicly available at `https://github.com/Hfkml/pos-tagged-scots-corpus`.
- An investigation of combining large English and small Scots data for PoS tagging, leading to the the first NLP tool of its kind developed for the Scots language, including an investigation of two different English training sets.

We believe that our study presents a method that can be used to create NLP tools for many other minority languages with scant resources that are closely related to a high-resource language, such as Low Saxon and Occitan. These NLP tools present a valuable lifeline that enables increased usage of minority languages in digital settings.

## 2 Scots

Scots is a West-Germanic language spoken by 1.5 million people.[1] While recognized as a language by the Scottish government, Scots' status as a language (Sebba, 2019) is disputed, with 64% of respondents in a 2010 Scottish Government survey seeing Scots as a way of speaking rather than a language.[2]

Scots diverged from English in the 13$^{th}$ century and use of Scots was widespread for both government and literary purposes until the 16$^{th}$ century (Kay, 1988). Since the Union of the Crowns, Scots has undergone a process of Anglicization, and has largely been superseded by Scottish English with which it is spoken on a bipolar spectrum. Currently, Scots is a low-resource language. Scots texts are readily available, but parallel corpora or gold-standard annotations are lacking. This severely complicates the analysis of Scots. In recent years, Scottish identity has experienced a revival, and with it interest in the Scots language. Scots is reemerging as an active language, especially amongst the young rural population of Scotland and amongst Scottish nationalists (Lemeshchenko-Lagoda, 2019). Although this has generated academic interest in the fields of phonology (Lawson et al., 2019) and socio-linguistics (Shoemark et al., 2017a,b), no NLP-tools are available specifically for Scots. The only Scots NLP efforts are two sets of word embeddings, namely Fasttext embeddings (Joulin et al., 2016) and Polyglot Word Embeddings (Al-Rfou et al., 2013).

There are several complications relating to NLP for Scots. Despite attempts at standardization by the *Report & Recommends o the Scots Spellin Comatee* (Allan et al., 1998), Scots does not have a *de jure* standard. As there is no official orthography, major variations in spelling exist, mostly depending on the written dialect of Scots. Additionally, almost no parallel English-Scots texts are publicly available, which excludes approaches relying on such data. Scots grammar and morphology, while similar to and not necessarily more or less complex, differs from English grammar and morphology. Some specific features of Scots is its irregular plural forms for nouns, e.g. *cauf/caur* for "calf/calves", and frequent use of progressive forms, e.g. *who's wantin tae go oot wi you* — "who wants to go out with you"[3],

## 3 Related Work

### 3.1 Non-standard Language PoS Tagging

State-of-the-art PoS taggers use a Bi-LSTM RNN configuration, often with a Conditional Random Field, e.g. in Akbik et al. (2018), who achieved an accuracy of 97.85% on the English-language

[1] `https://www.gov.scot/policies/languages/scots/`

[2] *Public Attitudes Towards the Scots Language:* `https://www.webarchive.org.uk/wayback/archive/3000/https://www.gov.scot/Resource/Doc/298037/0092859.pdf`

[3] Example from Anderson et al. (2007)

Wall Street Journal corpus. Most state-of-the-art taggers are trained on newswire articles, and perform poorly on non-standard language or informal language. Gui et al. (2017), for example noticed a performance drop from 97.0% to 73.4% for tweets and other out-of-domain data for a model trained on the Wall Street Journal. This drop was mitigated by training a novel neural network that used an adversarial discriminator on out-of-domain labeled data, unlabeled in-domain data, and labeled in-domain data to achieve 90.92% accuracy.

This performance drop is even greater when examining non-standard English such as African American Vernacular English. In Jørgensen et al. (2016), the performance of the newswire-trained tagger decreased to 63% for AAVE data. The researchers used word representations learned from unlabelled data, as well as partially labelled data generated from lexicons in combination with ambiguous supervision that weighs the probabilities from the partially labelled data to overcome the domain gap.

Similarly, historical texts contain spelling and grammatical constructions that deviate from modern standard language. Moon and Baldridge (2007) achieved an accuracy of 80.5% on Biblical texts and an accuracy 63.9% on other Middle English texts bootstrapping a maximum entropy tagger using alignment projections from the Wycliffe's Middle English Bible and the Modern English NET Bible.

Another method that has shown promise in PoS tagging out-of-domain data, especially historical data, is word normalization. Normalization of non-standard text improves tagging accuracy by lowering the number of out-of-vocabulary words which are substantially more difficult to tag compared to in-vocabulary words. This approached has successfully been applied for several languages including Slovenian (Zupan et al., 2019) and German (Bollmann, 2013). Zupan et al. (2019) further found that for small annotation efforts manual normalization was preferable to manual PoS tagging, while manual PoS tagging was useful in larger annotation projects.

While Scots has been analyzed as non-standard English in the field of sentiment analysis (Shoemark et al., 2017b), this approach is not straightforward for PoS tagging, as the approaches used to fine-tune the models for domain adaptation for PoS tagging in e.g. Gui et al. (2017); Jørgensen et al. (2016) rely on annotated corpora that are unavailable for Scots.

### 3.2 Low-Resource PoS tagging

Due to the complete absence of annotated data for Scots, its situation in terms of NLP is similar to the situation of many low-resource languages. Multiple methods have been used to alleviate data-scarcity, generally by using multi-task learning or cross-lingual transfer. While this project focuses on cross-lingual transfer learning, multi-task learning bears great relevance to low-resource PoS tagging and so will be briefly discussed.

#### 3.2.1 Multi-task learning

Kann et al. (2018) built on the work of Plank et al. (2016) to hierarchically combine a recurrent network and a character-based sequence-to-sequence model with a corpus of only 478 annotated tokens in the low-resource language. The researchers jointly trained for PoS tagging and alternately unsupervised lemmatization, character-based autoencoding, and character-based random string autoencoding to close the gap on the state-of-the-art PoS tagger by 43%.

#### 3.2.2 Cross-lingual tagging

Yarowsky et al. (2001) first showed the benefit of language transfer, a method in which multiple source languages are used to create a the target language model. Cross-lingual transfer especially improves results for closely related languages, highlighting the applicability for this study. Das (2011) expanded on this study by introducing an graph-based method using unsupervised label propagation that projects syntactic information across languages by constructing a bilingual graph over word-types resulting in a 16.7% error reduction compared to regular HMMs.

Cross-lingual PoS tagging utilizes annotated data available from high resource languages (Kim et al., 2017). This data is supplemented with e.g. parallel corpora to provide linguistic knowledge to aid the transfer. Before the neural era, cross-lingual PoS tagging was applied in low-resource PoS tagging by Agić et al. (2015), who used word alignment from Bible verses in 100 languages in combination with a small amount of manually tagged data to create PoS taggers. For the 25 languages for which test data was available, the models showed an improvement of 20-30% compared to unsupervised models.

Since neural networks perform especially poorly in low-data scenarios, cross-lingual transfer learning is indispensable in the creation of NLP tools such as PoS taggers for low-resource languages. As the aforementioned parallel corpora necessary for cross-lingual tagging are absent, cross-lingual transfer learning can be used. In cross-lingual transfer learning the lower layers of a hierarchical model share knowledge, such as parameters, between the differing input domains.

Yang et al. (2016) explored cross-lingual transfer solely focussed on model transfer between similar languages. Character embeddings were used in order to take advantage of morphological and lexical similarities. The researchers note, however, that fewer parameters were shared for cross-lingual transfer than for cross-domain transfer. Kim et al. (2017) built on Yang et al. (2016)'s research by using knowledge transfer in the Bi-LSTM layers that have as input the character and word embeddings from both a common (shared) Bi-LSTM and a private BiLSTM, as well as language adversarial training. Their best-performing model achieves an average of 88.37% for Germanic languages.

Magistry et al. (2019) attempted to create PoS taggers for three regional French languages. They had no training data available for these languages and relied on data for related high-resource languages, focusing on strategies based on delexicalization and transposition. They use a BiLSTM-CRF tagger, which overall gave better results than two alternative taggers.

# 4 Data

In this section we describe the data used. We use two existing English data sets, and create a small new set for Scots. The tags used for all the data are the Universal UD PoS tags (Zeman et al., 2019). This tagset consists of 17 tags and only uses basic lexical categories that are seen as applicable to all languages. Examples of categories are: Noun (NOUN), adjective (ADJ), and adposition (ADP).

## 4.1 English Data

For the English data, we used two data sets: the GUM Treebank (Zeldes, 2017) and the EWT Treebank (Silveira et al., 2014) from the Universal Dependencies project, version 2.5. Both data sets came pre-tagged using the Universal PoS tags from Universal Dependencies (Zeman et al., 2019) and pre-split into a training, development, and test set. The training set of the GUM Treebank contains 73513 tokens and the training set of the EWT Treebank contains 217152 tokens. The GUM Treebank contains multiple text types, including Wiki data, fiction and news, and the EWT Treebank contains web data including blogs and reviews.

## 4.2 Scots data

No pre-tagged PoS data were available for Scots, so instead the untagged Scots language data were harvested from the Scottish Corpus of Text and Speech (SCOTS) (Anderson et al., 2007), and Scots-language blog Mak Forrit (Clark et al., 2020). SCOTS is a corpus containing 949 texts which were written between 1945 and 2011. It includes texts in Scottish English, Scots, and Scottish Gaelic. The Scots data had to be classified as Broad Scots for it to be included in this project, which was achieved by searching for Scots function words such as *fae* — "from" and *tae* — "to". Mak Forrit is written in a standardized version of Scots. A final selection was arbitrarily made by picking a variety of sentences from the combination of sources. Multiple dialects of Scots, including Doric and West-Central Scots are represented. This results in multiple orthographies being present in the Scots language data. In total, 37 sentences with 1040 tokens were extracted which were divided into a training set of 536 tokens and a test set of 504 tokens. This ratio was chosen to maximize the amount of training and test data available given the time-scope of the project. Due to the small amount of data, we do not have a separate development set, and for the combined models we use the training data also as a development set.

While the option of normalization was considered, we decided on manually PoS tagging Scots as a result of the large number of spelling differences between Scots and English compared to Zupan et al. (2019). The Scots language data were manually tagged by the first author, a non-native Scots speaker with training in linguistics and computational linguistics. The data was tagged on two separate occasions, approximately one month apart so that intra-annotator agreement could be measured. The intra-annotator agreement was 96.5%. The two separate taggings were then consolidated into the final version, by the same annotator. Five retagged tokens were considered outright mistakes due to misinterpretations, for example in: *Kinnin the English airmy wad hae need o great stores*

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| The | problem | wis | , | thair | Scots | wisnae | the | best | , | an | thay | war | relyin | on | a | dictionar | tae | help | thaim | oot | . |
| DET | NOUN | VERB | PUNCT | DET | PROPN | VERB | DET | ADJ | PUNCT | CCONJ | PRON | AUX | VERB | ADP | DET | NOUN | PART | VERB | PRON | ADV | PUNCT |

Figure 2: Example of a tagged Scots sentence

*o supplies* —"Knowing the English army would need great stores of supplies", need was tagged as a VERB due to the proceeding verbs that appeared to be auxiliary verbs. The word *o* "of", however makes the construction more akin to "Knowing the English army would have need of great stores of supplies", making *need* a NOUN and was therefore tagged as such. Other differences were attributed to slight ambiguity. In: *The ploy wasnae faur awa fae wirkin* — "The ploy nearly worked", the tokens *faur* "far" and *awa* "away" were initially tagged as ADJ, as *wasnae* — "was not" is a frequently used copula verb that is often followed by an adjective phrase. *Awa* only has ADV entries in *the Dictionary of the Scots Language*,[4] however, and *faur* modifies *awa* making *faur* another ADV. *Faur awa* was therefore reinterpreted as an adverbial phrase and both words were tagged ADV. An example of a tagged Scots sentence can be found in Figure 2.

### 4.3 Word Embeddings

Both the Scots and English language data were supplemented with Polyglot Word Embeddings (Al-Rfou et al., 2013) which are cross-lingual word embeddings available for over 100 languages trained on Wikipedia. While the Scots Wikipedia has attracted some controversy recently due to the questionable quality of the Scots featured on its pages[5], both sets of word embeddings available for Scots were trained on the Scots Wikipedia. Polyglot Word Embeddings were chosen over Fasttext word embeddings as its usage was recommended for the PoS tagger by Plank et al. (2016). No quality assessment had ever been done on these embeddings, however, and thus the performance of the embeddings could not be verified in advance. We leave a further investigation of word embedding quality to future work.

## 5 Models & Experiments

All models were run on the Bi-LSTM PoS tagger from Plank et al. (2016) running on the standard settings, bar the addition of the polyglot embeddings (Al-Rfou et al., 2013). Plank et al. (2016)'s tagger uses a Bi-LSTM for character embeddings, the output of which is concatenated with the word embeddings. These embeddings are the input for the BiLSTM which has the softmax for the PoS tags as an output layer.

The network used 100-dimensional character embeddings, 100 dimensional polyglot embeddings, 100-dimensional LSTM hidden states and had one hidden-layer for the encoder and the decoder. During training, stochastic gradient descent was applied with a dropout rate of 0.25 as well as a sigma of 0.2 Gaussian noise. The network used a softmax activation function in the decoder to generate the tags.

All models were trained for 20 iterations with an early stop of four iterations. These hyperparameters were chosen to prevent training from finishing pre-maturely, although no model trained for more than 19 iterations.

For our experiments, we first trained monolingual baseline models for English and Scots. Then we explored three ways of combining data across languages. In all cases we also compare the results with the two English corpora.

### 5.1 Model 1: Zero-shot

To compare the effect of the different types of model- and data transfer to the monolingual models, we first obtained a zero-shot baseline by training models on the two English training and development sets, and testing the model on the Scots test set. To examine the accuracy of the zero-shot models for English, it was first tested on the English test set from the GUM (Zeldes, 2017) and EWT (Silveira et al., 2014) obtaining an accuracy of of 95.7% and 95.1% respectively.

### 5.2 Model 2: Scots Only

The second monolingual baseline model was only trained on the Scots training data from the SCOTS corpus (Anderson et al., 2007) and Mak Forrit (Clark et al., 2020). The training set, 536 tokens, was, therefore, much smaller than for the other models, and it did not feature a development set. The purpose of this model was to elucidate fur-

[4] Scottish Language Dictionaries Ltd., `https://dsl.ac.uk/`

[5] https://www.theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia

ther where transfer learning could bring benefits for small data sets.

### 5.3 Model 3: Concatenated Data

The concatenated data model used a joint-training method. The English and Scots training data were simply concatenated. The Scots training set was additionally also used as the development set during training, as empirical evidence showed that using the English evaluation set resulted in overfitting on the English part of the data set and had poorer results.

### 5.4 Model 4: Mixed training

In the mixed-training model, the model was first allowed to train to convergence on the English data set. After convergence, an even distribution of the Scots and the English data was used to further fine-tune the parameters. This method was used since it had proved beneficial in improving Neural Machine Translation in low-resource settings (Dabre et al., 2019; Chu and Wang, 2018).

### 5.5 Model 5: Pure fine-tuned model

In pure fine-tuning (Dabre et al., 2019), the model was first allowed to fully converge on the English training set. These parameters were then transferred to the Scots model by training the converged model on the small Scots training set, to allow convergence on the Scots set. This model used the Scots development set for pre-training and no development set for fine-tuning.

## 6 Results

Table 1 shows the main results. Training on only the small Scots data gave an accuracy of 60.5%, which is higher than the zero-shot results trained on much larger English data. All cross-lingual models performed remarkably better than the zero-shot models, seeing an improvement in accuracy of 16.9-23.1% for the GUM data set, and an improvement of 5.2-9.9% for the EWT data set. While for both data sets the zero-shot models obtain a low accuracy, the difference is starker for the GUM data set, which sees the best performance for the cross-lingual models, while it is out-performed by EWT for the zero-shot model. A possible explanation for this is the genres of the data. The GUM data set consists of largely curated data, for instance Wikipedia and news, with some non-curated Reddit data, while the EWT data set largely contains non-curated user-generated blogs. This has, most likely, caused the EWT zero-shot model to be more robust. The low accuracy obtained by the zero-shot models is indicative of a large discrepancy between the training and the test data. This strongly suggests that Scots, for NLP purposes, cannot merely be treated as out-of-domain English data. The zero-shot accuracy is much lower and the difference in accuracy between the zero-shot model and the cross-lingual models is a lot larger than in, e.g. Gui et al. (2017)'s models for the PoS tagging of out-of-domain data that obtained an accuracy of 73.4%, and than the 63% accuracy for AAVE-like language in Jørgensen et al. (2016).

| | GUM | EWT |
|---|---|---|
| Scots only | 60.5 | |
| Zero-shot | 42.4 | **55.7** |
| Concatenated | **67.4** | 65.6 |
| Mixed training | 62.2 | **63.8** |
| Pure fine-tune | **63.5** | 60.9 |

Table 1: The accuracy results of each model. Boldface indicates the best model of each type.

As can be seen in Table 1, the cross-lingual models display a slight to moderate improvement on the Scots-only model. The similarity in accuracy scores between the Scots-only and the cross-lingual models could be a sign that the domain overlap between the Scots training and test set was too great. Even though the annotated Scots corpus was created from sources that spanned multiple domains, the division between the training- and test data was made randomly, meaning that segments of the training- and test set could have been taken from the same text. The best-performing model for both data sets were the concatenated models. The concatenated models appear to benefit from first converging to the English training data, with the parameters then being adjusted to the Scots training data. The other models show more frequent under- and overfitting. The GUM Mixed-training mode for example on five occasions predicted an X tag, for example for *faa* "fall" and *e* "he" or "the". The X tag occurs in the Scots training data, as it contains some Catalan words. In the GUM data set the X symbol is nonetheless more sparse. Conversely, *wisnae* "was not" was tagged as a NOUN, despite occurring in the training data as a VERB and AUX.[6]

[6] In this work we have not segmented words not sep-

| | Scots Only | | Zero-shot GUM | | Zero-shot EWT | | Concatenation GUM | | Concatenation EWT | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| INTJ | N/A | 0 | **100** | *100* | 33.0 | *100* | 0 | 0 | **50.0** | **100** |
| PUNCT | 100 | 94.0 | **100** | 100 | 89.3 | 100 | **100** | 100 | 89.3 | 100 |
| NOUN | 61.5 | 64.9 | 33.7 | 81.1 | **47.2** | **90.5** | **64.3** | **87.8** | 55.9 | 83.8 |
| PROPN | 50.0 | 8.3 | 17.9 | **100** | **26.2** | 91.7 | **50.0** | **91.7** | 38.9 | 58.3 |
| ADP | 79.6 | 78.0 | **100** | 32.0 | 85.3 | **58.0** | 72.3 | 68.0 | **77.8** | **70.0** |
| DET | 86.6 | 75.0 | **100** | 48.1 | 90.6 | **55.7** | 78.0 | **75.0** | **92.3** | 69.2 |
| ADJ | 36.4 | 16.0 | 13.0 | 12.0 | **25.8** | **32.0** | **50.0** | **32.0** | 35.0 | 28.0 |
| VERB | 52.2 | 61.2 | **52.2** | 31.0 | 52.1 | **35.2** | 55.3 | 43.7 | **61.2** | **57.8** |
| PRON | 50.0 | 69.4 | 17.9 | **100** | **26.2** | 91.7 | **80.0** | 65.3 | 78.0 | 65.3 |
| PART | 55.6 | 38.5 | 20.0 | 7.7 | **40.0** | **15.4** | **81.8** | **69.2** | 40.0 | 15.4 |
| ADV | 100 | 15.2 | 26.6 | 24.2 | **42.4** | **42.4** | 47.4 | **54.5** | **48.4** | 45.5 |
| AUX | 20.3 | 59.1 | 0 | 0 | **18.2** | **9.1** | 47.8 | 50 | **57.1** | **54.5** |
| CCONJ | 66.7 | 33.3 | 60.0 | 25.0 | **71.4** | **41.7** | 81.8 | **75.0** | **100** | 58.3 |
| SCONJ | N/A | 0 | 0 | 0 | **25.0** | **20.0** | **66.7** | 40.0 | 33.3 | 40.0 |
| NUM | 100 | 28.6 | 0 | 0 | **50.0** | **57.1** | **66.7** | 28.6 | 36.4 | **57.1** |

Table 2: The per-tag precision and recall for the models. For each model pair, boldface indicates the best performing model for each tag.

Table 2 shows the precision and recall for each tag for Scots only and for the zero-shot and the best cross-lingual models: the concatenated models. Comparing the GUM concatenated model to the baselines, one can see that the model shows improvements compared to both the Scots-only and zero-shot baselines, especially when it comes to nouns (NOUN), adjectives (ADJ), pronouns (PRON), particles (PART), coordinating conjunctions (CCONJ), and subordinating conjunctions (SCONJ) for which both precision and recall increased. Interjections (INTJ) were surprisingly not predicted correctly, achieving 0% for both precision and recall, despite the zero-shot model achieving 100% precision and recall. Apart from the precision for INTJ, adpositions (ADP), determiners (DET), and the recall for INTJ, and proper nouns (PROPN), the model improved for every PoS category. Some surprising errors remain, however. The verb or auxiliary verb *are* was consistently tagged as an adverb (ADV), despite occurring as such in neither the English nor the Scots data. Other (auxiliary) verbs, such as *micht* "might" and *'ve* were also tagged as adverbs. Many sentence-initial words such as *Syne* "since", *Reading* and *They* were tagged as PROPN, despite the two latter words also occurring in both languages. Other incorrect predictions are more similar to what one would expect from English PoS taggers for English, such as the labelling of VERB as auxiliary verb (AUX) and vice-versa, e.g. in *Ye jist had tae **be** maist aafa partickler* "You just had to be most awfully particular", be was labelled as an AUX instead of a VERB.

The differences between the EWT concatenated model and the EWT zero-shot baseline are much smaller, as can be seen in Table 2. While there are some improvements in both precision and recall for e.g. DET, VERB, ADV, AUX, CCONJ,and SCONJ, the difference generally not as stark as for the GUM concatenated model. Many other categories see a drop in either precision, such as INTJ, ADP, VERB, and PRON or recall, such as NOUN, PROPN, and PRON. Compared to the Scots-only baseline, the EWT concatenated model shows consistent improvement for INTJ, PRON, CCONJ and SCONJ, while also notably improving the precision for VERB and the recall for NOUN.

The EWT concatenated model also contains unexpected errors. Despite occurring in both English and Scots, auxiliary verb *can* was tagged as both an ADV and a NOUN. Another error can possibly, though not necessarily, be attributed to the quality

arated by whitespace. An option would be to separate a word like wisnae into wis+nae, according to UD guidelines (https://universaldependencies.org/u/overview/tokenization.html). However, this would lead to a call for a tokenizer especially adapted to Scots, which is currently not available.

of the word embeddings. *E* "he" or "the", is consistently tagged as NUM, despite it not having this tag in the training data, although this could also be due to the model parameters.

Comparing the GUM and EWT concatenated models in table 4, one can see that both models generally have comparable per-tag precision and recall, and that neither model scores consistently better than the other, with each model outperforming the other in both precision and recall in four PoS categories. The main reason for the slightly better performance of the GUM concatenated model appears to stem from better precision in PUNCT, NOUN, ADJ, and PART, all frequently occurring tags although the performance for VERB is much better for the EWT concatenated model.

## 7 Conclusion

We have annotated a small amount of Scots data with PoS tags, and shown how this data can be used to improve tagging for Scots. The best results were obtained by a simple concatenation model, where a large English data set was concatenated with a small amount of Scots data. We also found that while the English data EWT data set containing web data, performed better than the GUM Treebank with mostly edited data in a zero-shot setting, the GUM Treebank, trained on a mix of genres performed better in the cross-lingual setting.

While greatly improving on the zero-shot model and slightly improving on the Scots-only model, the cross-lingual models for both data sets scored lower than other studies that examine low-resource settings for PoS tagging. Longer training, higher dimensional embeddings and modifications to the model architecture such as private LSTMs and the use of adversarial training could be made. Kim et al. (2017) achieve higher accuracies for all languages using 320 tokens per language, showing the efficacy of the combination of these factors. An improvement on the model architecture can be made as well by, rather than weighting English and Scots data equally in the loss function as was done in these experiments, weighting the Scots data more heavily in the loss function of the Bi-LSTM in order to prevent underfitting on the Scots data.

Another possible cause of these results is the previously mentioned questionable quality of the word embeddings, and we think it would be interesting to train Scots word embeddings on verified Scots data, to be able to compare to the current embeddings trained on Wikipedia data with questionable quality. We also think it would be interesting to explore text normalization strategies for Scots.

Lastly and perhaps most importantly, an increase in the size of the Scots data set and an interest in developing NLP tool for minority languages is required to improve performance of the tagger. Other models that examine cross-lingual transfer, such as Yang et al. (2016) use 900 training tokens minimally, nearly double the tokens used in this paper. Yang et al. (2016) also note that improvements are smaller for cross-lingual transfer than in cross-domain training, strengthening the case for the treatment of Scots as a language in terms of the required NLP resources. Combined with the low accuracy for the zero-shot model this indicates the need for additional resources for Scots, such as dictionaries and parallel corpora.

## References


Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning PoS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Alasdair Allan, Andy Eagle, John Law, J Derrick McClure, George Philp, Iseabail Macleod, Liz Niven, David Purves, and John Tait. 1998. Report & recommends o the Scots spellin comatee.

Jean Anderson, Dave Beavan, and Christian Kay. 2007. Scots: Scottish corpus of texts and speech. In Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, editors, *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 17–34. Palgrave Macmillan UK, London.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association* for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany. Association for Computational Linguistics.

Marcel Bollmann. 2013. POS tagging for historical texts with sparse training data. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 11–18, Sofia, Bulgaria. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Thomas Clark, Alistair Heather, James McDonald, Jamie Smith, Elizabeth Thoumire, and Antonia Uri. 2020. Mak Forrit. Blog: `https://www.makforrit.scot`.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

Amitava Das. 2011. PsychoSentiWordNet. In *Proceedings of the ACL 2011 Student Session*, pages 52–57, Portland, OR, USA. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593.

Dan Garrette, Jason Mielens, and Jason Baldridge. 2013. Real-world semi-supervised learning of PoS-taggers for low-resource languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–592.

Tao Gui, Qi Zhang, Haoran Huang, Minlong Peng, and Xuan-Jing Huang. 2017. Part-of-Speech tagging for twitter with adversarial neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2420.

Erhard Hinrichs, Marie Hinrichs, Sandra Kübler, and Thorsten Trippel. 2019. Language technology for digital humanities: introduction to the special issue. *Language Resources & Evaluation*, 53:559–563.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a PoS tagger for AAVE-like language. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1115–1120.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Katharina Kann, Johannes Bjerva, Isabelle Augenstein, Barbara Plank, and Anders Søgaard. 2018. Character-level supervision for low-resource PoS tagging. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 1–11.

Billy Kay. 1988. *The mither tongue*. Grafton, London, United Kingdom.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for PoS tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Eleanor Lawson, Jane Stuart-Smith, and Lydia Rodger. 2019. A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes. *The Journal of the Acoustical Society of America*, 146(6):4363–4381.

Viktoriia Lemeshchenko-Lagoda. 2019. What is Scots? In Halyna Matiukha and Maria Karpinska, editors, *Novi perspektyvy v angliys'kiy filolohiyi ta navchannya angliys'kiy movi*. TOB, Kolor Print, Melitopol.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset. 2019. Exploiting languages proximity for Part-of-Speech tagging of three French regional languages. *Language Resources & Evaluation*, 53:865–888.

Taesun Moon and Jason Baldridge. 2007. Part-of-Speech tagging for middle English through alignment and projection of parallel diachronic texts. In *Proceedings of the 2007 joint conference on empirical methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 390–399.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Mark Sebba. 2019. Named into being? language questions and the politics of Scots in the 2011 census in Scotland. *Language Policy*, 18(3):339–362.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2017a. Topic and audience effects on distinctively scottish vocabulary usage in twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017b. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1239–1248, Valencia, Spain. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2016. Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:* 1703.06345v1.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. Technical report, Johns Hopkins University Baltimore, MD Dept. of Compute Science.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, et al. 2019. Universal dependencies 2.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Katja Zupan, Nikola Ljubešić, and Tomaž Erjavec. 2019. How to tag non-standard language: Normalisation versus domain adaptation for Slovene historical and user-generated texts. *Natural Language Engineering*, 25(5):651–674.