Joint Part-of-Speech and Language ID Tagging for Code-Switched Data

Victor Soto
Department of Computer Science
Columbia University
New York, NY 10027
vsoto@cs.columbia.edu

Julia Hirschberg
Department of Computer Science
Columbia University
New York, NY 10027
julia@cs.columbia.edu

Abstract

Code-switching is the fluent alternation between two or more languages in conversation between bilinguals. Large populations of speakers code-switch during communication, but little effort has been made to develop tools for code-switching, including part-of-speech taggers. In this paper, we propose an approach to POS tagging of code-switched English-Spanish data based on recurrent neural networks. We test our model on known monolingual benchmarks to demonstrate that our neural POS tagging model is on par with state-of-the-art methods. We next test our code-switched methods on the Miami Bangor corpus of English-Spanish conversation, focusing on two types of experiments: POS tagging alone, for which we achieve 96.34% accuracy, and joint part-of-speech and language ID tagging, which achieves similar POS tagging accuracy (96.39%) and very high language ID accuracy (98.78%). Finally, we show that our proposed models outperform other state-of-the-art code-switched taggers.

1 Introduction

Code-switching (CS) is the phenomenon by which multilingual speakers switch between languages in written or spoken communication. For example, a English-Spanish speaker might say "El teacher me dijo que Juanito is very good at math." CS can be observed in various linguistic levels: phonological, morphological, lexical, and syntactic and can be classified as *intra-sentential* (if the switch occurs within the boundaries of a sentence or utterance), or *inter-sentential* (if the switch occurs between two sentences or utterances). The importance of developing NLP technologies for CS data is immense. In the US alone there is an estimated population of 56.6 million Hispanic people (US Census Bureau, 2014), of which 40 million are native speakers (US Census Bureau, 2015). Most of these speakers routinely code-switch. However, very little research has been done to develop NLP approaches to CS language, due largely to the lack of sufficient corpora of high-quality annotated data to train on. Yet CS presents serious challenges to all language technologies, including part-of-speech (POS) tagging, parsing, language modeling, machine translation, and automatic speech recognition, since techniques developed on one language quickly break down when that language is mixed with another.

One of Artificial Intelligence's ultimate goals is to enable seamless natural language interactions between artificial agents and human users. In order to achieve that goal, it is imperative that users be able to communicate with artificial agents as they do with other humans. In addition to such real time interactions, CS language is also pervasive in social media (David, 2001; Danet and Herring, 2007; Cárdenas-Claros and Isharyanti, 2009). So, any system which attempts to communicate with these users or to mine their social media content needs to deal with CS language.

POS tagging is a key component of any Natural Language Understanding system and one of the first researchers employ to process data. As such, it is crucial that POS taggers be able to process CS content. Monolingual POS taggers stumble when processing CS sentences due to out-of-vocabulary words in one language, confusable words that exist in both language lexicons, and differences in the syntax of the two languages. For example, when running monolingual English and Spanish taggers on the CS English-Spanish shown in Figure 1, the English tagger erroneously tagged most Spanish

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 1–10
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

Words:	Ella	lo	había	leído	when	she	was	in	third	grade
Translation:	*She*	*it*	*had*	*read*	-	-	-	-	-	-
Gold:	PRON	PRON	AUX	VERB	SCONJ	PRON	VERB	ADP	ADJ	NOUN
EN Tagger:	NOUN	ADV	NOUN	VERB	ADV	PRON	VERB	ADP	ADJ	NOUN
ES Tagger:	PRON	PRON	AUX	VERB	PROPN	PROPN	PROPN	ADP	X	PROPN
EN+ES Tagger:	PRON	PRON	AUX	VERB	ADV	PRON	VERB	ADV	ADJ	NOUN
CS Tagger:	PRON	PRON	AUX	VERB	SCONJ	PRON	VERB	ADP	ADJ	NOUN

Figure 1: Example of an English-Spanish code-switched sentence. The figure shows the original code-switched sentence, English translations of each token, gold POS tags and the tagging output of an English tagger, a Spanish tagger, a tagger trained on English and Spanish sentences, and a tagger trained on a corpus of code-switched sentences, in that order. Errors made by each tagger are underlined.

tokens, and similarly the Spanish tagger mistagged most English tokens. A tagger trained on monolingual English and Spanish sentences (EN+ES tagger) fared better, making only two mistakes: on the word "when", where the switch occurs (confusing the subordinating conjunction for an adverb), and the word "in" (which exists in both vocabularies). A tagger trained on CS instances of English-Spanish, however, was able to tag the whole sentence correctly.

In this paper, we present a comprehensive study of POS tagging for CS utterances that includes the following: a) use of a state-of-the-art bidirectional recurrent neural network b) use of a large CS English-Spanish corpus annotated with high-quality labels from the Universal POS tagset; c) extensive analyses of the performance of our taggers on monolingual and CS sentences; d) study of the performance of a tagger trained on the subset of the monolingual sentences of the CS corpus (in-genre baseline); e) examination of the effect of language identifiers both as feature inputs and for joint language identification and POS tagging; and f) comparison to state-of-the-art taggers for code-switching on the same corpus.

2 Related Work

A variety of tasks have been studied in CS data. For language identification (LID), Rosner and Farrugia (2007) proposed a word-level Hidden Markov Model and a character-level Markov Model to revert to when a word is out-of-vocabulary, and tested these on a corpus of Maltese-English sentences, achieving 95% accuracy. Working on a Bengali-Hindi-English dataset of Facebook posts, Barman et al. (2014) employed classifiers using n-gram and contextual features to obtain 95% accuracy.

In the first statistical approach to POS-tagging on CS data, Solorio and Liu (2008) collected the Spanglish corpus, a small set of 922 English-Spanish sentences. They proposed several heuristics to combine monolingual taggers with limited success, achieving 86% accuracy when choosing the output of a monolingual tagger based on the dictionary language ID of each token. However, an SVM trained on the output of the monolingual taggers performed better than their oracle, reaching 93.48% accuracy. On the same dataset, Rodrigues (2013) compared the performance of a POS-tagger trained on CS sentences with a dynamic model that switched between taggers based on gold language identifiers; they found the latter to work better (89.96% and 90.45% respectively). Note, however, that the monolingual taggers from (Solorio and Liu, 2008) were trained on other larger corpora, while all the models used in (Rodrigues, 2013) were trained on the Spanglish corpus.

Jamatia et al. (2015) used CS English-Hindi Facebook and Twitter posts to train and test POS taggers. They found a Conditional Random Field model to perform best (71.6% accuracy), and a combination of monolingual taggers similar to the one in (Solorio and Liu, 2008) achieved 72.0% accuracy. Again using Hindi-English Facebook posts, Vyas et al. (2014) ran Hindi and English monolingual taggers on monolingual chunks of each sentence. Sequiera et al. (2015) tested algorithms from (Solorio and Liu, 2008) and (Vyas et al., 2014) on the Facebook dataset from (Vyas et al., 2014) and the Facebook+Twitter dataset from (Jamatia et al., 2015), and found that (Solorio and Liu, 2008) yielded better results. Similarly, Barman et al. (2016) compared the methods proposed in (Solorio and Liu, 2008) and (Vyas et al., 2014) on a subset of 1,239 code-mixed Facebook posts from (Barman et al., 2014) and found that a modified version of (Solorio and Liu, 2008) performed best. They also experimented with performing joint POS and LID tagging using 2-level factorial Conditional Random Field and achieved statistically similar results.

AlGhamdi et al. (2016) tested seven different POS tagging strategies for CS data: four consisted of combinations of monolingual systems and the other three were integrated systems. They tested them on MSA-Egyptian Arabic and English-Spanish. The first three combined strategies consisted of running monolingual POS taggers and language ID taggers in different order and combining the outputs in a single multilingual prediction. The fourth approach involved training an SVM on the output of the monolingual taggers. The three integrated approaches trained a supervised model on a) the Miami Bangor corpus (which contains switched and monolingual utterances), b) the union of two monolingual corpora (Ancora-ES and Penn Treebank), c) the union of the three corpora. The monolingual approaches consistently underperformed compared to the other strategies. The SVM approach consistently outperformed the integrated approaches. However, this method was trained on both monolingual and multilingual resources – the Penn Treebank Data for the English model, and the Ancora-ES dataset for the Spanish model. In Section 6.4, we run experiments in similar conditions to the integrated approaches from (AlGhamdi et al., 2016), which we will compare to our work. The main contributions of this paper over this previous research on POS tagging for CS data, are the following: a) Our tagger is a bidirectional LSTM that achieves POS tagging accuracy comparable to state-of-the-art taggers on benchmark datasets like the Wall Street Journal corpus and the Universal Dependencies corpora. It is the first such model used to train code-switched POS taggers; b) Our model can simultaneously perform POS and LID tagging without loss of POS tagging accuracy; c) We run experiments on the Miami Bangor corpus of Spanish and English conversational speech. However, unlike (AlGhamdi et al., 2016) which used POS tags obtained from an automatic tagger and then mapped to a deprecated version of the Universal POS tagset, our experiments are run on newly crowd-sourced Universal POS tags (Soto and Hirschberg, 2017), which were obtained with high accuracy and inter-annotator agreement.

3 A Model for Neural POS Tagging

For our experiments we use a bi-directional LSTM network similar to the one proposed by Wang et al. (2015) with the following set of features: 1) word embeddings, 2) prefix and suffix embeddings of one, two and three characters, and 3) four boolean features that encode whether the word is all upper case, all lower case, formatted as a title, or contains any digits. In total, the input space consists of seven embeddings and four boolean features. For the embeddings, we compute word, prefix and suffix lexicons, excluding tokens that appear less than five times in the training set, and then assign a unique integer to each token. We also reserve two integers for the padding and out-of-lexicon symbols.

We present two architectures for POS tagging and one for joint POS and LID tagging. In the most basic architecture the word, prefix and suffix embeddings and the linear activation units are concatenated into a single layer. The second layer of the network is a bidirectional LSTM. Finally, the output layer is a softmax activation layer, whose i-th output unit at time t represents the probability of the word w_t being the part-of-speech POS$_i$. We refer to this model as Bi-LSTM POS Tagger for the rest of the article and in our tables. For the second model, given the multilingual nature of our experiments, we modify the input space of our Bi-LSTM tagger to make use of the language ID information in our corpus. We add six more boolean features to represent the language ID (one for each label) and add six linear activation units in the first hidden layer, which are then concatenated with the rest of linear activation units and word embeddings in the basic model. This model is referred to as Bi-LSTM POS tagger + LID features.

Finally, our third model simultaneously tags words with POS and LID labels. The architecture of this model follows the Bi-LSTM POS architecture very closely adding a second output layer with softmax activations for LID prediction. Note that the POS and LID output layers are independent and are connected by their weight matrices to the hidden layer, and both loss functions are given the same weight. This model is referred to as joint POS+LID tagger. We implemented our code using the library for deep learning Keras (Chollet, 2015), on a Tensorflow backend (Abadi et al., 2015).

Corpus	Split	# Sents	# Toks
WSJ	Train	38.2K	912.3K
	Dev.	5.5K	131.7K
	Test	5.5K	129.7K
UD-EN	Train	12.5K	204.6K
	Dev.	2K	25.1K
	Test	2K	25.1K
UD-ES	Train	14.2K	403.9K
	Dev.	1.6K	43.5K
	Test	274	8.4K
Miami Bangor	Full	42.9K	333.1K
	Train	38.7K	300.3K
	Test	4.2K	32.8K
	Train Inter-CS	36.0K	267.3K
	Test Intra-CS	285	3.6K

Table 1: Datasets used for our experiments.

4 Datasets

Throughout our experiments we use three corpora for different purposes. The Wall Street Journal (WSJ) corpus is used to demonstrate that our proposed Bi-LSTM POS tagger is on par with current state-of-the-art English POS taggers. The Universal Dependencies (UD) corpus is used to train baseline monolingual POS taggers in English and Spanish that we can use to test on our CS data since both employ the Universal POS tagset (Petrov et al., 2012). The Miami Bangor corpus, which contains instances of inter- and intra-sentential CS utterances in English and Spanish, is used for training and testing CS models and comparing these to monolingual models. Table 1 shows the number of sentences/utterances and tokens in each dataset split. For the MB corpus, Inter-CS refers to the subset of monolingual sentences and Intra-CS refers to the subset of CS sentences.

4.1 Wall Street Journal Corpus

The WSJ corpus (Marcus et al., 1999) is a monolingual English news corpus comprised of 49208 sentences and over 1.1 million tokens. It is tagged with the Treebank tagset (Santorini, 1990; Marcus et al., 1993), which has a total of 45 tags. We use the standard training, development and test splits from (Collins, 2002) which span sections 0-18 19-21 and 22-24 respectively.

4.2 Universal Dependency Corpora

Universal Dependencies (UD) is a project to develop cross-linguistically consistent treebank annotations for many languages. The English UD corpus (Silveira et al., 2014) is built from the English Web Treebank (Bies et al., 2012). The corpus contains data from web media sources, including web logs, newsgroups, emails, reviews and Yahoo! answers. The trees were automatically converted into Stanford Dependencies and then hand-corrected to Universal Dependencies. The corpus contains 16,622 sentences and over 254K tokens. The Spanish UD corpus (McDonald et al., 2013) is built from the content head version of the Universal Dependency Treebank v2.0, to which several token-level morphology features were added. It is comprised of news blog data and has a total of 16,013 sentences and over 455k tokens.

Split	Full	Train	Test	CS
EN	53.48	53.41	54.14	38.98
ES	27.78	27.86	27.04	46.12
PUNCT	15.71	15.76	15.55	12.26
AMBIG	2.27	2.25	2.49	2.06
MIXED	0.01	0.01	0.00	0.01
OTHER	0.76	0.76	0.79	0.60

Table 2: Language composition (%) of the MB corpus.

4.3 Miami Bangor Corpus

The Miami Bangor (MB) corpus is a conversational speech corpus recorded from bilingual English-Spanish speakers living in Miami, FL. It includes 56 conversations recorded from 84 speakers. The corpus consists of 242,475 words (333,069 including punctuation tokens) and 35 hours of recorded conversation. The language markers in the corpus were manually annotated. Table 2 shows the language composition of the corpus. The dominant language in this corpus is English (53.48% of the tokens), followed by Spanish (27.78%). The ambiguous label includes words that are difficult to tag as either English or Spanish due to lack of context (e.g. “no”). Since, in the original corpus, punctuation tokens were labeled as ambiguous, we created an additional punctuation tag for our experiments. The mixed category contains tokens that are formed by morphemes and roots from both languages (e.g. “ripear”) and the category ’Other’ untranscribed tokens. However, the composition of the subset of CS sentences is different: Spanish becomes the dominant language, comprising 46.12% of the tokens compared to 38.98% of the English tokens.

The utterances in the original MB corpus were transcribed in the CHAT transcription and coding format (MacWhinney, 2000), which allows annotators to divide full utterances in chunks to repre-

	Full	Train	Test	CS
#Switches(K)	4.2	3.8	0.4	4.2
Avg.#swts/utt	0.098	0.098	0.095	1.41
Swt.words(%)	1.26	1.27	1.22	11.00
Swt.utts(#)	2980	2695	285	2980
Swt.utts(%)	6.94	6.96	6.79	100
0 swt.(%)	93.06	93.04	93.21	0.00
1 swt.(%)	4.79	4.78	4.83	69.03
2 swt.(%)	1.71	1.73	1.55	24.62
Max#Swt.	8	8	7	8

Table 3: CS in the Miami Bangor Corpus. The top subtable shows the number of switches, the average number of switches per utterances, the amount of switched words (word after which a switch occurs), and the amount of switched utterances in each partition. The bottom subtable shows the percentage of utterances that contain n switches.

sent citations and other speech discourse phenomena. However, working on full utterances is more suitable in the context of POS tagging. Therefore, following the guidelines in (MacWhinney, 2009), we used the utterance linkers and utterance terminators to reconstruct full utterances when possible. After this, the corpus had a total of 16013 sentences and 333K tokens.

The original MB corpus was automatically glossed and tagged with POS tags using the Bangor Autoglosser (Donnelly and Deuchar, 2011a,b). The autoglosser finds the gloss for each token in the corpus and assigns the tag or group of tags most common for that word in the annotated language. However, here we use the Universal POS tags obtained by (Soto and Hirschberg, 2017). These tags were collected using crowdsourcing tasks and automatic labeling, with high annotation accuracy and label recall. We split the MB corpus into training and test. For the test split we randomly drew 4,200 utterances. The training split is used for 4-fold cross-validation. Table 3 shows the degree of multilingualism in the MB corpus and the two splits. In the full dataset, about 6.94% of the utterances contain intra-sentential switches. Note that full dataset and its train and test splits (columns 2 to 4) have very similar degrees of multilingualism according to the reported measures, whereas the subset of intra-sentential CS sentences (column 5) has a much higher rate of switched tokens (11%, from 1.26%) and average number of switches per sentence (1.41, from 0.098). More than 93% of CS utterances contain one or two switches; some contain up to eight switches. For example, the following sentence contains five switches (marked with '|'): *"... y en | summer | y en | fall | tengo que hacer | one class."*

5 Methodology

For the experiments involving the Bangor corpus, we perform 4-fold cross-validation (CV) on the training corpus to run grid search and obtain the best learning rate and decay learning rate parameter values. For the experiments on WSJ and UD, we use the official development set. The performance of the best parameter values is reported as "Dev" accuracy. We then train a model using the best parameter values on the full train set and obtain predictions for the test set (reported as "Test"). When pertinent we also report results on the subset of intra-sentential CS utterances of the test set (reported as "Intra-CS Test").

During CV, each model is trained for a maximum number of 75 epochs using batches of 128 examples. We use early stopping to halt training when the development POS accuracy has not improved for the last three epochs, and keep only the best performing model. However, when training the final model, we stop training after the number of epochs that the best model trained for during CV. The loss function used is categorical cross-entropy and we use ADAM (Kingma and Ba, 2015) with its default β_1, β_2 and ϵ parameter values as the stochastic optimization method.

The word embeddings (Bengio et al., 2003) we use are trained with the rest of the network during training following the Keras implementation (Gal and Ghahramani, 2016). The size of the embedding layers is 128 for the word embeddings and 4, 8 and 16 for the prefix and suffix embeddings of length 1, 2 and 3 respectively. The Bi-LSTM hidden layer has 200 units for each direction.

Finally, we run McNemar's test (McNemar, 1947) to show significant statistical difference between pairs of classifiers when the accuracy of the classifiers is similar, and report statistical significance for p-values smaller than 0.05.

6 Experiments & Results

In this section, we present our experiments using the three Bi-LSTM models introduced in Section 3 and the datasets from Section 4. Our goal is a) to show that the basic Bi-LSTM POS tagger performs very well against known POS tagging benchmarks; b) to obtain baseline performances for monolingual taggers when tested on CS data;

and c) to train and test the proposed models on CS data and analyze their performance when trained on different proportions of monolingual and CS data.

6.1 WSJ results

We begin by evaluating the performance of the Bi-LSTM POS tagger on the benchmark WSJ corpus to show that it is on par with current state-of-the-art English POS taggers. We train taggers on three incremental feature sets to measure how much each feature adds. Using only word embeddings we achieve 95.14% accuracy on the test set; adding word features increases accuracy to 95.84%; and adding the prefix and suffix embeddings further increases accuracy by up to 97.10%. This demonstrates that our tagger is on par with current state-of-the-art systems which report 97.78% (Ling et al., 2015), 97.45% (Andor et al., 2016), 97.35% (Huang et al., 2012), 97.34% (Moore, 2014) and 97.33% (Shen et al., 2007) accuracy on their standard test set. Systems most similar to our Bi-LSTM tagger with basic features reported 97.20% in (Collobert et al., 2011) and 97.26% (Wang et al., 2015).

6.2 Universal tagset baseline

In the second set of experiments we train baseline monolingual Spanish and English taggers on the UD corpora: one monolingual Spanish and one monolingual English tagger, and one tagger trained on both corpora. The goal of these experiments is to obtain taggers trained on the Universal tagset that we can use to obtain a baseline performance of monolingual taggers on the CS Bangor corpus. The results are shown in Table 4. The accuracy of the baseline UD taggers is slightly worse than the WSJ taggers, probably due to the smaller size of the UD datasets. The accuracy of the taggers on their own test sets is 94.78% and 95.02% for English and Spanish respectively. In comparison, Stanford's neural dependency parser (Dozat et al., 2017) reports accuracy values of 95.11% and 96.59% respectively.

In order to approximate how a monolingual tagger trained on established datasets performs on a conversational CS dataset, we test the baseline UD taggers on the MB test set and observe a dramatic drop in accuracy, due perhaps to the difference in genre (web blog data vs. transcribed conversation) and the bilingual nature of the Miami corpus. Note that, when training on both EN and ES UD, the Bi-LSTM taggers reach 88.17% accuracy, from only 69.97 and 45.13% by the monolingual taggers. When looking at the multilingual subset of sentences from the test set (CS Test in Table 4), we observe that the English model decreases in accuracy further, whereas the Spanish tagger has better performance. This is due to the CS sentences having more Spanish than English.

	UD		MB	
Training	Dev	Test	Test	CS Test
UD EN	94.53	94.78	69.97	56.20
UD ES	96.20	95.02	45.13	55.32
UD EN&ES	94.88	94.25	88.17	87.18

Table 4: Bi-LSTM POS tagging accuracy (%) on the Universal Dependency corpora. The left subtable shows the accuracy on the UD dev and test sets. The right subtables shows the accuracy on the MB test set and on the subset of CS utterances.

Training	Task	Dev	Test	CS Test
MB	Tagger	96.27	96.34	96.10
	Tagger+LID	96.35	96.49	**96.44**
	Joint Tagger	96.30	96.39	95.97
MB + UD	Tagger	96.34	96.47	95.99
	Tagger+LID	**96.40**	**96.63**	**96.44**
	Joint Tagger	96.39	96.61	96.35
MB Inter-CS	Tagger	96.24	96.03	95.27
	Tagger+LID	96.26	96.16	95.55
	Joint Tagger	96.25	96.11	95.22

Table 5: POS tagging accuracy (%) on the MB corpus. Underlined font indicates best result in test set by each training setting across different tagging models. Bold results indicate best overall result in that test set.

6.3 Miami Bangor results

In the third set of experiments we train the three proposed models (Bi-LSTM tagger, Bi-LSTM tagger with LID features and joint POS and LID tagger) on: a) the full MB corpus, b) the joint MB and UD ES&EN corpora, and c) instances of inter-sentential CS utterances from the MB corpus. The LID features were obtained from the MB corpus language tags. POS and LID accuracy results are shown in Table 5 and Table 6 respectively.

When training on the full MB corpus (top subtable from table 5), the POS tagger achieves 96.34% accuracy, a significant improvement from the 88.17% of the UD EN&ES. The improvement holds up on the subset of CS utterances, achieving 96.10% accuracy. Adding the LID features improves performance by 0.15 and 0.34 absolute percentage points. In both cases these differences are statistically significant ($p = 0.03$). Furthermore, when running joint POS and LID tagging, we see that tagging accuracy decreases slightly with respect to the POS tagger with LID features. This result reaffirms the contribution of the LID features. The difference in performance between the joint tagger and the basic tagger is slightly higher but not statistically significant ($p \sim 0.5$), showing that joint decoding does not harm overall performance. The best POS tagging accuracy is always achieved by the Bi-LSTM tagger with LID features on both Test and CS Test; however, the joint Tagger is very close at no more than 0.1 percentage points on Test. When adding the UD corpora during training (middle subtable from Table 5) we see some improvements for the three models (0.13, 0.14 and 0.22 absolute percentage points respectively), and once again the difference in performance between the basic tagger and the tagger with LID features is statistically significant ($p < 0.05$).

We performed statistical tests to measure how different the models trained on MB are from the models trained on MB+UD and found that the addition of more monolingual data only makes a difference for the joint tagger ($p < 0.01$) when looking at the performance on the Test set. On the CS test set, these models achieve about the same performance in POS tagging with a slight decrease for the basic tagger (-0.11 points, not significant) and a slight increase in accuracy for the joint tagger (0.38 percentage points, again not significant). Thus, it is clear that our model is able to learn from a few CS examples – even when many more monolingual sentences, from a different genre, are added to the train set.

Finally, we trained models on the subset of monolingual English and Spanish sentences from the MB training set to measure how a model trained on the same genre would be able to generalize on unseen intra-sentential CS sentences (bottom subtable from Table 5, marked as Inter-CS). This model would be closer to an in-genre inter-sentential CS tagger, tested on intra-sentential CS. Compared to the models trained on UD EN&ES, this model performs much better: 96.03% compared to 88.17% on the MB test set. This is mainly due to the fact that the UD corpus is *not* conversational speech. When comparing this result to the taggers trained on the full MB corpus, it can be seen that these new models achieved the lowest test accuracy across all models, probably due to the lack of bilingual examples in their training set. The difference in performance is more pronounced on the subset of CS utterances. Again, we ran statistical tests to compare these three new taggers to the taggers trained on the full MB corpus, and we found that their differences were statistically significant in the three cases ($p < 0.001$).

Training	Dev	Test	CS Test
MB	98.82	98.78	98.01
MB + UD EN&ES	98.60	98.49	97.93
MB Inter-CS Subset	98.53	97.99	90.25

Table 6: LID tagging accuracy by the Bi-LSTM joint POS+LID Tagger on the MB corpus.

With respect to the LID accuracy of the joint Tagger, the best model is the one trained on the MB corpus, followed very closely by the model trained on MB and UD data. In both cases, the test set accuracy is above 98.49%. The accuracy on the CS test subset is sightly lower at 98.01% and 97.93%. The monolingual Bangor tagger sees a decrease in test accuracy (97.99%) and a bigger drop, down to 90.25%, on the CS subset. All the differences in performance between every pair of the three LID taggers are statistically significant ($p < 10^{-5}$).

6.4 Comparison to Previous Work

We compare the performance of our models to the Integrated and Combined models proposed in (AlGhamdi et al., 2016). In that paper, POS tagging results are reported on the MB corpus, but using a preliminary mapping to the first iteration of the Universal tagset (12 tags, as opposed to the current 17); furthermore, the train and test splits were different. Therefore, we decided to replicate their experiments using our data configuration and compare them to our own classifiers. With respect to their "Integrated" models, INT3:AllMonoData+CSD is comparable to our POS Tagger trained on the full MB set and UD EN&ES (ours at 96.47% compared to 92.33%); INT2:AllMono is comparable to our POS Tagger trained on UD EN&ES (ours at 88.17% compared to 84.47%) and INT1:CSD is comparable to our POS Tagger trained on Bangor (ours at 96.34% versus 92.71%). For their "Combined" models, COMB4:MonoLT-SVM trained two monolingual taggers on the UD-EN and UD-ES corpora and then a SVM on top from the output of the taggers on the MB corpus. We do not perform system

	EN	ES	EN&ES	Bangor
OOV	40.9	32.7	10.7	7.9
SAcc.	2.5	4.2	21.8	60.7
WAcc.	56.2	55.3	87.2	96.1
CSFAcc.	10.9	12.6	57.5	84.2
CSFWAcc.	12.6	16.1	63.3	86.7
AvgMinDistCSF	4.0	5.4	3.9	3.5
%ErrorsInCSF	26.9	24.3	32.5	36.9

Table 7: Out-of-vocabulary (OOV) rate, sentence (Sacc) and word accuracy (Wacc) at the sentence level, fragment (CSFAcc) and word accuracy (CSFWacc) at the fragment level, average minimum distance from tagging error to CSF (AvgMinDistCSF), and percentage of errors that occur within a CSF (%ErrorsInCSF).

combination in this paper, but in terms of data, this model would be most similar to our POS tagger trained on Miami and EN&ES UD, in which we reached 96.47% compared to their 92.20%. Furthermore, we note that our joint POS+LID tagger also has better POS accuracy than its counterparts Integrated systems from (AlGhamdi et al., 2016) in addition to performing LID tagging.

7 Error Analysis

In this section we aim to analyze the performance of the POS taggers on the CS sentences of the Bangor test set and more specifically, on the CS fragments (CSF) of those test sentences. We define a CSF as the minimum contiguous span of words where a CS occurs. Most often a CSF will be two words long, spanning a Spanish token and an English one or vice versa, but it is also possible for fragments to be longer than that, given that a Mixed or Ambiguous token could occur within a fragment. The average CSF length in the Bangor test set is 2.16. We compare the performance of the UD-EN, UD-ES, UD-EN&ES and the Bangor-trained taggers on the Bangor CS Test set to understand the difference in errors made by monolingual and CS taggers. Table 7 shows the following measures: OOV rate, POS tagging accuracy at the sentence and word level, POS tagging accuracy in CS fragments at the fragment and word level, the average distance from a POS tagging error to the nearest CSF (AvgMinDistCSF) and the percentage of POS tagging errors that occur within the boundaries of a CS utterance (%ErrorsInCSF). All measures are computed on the CS subset of test sentences of the Bangor corpus using the basic POS taggers trained on UD-EN, UD-ES, UD EN&ES and the Bangor corpus. In the table, we see that the multilingual models have much lower OOV rates, which translates into much higher sentence-level and word-level POS tagging accuracy. The CS Bangor-trained model fares much better than the UD EN&ES model in terms of word-level accuracy (96.1 versus 87.2%), especially when looking at the sentence-level accuracy (60.7 versus 21.8%), because the Bangor model is able to deal with code-switches. When looking at the tagging accuracy on the CS utterances the relative gains at the word level are even larger. This demonstrates that training on CS sentences is an important factor to achieving high-performing POS tagging accuracy.

It can also be seen from the table that, as the models achieve CS tagging accuracy, tagging errors are still concentrated near or within CSFs – for the UD EN&ES and Bangor models, AvgMinDistCSF and %ErrorsInCS decrease as the CSF-level accuracies increase. This shows that even as the models improve at tagging CS fragments, CS fragments still remain the most challenging aspect of the task.

8 Conclusions

In this paper, we have presented a neural model for POS tagging and language identification on CS data. The neural network is a state-of-the-art bidirectional LSTM with prefix, suffix and word embeddings and four boolean features. We used the Miami Bangor corpus to train and test models and showed that: a) monolingual taggers trained on benchmark training sets perform poorly on the test set of the CS corpus, b) our CS models achieve high POS accuracy when trained and tested on CS sentences, c) expanding the feature set to include language ID as input features yielded the best performing models, d) a joint POS and language ID tagger performs comparably to the POS tagger and its LID accuracy is higher than 98%, and e) a model trained on instances of in-genre inter-sentential CS performs much better than the monolingual baselines, but yielded worse test results than the model trained on instances of inter-sentential and intra-sentential code-switching. Furthermore, we compared to our results to the previous state-of-the-art POS tagger for this corpus and showed that our classifiers outperform them in every configuration.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proc. of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proc. of The First Workshop on Computational Approaches to Code Switching*, pages 13–23.

Utsab Barman, Joachim Wagner, and Jennifer Foster. 2016. Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In *Proc. of The Second Workshop on Computational Approaches to Code Switching*, pages 42–51.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank LDC2012T13. `https://catalog.ldc.upenn.edu/LDC2012T13`.

Mónica Stella Cárdenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in Internet chatting: between 'yes', 'ya', and 'si' – a case study. *The JALT CALL Journal*, 5(3):67–78.

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

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Brenda Danet and Susan C Herring. 2007. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press on Demand.

Crystal David. 2001. *Language and the Internet*. Cambridge, CUP.

Kevin Donnelly and Margaret Deuchar. 2011a. The Bangor Autoglosser: a multilingual tagger for conversational text. *ITA11, Wrexham, Wales*.

Kevin Donnelly and Margaret Deuchar. 2011b. Using constraint grammar in the Bangor Autoglosser to disambiguate multilingual spoken text. In *Constraint Grammar Applications: Proceedings of the NODALIDA 2011 Workshop, Riga, Latvia*, pages 17–25.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanfords graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1019–1027.

Liang Huang, Suphan Fayong, and Yang Guo. 2012. Structured perceptron with inexact search. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–151. Association for Computational Linguistics.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. In *Proc. of Recent Advances in Natural Language Processing*, pages 239–248.

Diederik P Kingma and Jimmy Lei Ba. 2015. ADAM: A method for stochastic optimization. In *Proceedings of the Third International Conference on Learning Representations (ICLR)*.

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk.*, 3rd edition. Lawrence Erlbaum Associates, Inc.

Brian MacWhinney. 2009. The CHILDES project part 1: The chat transcription format. *Department of Psychology*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3 LDC99T42. https://catalog.ldc.upenn.edu/ldc99t42.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pages 92–97.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Robert Moore. 2014. Fast high-accuracy part-of-speech tagging by independent classifiers. In *COLING*, pages 1165–1176.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Paul Rodrigues. 2013. Part of speech tagging bilingual speech transcripts with intrasentential model switching. In *AAAI Spring Symposium*, pages 56–63.

Mike Rosner and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *Proc. of INTERSPEECH*, pages 190–193.

Beatrice Santorini. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)*, 3 edition. LDC, UPenn. 2nd Printing.

Royal Sequiera, Monojit Choudhury, and Kalika Bali. 2015. POS tagging of Hindi-English code mixed text from social media: Some machine learning experiments. In *Proc. of ICON*.

Libin Shen, Giorgio Satta, and Aravind Joshi. 2007. Guided learning for bidirectional sequence classification. In *ACL*, volume 7, pages 760–767.

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)*.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proc. of EMNLP*, pages 1051–1060.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing universal part-of-speech tags for code-switching. *Interspeech*.

US Census Bureau. 2014. Annual estimates of the resident population by sex, age, race, and Hispanic origin for the United States: April 1, 2010 to July 1, 2014. https://factfinder.census.gov/bkmk/table/1.0/en/PEP/2014/PEPASR6H?slice=hisp~hisp!year~est72014.

US Census Bureau. 2015. American community survey 1-year estimates: S1601 - language spoken at home. https://factfinder.census.gov/bkmk/table/1.0/en/ACS/15_1YR/S1601.

Yogarshi Vyas, Spandana Gella, and Jatin Sharma. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proc. of EMNLP*, pages 974–979.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. A unified tagging solution: Bidirectional LSTM recurrent neural network with word embedding. *arXiv preprint arXiv:1511.00215*.