Leveraging Pretrained Word Embeddings for Part-of-Speech Tagging of Code Switching Data

Fahad AlGhamdi [1], **and Mona Diab**[1,2]

[1] Department of Computer Science, The George Washington University
[2] AWS, Amazon AI
{fghamdi, mtdiab}@gwu.edu

Abstract

Linguistic Code Switching (CS) is a phenomenon that occurs when multilingual speakers alternate between two or more languages/dialects within a single conversation. Processing CS data is especially challenging in intra-sentential data given state-of-the-art monolingual NLP technologies since such technologies are geared toward the processing of one language at a time. In this paper, we address the problem of Part-of-Speech tagging (POS) in the context of linguistic code switching (CS). We explore leveraging multiple neural network architectures to measure the impact of different pre-trained embeddings methods on POS tagging CS data. We investigate the landscape in four CS language pairs, Spanish-English, Hindi-English, Modern Standard Arabic- Egyptian Arabic dialect (MSA-EGY), and Modern Standard Arabic-Levantine Arabic dialect (MSA-LEV). Our results show that multilingual embedding (e.g., MSA-EGY and MSA-LEV) helps closely related languages (EGY/LEV) but adds noise to the languages that are distant (SPA/HIN). Finally, we show that our proposed models outperform state-of-the-art CS taggers for MSA-EGY language pair.

1 Introduction

Code Switching (CS) is a common linguistic behavior where two or more languages/dialects are used interchangeably in either spoken or written form. CS is typically present at various levels of linguistic structure: across sentence boundaries (i.e., inter-sentential), within the same utterances, mixing two or more languages (i.e., intra-sentential), or at the words/morphemes level. The CS phenomenon is noticeable and common in countries that have large immigrant groups, naturally leading to bilingualism. Typically people who code switch master two (or more) languages: a common first language (lang1) and another prevalent language as a second language (lang2). The languages could be completely distinct such as Mandarin and English, or Hindi and English, or they can be variants of one another such as in the case of Modern Standard Arabic (MSA) and Arabic regional dialects (e.g. Egyptian dialect-EGY). CS is traditionally prevalent in the spoken modality but with the ubiquity of the Internet and proliferation of social media, CS is becoming ubiquitous in written modalities and genres (Vyas et al., 2014a; Danet and Herring, 2007; Cárdenas-Claros and Isharyanti, 2009). This new situation has created an unusual deluge of CS textual data on the Internet. This data brings in its wake new opportunities, but it poses serious challenges for different NLP tasks; traditional techniques trained for one language tend to break down when the input text happens to include two or more languages. The performance of NLP models that are currently expected to yield good results (e.g., Part-of-Speech Tagging) would degrade at a rate proportional to the amount and level of mixed-language present. This is a result of out-of-vocabulary words in one language and new hybrid grammar structures, and in some cases shared cognates or ambiguous words that exist in both language lexicons.

POS tagging is a vital component of any Natural Language Understanding system, and one of the first tasks researchers employ to process data. POS tagging is an enabling technology needed by higher-up NLP tools such as chunkers and parsers – syntactic, semantic and discourse level processing; all of which are used for such applications as sentiment analysis and subjectivity, text summarization, information extraction, automatic speech recognition, and machine translation among others. As such, it is crucial that POS taggers be able to process CS textual data.

In this paper, we address the problem of Part-of-Speech tagging (POS) for CS data on the intra-sentential level for multiple language pairs. We explore the effect of using various embeddings setups and multiple neural network architectures in order to mitigate the problem of the scarcity of CS annotated data. We propose multiple word embedding techniques that could help in tackling POS tagging of CS data.

In order to examine the generalization of our approaches across language pairs, we conduct our study on four different evaluation CS data sets, covering four language pairs: Modern Standard Arabic and the Egyptian Arabic dialect (MSA-EGY), Modern Standard Arabic- and the Levantine Arabic dialect (MSA-LEV), Spanish-English (SPA-ENG) and Hindi-English (HIN-ENG). We use the same POS tag sets for all language pairs, namely, the Universal POS tag set (Petrov et al., 2011). Our contributions are the following: a) We use a state-of-the-art bidirectional recurrent neural networks; b) We explore different strategies to leverage raw textual resources for creating pre-trained embeddings for POS tagging CS data; c) We examine the effect of language identifiers for joint POS tagging and language identification; d) We present the first empirical evaluation on POS tagging with four different language pairs. All of the previous work focused on a single or two language pair combinations.

2 Related Work

Developing CS text processing NLP techniques for analyzing user generated content as well as cater for needs of multilingual societies is vital (Vyas et al., 2014b). Previous studies that address the problem of POS tagging of CS data first attempt to identify the correct language of a word before feeding it into an appropriate monolingual tagger (Solorio and Liu, 2008; Vyas et al., 2014a; AlGhamdi et al., 2016). As is typically the case in NLP, such pipelines suffer from the problem of error propagation; e.g., failure of the language identification will cause problems in the POS tag prediction. Other approaches have trained supervised models on POS-annotated, CS data resources which are expensive to create and unavailable for most language pairs. (AlGhamdi et al., 2016; Solorio and Liu, 2008; Jamatia et al., 2015; Barman et al., 2016) Solorio and Liu (2008), proposed the first statistical approach to POS tagging

of CS data where they employ 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. Royal Sequiera (2015) introduces a ML-based approach with a number of new features for HIN-ENG POS tagging for Twitter and Facebook chat messages. The new feature set considers the transliteration problem inherent in social media. Their system achieves an accuracy of 84%. Jamatia et al. (2015) use both a fine-grained and coarse-grained POS tag set in their study. They introduce a comparison between the performance of a combination of language specific taggers and a machine learning based approach that uses a range of different features. They conclude that the machine learning approach failed to outperform the language specific combination taggers. AlGhamdi et al. (2016) examine seven POS tagging strategies to leverage the available monolingual resources for CS data. They conducted their study on two language pairs, namely MSA-EGY and SPA-ENG. The proposed strategies are divided into combined conditions and integrated conditions. Three of the combined conditions consist of running monolingual POS taggers and language ID taggers in a different order and combining the outputs in a single multilingual prediction. The fourth combined condition involved training an SVM model using the output of the monolingual taggers. The three integrated approaches trained a monolingual state-of-the-art POS tagger on a) CS corpus; b) the union of two monolingual corpora of the languages included in the CS; c) the union of the corpora used in a and b. Both combined and integrated conditions outperformed the baseline systems. The SVM approach consistently outperformed the integrated approaches achieving the highest accuracy results for both language pairs. Soto and Hirschberg (2018) propose a new approach to POS tagging for code switching SPA-ENG language pair based on recurrent neural network (RNN) as a way of providing better tools for better code switching data processing, including POS taggers. The authors use an experimental approach of POS tagging for CS utterances entailing use of state-of-the-art bi-directional RNN to extensively study effects of language identi-

fiers. For the Boolean features, a bi-directional Long short-term memory (BiLSTM) state-of-the-art neural network with suffix, prefix, and word embeddings, the results show that the neural POS tagging model proposed by the authors performs comparatively higher than other state of the art CS taggers, their system yields a POS tagging accuracy of 96.34%, while joint POS and language ID tagging yields an accuracy of 96.39% for POS tagging and language ID accuracy of 98.78% (Soto and Hirschberg, 2018). Language Modeling (LM) on HIN-ENG language texts has been proposed by (Chandu et al., 2018). (Pratapa et al., 2018) propose the use of a computational technique to create artificial code mixed data that is grammatically valid based on the ECT (Equivalence Constraint Theory) to solve the challenge of scarcity of CS training language models using an experimental approach. (Pratapa et al., 2018) uses two bilingual embedding techniques, namely Bilingual Compositional Model (BiCVM) and Correlation Based Embeddings (BiCCA) (Faruqui and Dyer, 2014; Blunsom and Hermann, 2014). Word embedding results in improved semantic and syntactic CS processing tasks. BiCVM at the sentence level only yields better performance for semantic tasks. BiCCA also only do well on semantic tasks because they make use of word alignments. Furthermore, g-skip embedding outperformed BiCCA and BiCVM, performing well across syntactic and semantic tasks. The study by Pratapa et al. (2018) illustrates that using pretrained embeddings learned from CS data outperforms pretrained embeddings learned from standard bilingual embeddings. Bhat et al. (2018) introduce a dependency parser developed specifically for HIN-ENG CS data. They adopted the neural stacking architecture proposed by (Zhang and Weiss, 2016; Chen et al., 2016) for learning POS tagging and parsing and for transferring the knowledge from bilingual models trained on Hindi and English UD treebanks. The stack-prop tagger, the model proposed by (Zhang and Weiss, 2016), achieves the highest accuracy of 90.53. To identify the language ID, they train a multilayer perceptron (MLP) stacked on top of a recurrent bidirectional LSTM (Bi-LSTM) network. The results of their system is 97.39 %.

3 Approach

Pretrained word embeddings enable models to exploit the raw textual data which is in all languages larger than annotated data. In recent times, there has been some interest in embedding approaches, e.g., multilingual embeddings and bilingual word embeddings, where two monolingual embeddings of two languages are mapped to a shared embeddings space. The main advantage of bilingual and multilingual embeddings is in solving tasks involving reasoning across two languages, such as Machine Translation (MT) (Vulić and Moens, 2016; Zou et al., 2013), as well as enabling transfer of models learned on a resource-rich language onto a resource-poor language (Adams et al., 2017). One of the potential applications of bilingual and multilingual embeddings is in the processing of code switching language. In this section, we compare leveraging multiple neural network architectures for POS tagging CS data. Moreover, we explore different embedding setups to investigate the optimal way of tackling the POS tagging of CS data. First, we illustrate the tool used for training the embeddings layer. Second, we present the neural network models. Then we list the embedding setups.

Pretrained Word Embedding Model Most successful methods for learning word embeddings (Mikolov et al., 2013c; Pennington et al., 2014; Bojanowski et al., 2017) rely on the distributional hypothesis (Mikolov et al., 2013b), i.e., words occurring in similar contexts tend to have similar meanings. Among all word embedding techniques, we choose the FastText tool developed by Facebook (Bojanowski et al., 2017). Our choice of FastText is motivated by the fact that social media networks are the primary source of our unlabeled data. This type of data exhibits huge variations of spelling and misspellings. FastText takes advantage of subword (i.e., n-gram) information. It creates vector representations for misspelling replacement candidates absent from the trained embedding space, by summing the vectors of the character n-grams. All word embedding techniques aim to capture the relation of the words that tend to appear in similar context. These relations occur on the sentential level for most of the NLP applications trained for monolingual data. However, the lack of having sufficient context in CS data makes learning these kinds of relations a diffi-

cult task as they occur on the sub-sentential (intra-sentential) level. Hence, we resort to using Fast-Text due to its principle approach of leveraging subword (i.e., n-gram) information. The n-gram approach resolves the problem of modeling languages with rare word inflections by using character n-grams. On the other hand, other embedding techniques, e.g., word2vec and glove, lack subword information and hence struggle with morphologically rich languages such as Arabic and noisy data such as Twitter data (Mikolov et al., 2013a; Pennington et al., 2014).

A Model for Neural POS Tagging For our experiments we use three neural networks architectures: a) a BiLSTM-CRF architecture similar to the one proposed by (Reimers and Gurevych, 2017) for POS tagging; b) a multi-task learning model for learning jointly POS tagging for related language pairs; and, c) a multi-task learning model for learning jointly POS and Language ID tagging.

- We train a BiLSTM network with a conditional random field objective (Reimers and Gurevych, 2017) that obtain the probability distribution over all labels by jointly modeling the probability of the entire tag sequence score. We initialize the embedding layer with the pre-trained FastText word embeddings and feed the output sequence from this layer to the BiLSTM layer. The BiLSTM hidden layer has 200 units for each direction and dropout of 0.2. We use early stopping (Chollet et al., 2015) based on performance on validation sets. We refer to this model as BiLSTM-CRF POS Tagger for the rest of the article and in our tables.

- Our second model is a multi-task learning model that learns simultaneously POS tagging for related code switching language pairs. The architecture of this model follows the BiLSTM-CRF architecture (Reimers and Gurevych, 2017). We add a second CRF output layer for predicting the POS tag for the related language pair (e.g., MSA-LEV), while the first CRF output layer is for predicting the POS tag for the other related language pair (e.g., MSA-EGY). The two output layers of POS tagging tasks for the two language pairs are independent and are connected by their weight matrices to the hidden layer, and both loss functions are given the same weight. We

refer to this model as MTL-POS Tagger for the rest of the article and in our tables.

- The third model is a multi-task learning model. The model learns simultaneously POS and language id tags with the aim of boosting the performance of POS tagging task. The architecture of this model follows MTL-POS Tagger architecture. The difference is that one output layer is for predicting the POS tagging, and the other output layer is for predicting language id task. This model is referred to as MTL-POS+LID Tagger.

Experimental Conditions

Monolingual embedding (baseline): We train word embeddings using the monolingual corpora for each language involved in the four language pairs. The results of this approach are six separate pre-trained embeddings, MSA, EGY, LEV, ENG, SPA, and HIN pre-trained embeddings. For each language, we train a BiLSTM-CRF model using one of the six pre-trained embeddings. We consider these models as baseline systems. The baseline performance is the POS tagging accuracy of the monolingual models with no special training for CS data.

Merged Bilingual embeddings:

Combined filtered monolingual corpora (CFM): To leverage the inter-sentential code switching type, we train a model using a pre-trained word embedding trained on monolingual data only. The assumption is that the data in MSA is purely MSA and that in EGY is purely EGY. This condition yields an inter-sentential CS pre-trained embeddings. None of the sentences reflect intra-sentential CS data.

Pure Code switched corpora (PCS): To leverage the intra-sentential code switching type, we train a model using a pre-trained word embedding trained on CS data only. The assumption is that the data used to train the embeddings exhibit the CS phenomenon. This condition yields an intra-sentential CS type. None of the sentences reflect inter-sentential CS data.

(Pseudo) Combined monolingual and CS corpora (PseudoCS): To address both code switching types, the intra-sentential and inter-sentential, we combine the pure code switched corpora and combined filtered monolingual corpora to form

unified to be used to train a unified word embeddings.

Dataset	Train	Dev	Test
ARZ (MSA-EGY)	133,357	21,146	20,464
LEV (MSA-LEV)	45,167	5,749	5,779
Miami Bangor (SPA-ENG)	268,464	67,114	67,114
UD-HIN-ENG	19,695	3,339	3,190

Table 1: Datasets distribution for the four language pairs

Merged multilingual embeddings: We train a multilingual embedding for language pairs that have one language in common (pivot language). To do so, we combine the corpora used to train the word embeddings of the language pairs that share one common language. For MSA-EGY and MSA-LEV language pairs, we have a common language, which is MSA, while ENG is a common language between SPA-ENG and HIN-ENG. To leverage this commonality between each of the two language pairs, we merge all the previous corpora: PCS, Mono, and PseudoCS for each language pair to form one corpus used to train merged multilingual embeddings model for MSA-EGY-LEV languages and another corpus for SPA-HIN-ENG. The intuition of this embeddings is to capture word usage in the context of each language and eliminates the ambiguity for the words that have the same surface form in multiple languages.

Projected bilingual embedding Projected bilingual embeddings are vector representations of two languages mapped into shared space, such that translated word pairs have similar vectors. There are three approaches to learn bilingual embeddings: 1) by mapping the space of both monolingual embeddings into a single shared space; 2) monolingual adaptation of one language's embedding space into another's; 3) Bilingual Training by bootstrapping the target representations learned from well-trained embeddings space of a source language. We train individual CS and monolingual embedding models separately before mapping them into a shared embedding space. To do so, We use MUSE (Conneau et al., 2017), state-of-the-art model for creating a projected bilingual embedding that uses the monolingual adaptation technique to create the shared space embedding. MUSE is equipped to learn either via supervision or no supervision. In our study, we utilize the unsupervised version of MUSE.

4 Evaluation

4.1 Data sets

Throughout our experiments, we use one evaluation dataset for each language pair and various corpora for training the embeddings layer. Table-1 shows the distribution of the evaluation data sets.

MSA-EGY We use the LDC Egyptian Arabic Treebanks 1-5 (ARZ1-5) (Maamouri et al., 2012). The ARZ1-5 data is from the discussion forums genre mostly in the Egyptian Arabic dialect (EGY). The total number of sentences in the corpus is 13,698 while the number of words is 174,967 words.

To train pre-trained embeddings, we crawl Egyptian tweets from some of the Egyptian public figures' Twitter accounts. The rest of the Egyptian raw textual data comes from the following sources: (Zaidan and Callison-Burch, 2011)'s Egyptian online news commentary corpus, the Egyptian tweets used in the CS shared tasks (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018), portion of MSA Gigaword (Parker et al., 2011), and LDC Arabic tree bank corpora (MSA) (Maamouri et al., 2004; Diab et al., 2013). To identify the language id of MSA-EGY tokens, we use the Automatic Identification of Dialectal Arabic (AIDA2) tool (Al-Badrashiny et al., 2015) to perform token level language identification for the EGY and MSA tokens in context.

MSA-LEV We use the Curras Corpus of Palestinian Arabic as the MSA-LEV textual CS data (Jarrar et al., 2017). Palestinian Arabic is a sub-dialect of Levantine Arabic. The corpus comprises 57,000 words, half of which come from transcripts of a TV show and the rest of which comes from various sources such as Facebook, discussing forums, Twitter, and blogs. The corpus is morphologically annotated by Eskander et al. (2016) using the same guidelines utilized for annotating the Egyptian ARZ corpus. We annotate the MSA-LEV evaluation data set with language id using the guidelines and tool proposed by (Diab et al., 2016; AlGhamdi and Diab, 2018).

To train pre-trained embeddings, we crawl tweets from Levantine public figures. Moreover, we compile Levantine and MSA raw textual data from multiple resources: online news commentary corpus from (Zaidan and Callison-Burch, 2011),

weblogs from COLABA (Diab et al., 2010), commentaries and tweets from Cotterell and Callison-Burch (2014), the Levantine portion of the PADIC data set (Meftouh et al., 2015), portion of MSA Gigaword (Parker et al., 2011), and LDC Arabic tree bank corpora (MSA) (Maamouri et al., 2004; Diab et al., 2013).

SPA-ENG The Miami Bangor (MB) corpus is a conversational speech corpus recorded from bilingual Spanish-English speakers living in Miami, FL. It includes 56 conversations recorded from 84 speakers (Soto and Hirschberg, 2017). 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.

To train pre-trained embeddings for Spanish and English language, we use the English Universal Dependencies (UD) corpus (Silveira et al., 2014) and the Spanish UD corpus (McDonald et al., 2013). Universal Dependencies (UD) is a project to develop cross-linguistically consistent treebank annotations for many languages. Moreover, we use some other English monolingual data from various resources. The English monolingual data contains around 250M sentences.

HIN-ENG The Hindi-English Code switching treebank is based on CS tweets of Hindi and English multilingual speakers (mostly Indian) (Bhat et al., 2017). The treebank is manually annotated using UD scheme. The corpus contains data from Twitter. The corpus contains 1,852 tweets and 26,224 tokens. To train pre-trained embeddings for Hindi and English language, we use the English Universal Dependencies (UD) corpus (Silveira et al., 2014) and the Hindi UD corpus (Bhat et al.; Palmer et al., 2009). Also, we use some other Hindi monolingual data from various resources. The word representations are learned using Skip-gram model with negative sampling which is implemented in FastText toolkit for all language.

Baseline Results The baseline performance is the POS tagging accuracy of the monolingual models with no special training for CS data. Therefore, our baselines are the neural network models trained using the monolingual embeddings. If CS data do not pose any particular challenge to monolingual POS taggers, then we should not expect a major degradation in performance.

Table-2 shows the performance of the different baseline POS tagging systems on the test data. For each language pair, there are two baseline systems.

For MSA-EGY, the baseline accuracies are 85.40 when the baseline system utilizes the MSA pre-trained embeddings, and 81.06 when BiLSTM-CRF uses the EGY pre-trained embeddings. Similarly, we report the baseline results for MSA-LEV, SPA-ENG and HIN-ENG language pairs.

4.2 Results

In this section, we present the results of our experiments using the neural network models and embeddings approach introduced in Section 3 and the datasets from Section 4.1. Also, we show the results of three neural network models when the embeddings are randomly initialized. Table-2 shows the POS tagging accuracy results of all language pairs, while Table-3 shows the LID accuracy results of all language pairs using MTL-POS+LID Tagger. To evaluate the performance of our approaches we report the accuracy of each condition by comparing the output POS tags generated from each condition against the available gold POS tags for each data set. We consistently apply the different experimental conditions on the same test set per language pair: for MSA-EGY we report results on MSA-EGY test set, for MSA-LEV we report results on MSA-LEV test set, and for SPA-ENG, we report results on Miami Bangor corpus (SPA-ENG) test set, and finally for HIN-ENG, we report on UD-HIN-ENG (HIN-ENG) test set. The highest accuracy results for MSA-EGY and MSA-LEV language pairs are 92.90% and 92.92%, respectively. These results are achieved by MTL-POS Tagger+Merged Bilingual PseudoCS embeddings. Our best model for MSA-EGY outperform state-of-the-art system by $\sim 2\%$ (AlGhamdi et al., 2016). We could not compare our best system for MSA-LEV language pair to any previous systems as we map the original POS tag set (Buckwalter POS tag set) of the MSA-LEV dataset into UD POS tag set. BiLSTM-CRF+Merged Bilingual PseudoCS embeddings yield the highest accuracy results for both SPA-ENG and HIN-ENG language pairs. The SPA-ENG model's accuracy, 96.55%, is comparable to the state-of-the-art system, 96.63% (Soto and Hirschberg, 2018). On the other hand, the accuracy of our best HIN-ENG model (86.01%) underperforms the state-of-the-

Embedding Condition	MSA-EGY	MSA-LEV	SPA-ENG	HIN-ENG
BiLSTM-CRF Tagger (1) + Random-Initi-Embed	89.12	88.90	95.33	69.87
(1) + mono (MSA/SPA/HIN) (Baseline)	85.40	83.25	50.30	63.30
(1) + mono (EGY/LEV/ENG) (Baseline)	81.06	80.21	71.21	67.11
(1)+ Merged Bilingual CFM	90.00	89.41	95.40	85.98
(1)+Merged Bilingual PCS	89.06	88.94	94.22	83.31
(1)+Merged Bilingual PseudoCS	91.96	91.92	*96.55*	*86.01*
(1)+ Merged Multilingual Embeddings(pivot)	92.81	92.91	94.81	85.87
(1)+Projected Bilingual	89.24	87.60	91.31	83.02
MTL-POS Tagger (2) + Random-Initi-Embed	89.91	90.02	92.89	79.51
(2)+Merged Bilingual PCS	90.01	90.51	93.21	84.30
(2)+Merged Bilingual PseudoCS	*92.90*	*92.92*	94.14	84.33
MTL-POS+LID Tagger (3) + Random-Initi-Embed	88.96	89.79	92.65	78.51
(3)+ Merged Bilingual CFM	90.00	89.01	95.42	84.01
(3)+Merged Bilingual PCS	90.41	90.48	95.29	84.41
(3)+Merged Bilingual PseudoCS	91.89	91.92	96.50	85.91
(3)+Projected Bilingual	88.61	88.09	92.91	82.39
State-of-the-art	90.56	–	**96.63**	**91.90**

Table 2: POS tagging accuracy (%) on the four corpora. Average over five runs with different random seeds. Bold and italics font indicates the best result in our experiments, while bold font indicates the best results compared to the state-of-the-art systems. We refer to BiLSTM-CRF Tagger as (1), MTL-POS Tagger that learns POS tag for related languages as (2), and MTL-POS+LID that learns jointly POS tagging and language identification as (3). Random-Initi-Embed refers to Random initialized embedding

Embedding Condition	MSA-EGY	MSA-LEV	SPA-ENG	HIN-ENG
MTL-POS+LID Tagger: (3) + Random-Initi-Embed	80.71	79.29	96.42	78.41
(3) + mono (MSA/SPA/HIN) (baselines)	77.41	78.08	88.11	71.46
(3) + mono (EGY/LEV/ENG) (baselines)	71.51	76.37	85.09	74.80
(3)+ Merged Bilingual CFM	80.06	78.39	95.49	92.54
(3)+Merged Bilingual PCS	81.33	79.28	95.02	93.20
(3)+Merged Bilingual PseudoCS	**82.15**	**81.32**	*97.20*	*94.92*
(3)+Projected Bilingual	78.17	79.11	90.01	87.69
State-of-the-art	–	–	**98.78**	**97.39**

Table 3: LID accuracy (%) on the four corpora. Average over five runs with different random seeds. Bold and italics font indicates the best result in our experiments, while bold font indicates the best results compared to the state-of-the-art systems. We refer to MTL-POS+LID that learns jointly POS tagging and language identification as (3). Random-Initi-Embed refers to Random initialized embeddings

art system (91.90%) (Bhat et al., 2018). For LID task, all our models underperform the state-of-the-art systems. Since there are no state-of-the-art systems for MSA-EGY and MSA-LEV, we compare the performance of our models against the baseline systems.

5 Discussion

Multilingual embedding (e.g., MSA-EGY and MSA-LEV) helps closely related languages (EGY/LEV) but adds noise to the languages that are distant (SPA/HIN). Similarly, learning jointly POS tagging for closely related languages yields the highest accuracy results for MSA-EGY and MSA-LEV as opposed to the languages that are distant (SPA/HIN). The accuracy results of MSA-EGY and MSA-LEV language pairs are the highest results in all experimental setups. The improvement could be attributed to the significant

number of homographs some of which are cognates.

The CS behavior can be different depending on the medium of communication, topic, speakers (or authors), and the languages being mixed among other factors. Hence, we believe that the difference in the genre of the evaluation data sets of SPA-ENG and HIN-ENG language pairs is one of the potential reasons that make both language pairs not to benefit from the multilingual embedding and learning jointly POS tagging for both language pairs. On the other hand, the MTL-POS-LID Tagger that learns simultaneously POS and language id tags with the aim of boosting the performance of POS tagging task benefit distant (SPA/HIN) more than closely related languages (EGY/LEV). We define code switching points as the points within a sentence where the languages of the words on the two sides are dif-

MSA-EGY			MSA-LEV			SPA-ENG			HIN-ENG	
Error Type	**Percentage**		**Error Type**	**Percentage**		**Error Type**	**Percentage**		**Error Type**	**Percentage**
ADJ >NOUN	19%		ADJ >NOUN	21%		NOUN >ADJ	10%		NOUN >VERB	27%
VERB >NOUN	15%		VERB >NOUN	19%		NOUN >PRON	8%		VERB >NOUN	24%
NOUN >VERB	11%		NOUN >ADJ	16%		VERB >NOUN	7%		NOUN >ADJ	19%
NOUN >ADJ	8%		NOUN >VERB	9%		ADJ >PRON	5%		ADJ >VERB	14%

Table 4: Most common errors for the best systems for all language pairs (Gold-POS > Predicted-POS)

ferent. We observe a sharp jump in the accuracy for SPA-ENG corpus. We believe the major factor of this jump is the low percentage of the CS points, $\sim 6\%$, while the percentage of CS points in the MSA-EGY, MSA-LEV, and HIN-ENG datasets are relatively high, 38.78%, 30.12%, and 15.17%. The low percentage of CS points in the SPA-ENG corpus leads the models that address the inter-sentential code switching type (BiLSTM-CRF+Merged BilingualMerged Bilingual CFM and MTL+Merged BilingualMerged Bilingual CFM) to score the second highest accuracy results, 95.40% and 95.42%.

The two key advantages of the Merged Bilingual PseudoCS embeddings and Multilingual embeddings are, 1) it enables the learned embeddings to capture the interactions between the words in different languages; 2) It captures the word usage in the context of each language and eliminates the ambiguity for the words that have the same surface form in multiple languages. Hence, the OOV percentage and ambiguity of words are reduced. Using multilingual embeddings for MSA-EGY and MSA-LEV, reduced the percentage of OOV of MSA-EGY and MSA-LEV from 10% and 13% to 8% and 10%, respectively. Similarly, with SPA-ENG and HIN-ENG language pairs, the rate of OOV is decreased from 12% and 15% to 9% and 11%, respectively.

One of the common CS intra-sentential patterns we notice in our data sets is insertion patterns. This pattern involves inserting material (lexical items, or entire constituents) from one language into a structure from the other language (Muysken et al., 2000). To evaluate the effect of the CS insertion pattern we define the CS fragment (CSF) of those test sentences. We define a CSF as the minimum contiguous span of words where a CS occurs (Soto and Hirschberg, 2018). The average length of the CS fragments in the SPA-ENG test set is 2.16, and 6.1 in the HIN-ENG test set. The average length of the CS fragments is 5.1 in MSA-LEV, and 5.8 in MSA-EGY test set. We observe that the length of the CS fragments impacts the

overall performance of the classifiers. For example, short CS fragments confuse the best classifiers of almost all language pairs. We noticed that a majority of CS sentences that have one or two lexical elements inserted had been miss-classified by almost all models in all language pairs.

Using bilingual embeddings outperform the baseline systems, but it did not achieve the highest results achieved by the other proposed models for all language pairs. The experiment results show a promising direction towards obtaining bilingual embeddings for CS tasks. Our explanation for this performance is that the sense distribution of polysemous words can differ widely between a monolingual (mono) corpus and Merged Bilingual PseudoCS corpus. For instance, the word 'bank' in English has several meanings such as (a) the land alongside or sloping, (b) a financial institution, and (c) a set or series of similar things. However, in a Spanish dominant sentence or corpus, 'bank' is primarily, if not only, used in sense (a). Our experiments show that standard bilingual embeddings are not well suited, in general, for CS tasks; embeddings learned from CS data yield better results which are aligned with our findings (Pratapa et al., 2018).

Table 4 shows the most common errors for the best systems for each language pairs. We observe almost the same trends across both MSA-EGY and MSA-LEV language pairs, while the common errors are relatively different between SPA-ENG and HIN-ENG.

6 Conclusion

In this paper, we present a detailed study of various strategies for POS tagging of CS data in four language pairs. We explore multiple strategies of measuring the impact of pretrained embeddings on POS tagging of CS data. We find that related language pairs, e.g., MSA-EGY and MSA-LEV, benefit from both jointly learning POS tagging as well as merged multilingual embeddings (i.e., pivot embedding), while distant language pairs, e.g., SPA-ENG and HIN-ENG, bene-

fit from a multi-task learning model that learns two different tasks, e.g., POS tagging and language identification. Furthermore, we compared our results to the previous state-of-the-art POS tagger for MSA-EGY, SPA-ENG, and HIN-ENG and showed that our classifiers outperform the MSA-EGY state-of-the-art system in every configuration (AlGhamdi et al., 2016). The results achieved by BiLSTM-CRF+Merged Bilingual PseudoCS embeddings model is comparable to Soto and Hirschberg (2018). We will explore several directions in the future. First, we will study the theoretical aspects of word embedding learning. Second, we will investigate the proposed word embeddings on other downstream NLP applications, such as segmentation and parsing.

Acknowledgments

References

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*, volume 1, pages 937–947.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147. Association for Computational Linguistics.

Mohamed Al-Badrashiny, Heba Elfardy, and Mona T. Diab. 2015. Aida2: A hybrid approach for token and sentence level dialect identification in arabic. In *Proceedings of the 19th Conference on Computational Natural Language Learning, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–51. ACL.

Fahad AlGhamdi and Mona Diab. 2018. Wasa: A web application for sequence annotation. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Fahad AlGhamdi, Giovanni Molina, Mona T. Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching@EMNLP 2016, Austin, Texas, USA, November 1, 2016*.

Utsab Barman, Joachim Wagner, and Jennifer Foster. 2016. Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 30–39. Association for Computational Linguistics.

Irshad Bhat, Riyaz A Bhat, Manish Shrivastava, and Dipti Sharma. 2017. Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 324–330.

Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava, and Dipti Misra Sharma. 2018. Universal dependency parsing for Hindi-English code-switching. *arXiv preprint arXiv:1804.05868*.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*. Springer Press.

Phil Blunsom and Karl Moritz Hermann. 2014. Multilingual models for compositional distributional semantics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W. Black. 2018. Language informed modeling of code-switched text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 92–97. Association for Computational Linguistics.

Hongshen Chen, Yue Zhang, and Qun Liu. 2016. Neural network for heterogeneous annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 731–741.

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

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

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

Mona Diab, Mahmoud Ghoneim, Fahad AlGhamdi, Nada AlMarwani, and Mohamed Al-Badrashiny. 2016. Creating a large multi-layered representational repository of linguistic code switched arabic data. In *10th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74.

Mona Diab, Nizar Habash, Owen Rambow, and Ryan Roth. 2013. LDC Arabic treebanks and associated corpora: Data divisions manual. *arXiv preprint arXiv:1309.5652*.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016. Creating resources for dialectal Arabic from a single annotation: A case study on egyptian and levantine. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3455–3465. The COLING 2016 Organizing Committee.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed English-Hindi Twitter and Facebook chat messages. *RECENT ADVANCES IN*, page 239.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn Arabic treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo.

Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi, and Sondos Krouna. 2012. Egyptian Arabic Treebank pilot.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on padic: A parallel Arabic dialect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49. Association for Computational Linguistics.

Pieter Muysken, Carmen Pena Díaz, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*, volume 11. Cambridge University Press.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi syntax: Annotating dependency, lexical predicate-argument structure, and phrase structure. In *The 7th International Conference on Natural Language Processing*, pages 14–17.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic Gigaword fifth edition LDC2011t11. *Philadelphia: Linguistic Data Consortium*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2011. A Universal Part-of-Speech Tagset. *CoRR*, abs/1104.2086.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3067–3072. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark.

Kalika Bali Royal Sequiera, Monojit Choudhury. 2015. POS tagging of Hindi-English Code Mixed Text from Social Media: Some Machine Learning Experiments. In *Proceedings of International Conference on NLP*. NLPAI.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview of the first shared task on language identification in code-switched data. In *First Workshop on Computational Approaches to Code Switching, EMNLP-2014*, Doha, Qatar. Association for Computational Linguistics.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing Universal Part-Of-Speech tags for Code-Switching. In *Interspeech 2017*, pages 77–81.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language id tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014a. 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, Doha, Qatar. Association for Computational Linguistics.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014b. 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.

Omar F Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 37–41. Association for Computational Linguistics.

Yuan Zhang and David Weiss. 2016. Stack-propagation: Improved representation learning for syntax. *arXiv preprint arXiv:1603.06598*.

Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398.