A Neural Network Model for Part-Of-Speech Tagging of Social Media Texts

Sara Meftah*, Nasredine Semmar*, Fatiha Sadat+
*CEA, LIST, Vision and Content Engineering Laboratory
F-91191, Gif-sur-Yvette, France
{sara.meftah, nasredine.semmar}@cea.fr
+Université du Québec à Montréal, UQÀM
201 Président Kennedy Avenue H2X 3Y7, Montréal, Canada
sadat.fatiha@uqam.ca

Abstract

In this paper, we propose a neural network model for Part-Of-Speech (POS) tagging of User-Generated Content (UGC) such as Twitter, Facebook and Web forums. The proposed model is end-to-end and uses both character and word level representations. Character level representations are learned during the training of the model through a Convolutional Neural Network (CNN). For word level representations, we combine several pre-trainned embeddings (Word2Vec, FastText and GloVe). To deal with the issue of the poor availability of annotated social media data, we have implemented a Transfer Learning (TL) approach. We demonstrate the validity and genericity of our model on a POS tagging task by conducting our experiments on five social media languages (English, German, French, Italian and Spanish).

Keywords: Part-of-Speech Tagging, Social Media Texts, Low-Resources Languages, Neural Networks, Transfer Learning

1. Introduction

Recent approaches based on end-to-end Deep Neural Networks (DNNs) have shown promising results for Natural Language Processing (NLP). Most of proposed neural models for sequence labeling (including POS taggers) use Recurrent Neural Networks (RNNs) and its variants (Long Short-Term Memory networks - LSTMs and Gated Recurrent Units - GRUs), and Convolutional Neural Networks (CNNs) for character-level representations. Indeed, previous studies (Jozefowicz et al., 2016) have shown that CNNs represent an effective approach to extract morphological information (root, prefix, suffix, etc.) from words and encode it into neural representations, especially for morphological rich texts (Chiu and Nichols, 2015; Ma and Hovy, 2016).

The actual performance of POS taggers trained from treebanks in the newswire domain, such as the Wall Street Journal (WSJ) corpus of the Penn TreeBank (PTB) (Marcus et al., 1993) and evaluated on in-domain data is close to human level, thanks to deep learning techniques trained on huge annotated datasets (97.64% accuracy by (Choi, 2016)). Contrariwise, approaching human-level accuracy on more complex domains such as User Generated Content (UGC) on social media is still a hard problem. Especially conversational texts (Twitter, Web blogs, SMS texts, etc.). This is due to the conversational nature of the text, the lack of conventional orthography, the noise, linguistic errors, spelling inconsistencies, informal abbreviations and the idiosyncratic style. Also, Twitter poses an additional issue by imposing 280 characters limit for each tweet.

The application of models trained on well-structured corpora such as WSJ fails to work effectively on noisy text. As illustrated in (Gimpel et al., 2011), the accuracy of the Stanford POS tagger (Toutanova et al., 2003) trained on WSJ falls from 97% on standard English to 85% accuracy on tweets. The main reason for this drop in accuracy is that tweets contain lot of Out-Of-Vocabulary (OOV) words compared to standard text. In addition, NLP's DNNs models often require to be trained on huge volumes of annotated data to produce powerful models and prevent over-fitting. Hence, the construction of a DNN model for UGC data needs huge amounts of annotated data with POS labels to provide high performances. However, available annotated in-domian datasets are very small.

In this paper, we present a POS tagger for multiple social media datasets, using a Transfer Learning (TL) based end-to-end neural model. In a TL scenario, the knowledge learned by handling one problem is used to help solving different but related problems.

The goal of this work is to examine the effectiveness of TL for POS tagging across domains and tasks. Experiments show significant improvements over several languages (English, French, German, Italian and Spanish).

2. Related Work

Our work is related to two lines of research: (1) Transfer Learning (2) POS tagging of social media texts. Below we discuss the state-of-the-art of each one.

2.1. Transfer Learning

As discussed in the introduction, high performing NLP's neural models often require huge volumes of annotated data to produce powerful models and prevent over-fitting. Consequently, in the case of social media content, it is difficult to achieve the performances of state-of-the-art models based on hand-crafted features by applying neural models trained on small amounts of annotated data. For this reason TL was proposed to exploit huge annotated out-of-domain data-sets. TL aims at performing a task on a target dataset using features learned from a source dataset (Pan and Yang, 2010).

Furthermore, the successes of neural models for many tasks over the last few years have intensified the interest for studying TL for neural networks.

In particular, TL was largely exploited in computer vision using pre-trained CNNs to generate representations for

novel tasks; some of the parameters learned on the source dataset are used to initialize the corresponding parameters of the CNNs for the target dataset.

In the past few years, few studies have been conducted on TL for neural based models in the field of NLP. It consists in performing a task on a low-resource target problem using features learned from a high-resource source problem. For instance, TL has been successfully applied in neural speech processing and machine translation (Zoph et al., 2016).

Two studies have been recently performed on TL for neural networks based models in sequence labeling: Yang et al. (2017) examined the effects of TL for deep hierarchical recurrent networks across domains, applications, and languages, and showed that significant improvement can be obtained. Lee et al. (2017) used cross-domain TL for Named Entity Recognition (NER) (specifically patient note de-identification), and showed that TL may be especially beneficial for a target dataset with small number of examples.

2.2. Part-Of-Speech Tagging of Social Media Texts

POS tagging is a sequence labeling problem, by assigning to each word its disambiguate part-of-speech (Verb, Noun, Adjective, etc.) in the sentential context in which the word is used. This information is useful for higher-level NLP applications such as semantic relations extraction, sentiment analysis, automatic summarization and machine translation.

Most performing traditional POS tagging models for social media content are linear statistical models, including Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMMs), Conditional Random Fields (CRF) and linear classifiers like SVM-based taggers.

There are two principal state-of-the-art works for English tweets POS tagging, both based on hand-crafted features, Ritter et al. (2011) published a set of 787 hand-annotated English tweets and proposed in (Derczynski et al., 2013) a model based on hidden Markov Models and a set of normalization rules, external dictionaries and lexical features. Gimpel et al. (2011) and Owoputi et al. (2013) constructed 1827 and 547 hand-annotated tweets, respectively, using the same tag-set. They proposed a model based on First-order maximum entropy Markov model (MEMM), engineered features like brown clustering and lexical features.

Nooralahzadeh et al. (2014) proposed a POS tagging system for French Social Media content using Conditional random fields (CRFs) with a set of several hand-crafted features.

These models rely heavily on hand-crafted features and task specific resources (morphological, orthographic and lexical features and external resources such as gazetteers or dictionaries). However, such task-specific knowledge is costly to develop and making sequence labelling models difficult to adapt to new tasks or new domains.

Recently a neural network model for English tweets POS tagging was proposed by Gui et al. (2017) (TPANN), they used Adversarial Neural Networks to leverage huge amounts of unlabeled tweets and labeled out-of-domain data (WSJ). TPANN achieves high performances compared to the former works. The model proposed in (Gui et al., 2017) requires that labeled in-domain-data and labeled out-of-domain data share the same tag-set (a mapping is necessary in case of tag-sets mismatch).

3. Contributions

This work is built on the basis of the recently published paper (Meftah et al., 2017), where cross-domain TL was successfully used for English tweets POS tagging by exploiting available huge amounts of POS labeled corpora of a similar domain (standard English). The knowledge learned on the parent neural network trained on enough standard English labeled data was transferred to initialize the child network, further fine-tuned on small annotated English Twitter corpus. Nevertheless, the present paper includes the following new contributions:

- We investigate a second scenario, cross-task TL, where the parent network is trained on in-domain data annotated with Named Entities (NE).
- We show that TL method is efficient on multiple social media languages (English, French, Spanish, German and Italian).
- We analyze how cross-task TL may address the issue of the low-availability of annotated data and improve performances.

4. Neural Model Architecture

The neural model that we use for TL experiments is the same used in (Meftah et al., 2017), based on bidirectional hierarchical Gated Recurrent Units (GRUs). Figure 2 shows an overview of the model's architecture[1].

4.1. Features Representation

In order to preserve both semantic and syntactic information of words, each word from the input sequence is represented by a combination of two vectors of features, character-level and word-level embedding. Therefore, each word in the input sentence is represented by a combination of two vectors:

1. Pre-trained words embedding: We initialize word-level embedding with a concatenation of different pre-retrained words embedding (details in section 6.3.) to accurately capture words' semantics.
2. Character level embedding: To learn orthographic features at the character level, we use a CNN architecture similar to that of Ma and Hovy (2016). As illustrated in figure 1. Each word is represented with a $v \times l$ dimensional matrix, next it's embedded into a $d \times l$ dimensional matrix, where v is character's vocabulary size, l is the maximal length of words and d is character embedding's dimension. Then, we take the character embeddings and apply (30×3)-stacked convolutional layers, followed by a max-pooling operation. Finally, the result is passed to a fully-connected layer using a Rectifier Linear Unit (ReLU) activation function.

[1]The model's architecture is the same among all datasets and tasks.

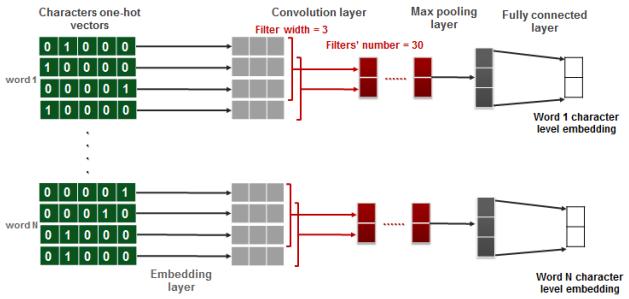

Figure 1: Convolutional Neural Network architecture for character-level embedding.

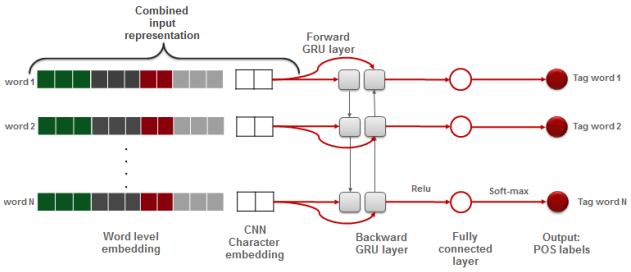

Figure 2: Overall system design. First, the system embeds each word of the current sentence into two representations: character level representation using a CNN network and a word level representation by combining different pre-trained models. Then, the two representations are combined and fed into a bidirectional GRU layer, the resulting vector is fed to a fully connected layer and finally a softmax layer to perform POS tagging.

4.2. Sequence Labelling with Gated Recurrent Units (GRUs) Layer

Word vectors (the combination between character level embedding and word level embedding CNN) are fed into a 100 dimension Gated Recurrent Units (GRUs) layer, a variant of RNNs.

Let $(x_1, x_2, ..., x_t, ..., x_n)$ the input sequence of the GRUs layer, which is in our case a sequence of n D-dimensional word vectors, where n is sentence's length and D is word vectors' dimension.

Let h_t be the GRU hidden state at time-step t. Formally, a GRU unit at a time-step t takes x_t and the previous hidden state h_{t-1} as input, and outputs the current hidden state h_t. Each gated recurrent unit can be expressed as follows:

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1}) \quad (1)$$

$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1}) \quad (2)$$

$$\hat{h}_t = \tanh(W_{hx}x_t + W_{hh}(r_t \otimes h_{t-1})) \quad (3)$$

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes \hat{h}_t \quad (4)$$

Where W's are model parameters of each unit, $\hat{h}_t$ is a candidate hidden state that is used to compute h_t, σ is an element-wise sigmoid logistic function defined as $\sigma(x) = 1/(1 + e^{-x})$, and $\otimes$ denotes element-wise multiplication of two vectors. The update gate z_t controls how much the unit updates its hidden state, and the reset gate r_t determines how much information from the previous hidden state needs to be reset.

4.3. Fully-connected Layer and Softmax Layer

The output of the forward GRUs and the backward GRUs at each time-step are combined and fed through a 80 dimension linear (fully connected) layer with a ReLU activation, followed by a final dense layer with a softmax activation to generate a probability distribution over the output classes at each time-step.

5. Transfer Learning Approach

TL is applied to address the problem of the need in annotated data for POS tagging of social media texts. It consists in learning a parent neural network on a source problem with enough data, then transferring a part of its weights to represent data of a target problem with few training examples.

We experiment two scenarios of TL. The first scenario is cross-domain transfer; knowledge is transferred from a source domain to a target domain. In our case, the source domain is the standard form (well-established) of a language and the target domain is the social media text of the same language. The source and the target problems are trained for the same task (POS tagging), even if source and target datasets do not share the same tag-set.

As illustrated in the figure 3, we have a parent neural network N_p with a set of parameters θ_p splitted into two sets: $\theta_p = (\theta_p^1, \theta_p^2)$. And a child network N_c with a set of parameters θ_c splitted into two sets: $\theta_c = (\theta_c^1, \theta_c^2)$.

(1) We learn the parent network on annotated data from the source problem on a source dataset D_s. (2) We transfer weights of the first set of parameters of the parent network N_p to the child network N_c: $\theta_c^1 = \theta_p^1$. (3) Then, the child network is fine-tuned to the target problem by training it on the target dataset D_c.

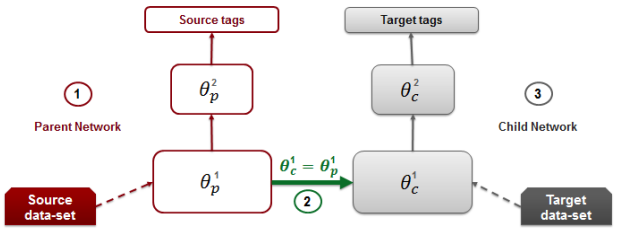

Figure 3: Cross-domain Transfer Learning scheme.

The second scenario is cross-task transfer; the source and the target problems share the same domain and the same language (social media text of the same language). However, tasks are different (The source problem's task is NER and the target's is POS tagging) to exploit the underlying similarities of the two tasks.

As illustrated in the figure 4, the parent neural network and the child network share the same first set of parameters (The feature extractor) : $\theta_p^1 = \theta_c^1 = \theta^1$.

θ^1 are jointly optimized by the two tasks, while task specific parameters θ_c^2 and θ_p^2 are trained for each task separately.

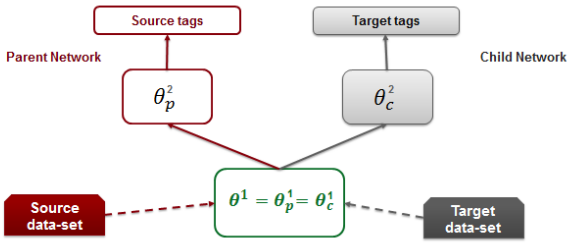

Figure 4: Cross-task Transfer Learning scheme.

6. Experimental Setup

6.1. Datasets

We use two types of source datasets for parent neural network training: (1) large out-of-domain POS-labeled data from resource-rich domains for the first TL scenario, and (2) NE-labeled in-domain data for the second scenario. For child neural network model fine-tuning, we use small POS labeled in-domain datasets.
In this section, we report the source and target datasets for each language, on which we perform our evaluations. The statistics of the datasets[2] are described in table 1.

6.1.1. English

As a source dataset for English experiments, we use a standard English corpus, the **Wall Street Journal (WSJ)** part of the PTB, annotated with 36 POS tags.
We evaluate our approach on three target datasets. The **NPS IRC Chat Corpus** (Forsythand and Martell, 2007) of 10,567 posts gathered from various online chat services. And two Twitter datasets:

- **The T-PoS corpus** of 787 hand-annotated English tweets, introduced by (Ritter et al., 2011), which uses the same tag-set as PTB's (Marcus et al., 1993), plus four Twitter special tags: URL for urls, HT for hashtags, USR for username mentions and RT for retweet signifier (40 tags in total). For our experiments on T-PoS, we use the same data splits used in (Derczynski et al., 2013); 70:15:15 into training, development and test sets named T-train, T-dev and T-eval.
- **The ARK corpus** was published on two parts, the first, Oct27 of 1827 hand-annotated English tweets, published in (Gimpel et al., 2011) and the second, Daily547 of 547 tweets published by Owoputi et al. (2013), using a novel and coarse grained tag-set (25 tags). For example, its V tag corresponds to any verb, conflating PTB's VB, VBD, VBG, VBN, VBP, VBZ, and MD tags. We split the Oct27 dataset into training-set and development-set (70:30) (data splits portions are not mentioned in original papers) and Daily547 as a test set.

6.1.2. French

As a source dataset for French experiments, we use a standard French corpus, **French-Tree-Bank (FTB)** (Abeillé et al., 2003), a POS-annotated French newspaper corpus.
We evaluate our approach on two publicly available POS-labeled User Generated (UG) French content datasets:

- **The French web 2.0 (Fr2.0) (Seddah et al., 2012)** is a set of 1700 sentences extracted from various types of French Web. (1) Micro-blogging: Facebook and Twitter. (2) web forums: French health forum DOCTISSIMO[3] and video games website JEUXVIDEOS[4].

  The tag-set includes 28 POS tags from FTB, plus combined tags for contracted tokens. For instance, the non-standard French contraction *tes* (widely used by French web's users), which stands for *tu es*, would have been tagged CLS and V (subject clitic and finite verb) in FTB. The non-standard contracted token *tes* is then tagged CLS+V. And specific tags to social media, including HT and RT. Twitter at-mentions as well as urls and e-mail addresses have been tagged NPP which is the main difference with other annotations of UG content.
- **ExtremeUGC dataset (UGC) (Alonso et al., 2016)**: contains user-generated content from three different sources. Two of them are logs of multi-player video-game chat sessions: MINECRAFT and LEAGUE OF LEGENDS, the last one is cooking-related user questions from MARMITON, a popular cooking French website. Datasets are annotated with the same scheme as the Fr2.0.

6.1.3. Spanish, Italian and German:

The xLiMe Twitter Corpus (Rei et al., 2016): is a Multilingual Social Media Linguistic Corpus, contains manually annotated Spanish, German and Italian tweets [5]. The corpus is annotated with POS tags and NE. The POS tag-set consists of the Universal Dependencies tag-set, plus Twitter specific tags based on (Gimpel et al., 2011). For NE, they used the same tag-set used in CoNLL-2003 Shared Task (Person, Location, Organization, and Miscellaneous). Since there is no standard training/dev/test data split for xLime corpora, we randomly split it 80:10:10 into training, development and test sets.

6.2. Baselines

We compare the performance of our system to performances of prior works described in section 2.2.:

6.2.1. English

- Derczynski et al. (2013) performed experiments on T-PoS corpus. For training, they used T-train (2.3K tokens), 50K tokens from the WSJ part of the PTB and 32K tokens from the NPS IRC corpus, achieving an accuracy of 88.69% on T-eval. Furthermore, they achieved 90.54% token accuracy using supplementary 1.5M training tokens annotated by vote-constrained bootstrapping.

[2] All corpora are in the CoNLL format, they are already tokenized.

[3] forum.doctissimo.fr

[4] www.jeuxvideo.com

[5] http://nl.ijs.si/janes/wp-content/uploads/2016/09/A-Multilingual-Social-Media-Linguistic-Corpus.html

Language	Domain	Corpus	Task	# Sentences	# Tokens
English	Source	WSJ	POS	67,786	1,2M
	Target	NPS	POS	10,567	45,000
	Target	T-POS	POS	787	15,000
	Target	Ark dataset (Oct27 + Daily547)	POS	1,827 + 547	26,594 + 7,707
French	Source	FTB	POS	21,634	624,187
	Target	French Web 2.0	POS	1,700	20,557
	Target	ExtremeUGC	POS	974	8,099
Spanish	Source	xLime Spanish NER	NER	7,668	140,852
	Target	xLime Spanish POS	POS	7,668	140,852
German	Source	xLime German NER	NER	3,400	60,873
	Target	xLime German POS	POS	3,400	60,873
Italian	Source	xLime Italian NER	NER	8,601	162,269
	Target	xLime Italian POS	POS	8,601	162,269

Table 1: Statistics of the different source and target datasets used in this paper.

- Owoputi et al. (2013) performed experiments on T-Pos, Ark and NPS corpora, achieving 90.40%, 93.20% and 93.4% accuracy respectively.

- Gui et al. (2017) performed experiments on T-PoS, ArK and NPS datasets, achieving 90.92%, 92.80% and 94.1% accuracy respectively. For training, they leverage 1,17M token from unlabeled tweets and more than 1,17M from labeled WSJ. In order to use WSJ labeled data in experiments on ARK dataset, they performed a mapping between PTB and ARK tag-sets.

6.2.2. French

- Nooralahzadeh et al. (2014) proposed a French POS tagging system using a discriminative sequence labeling model (CRF). They achieved 91.9% accuracy on Fr2.0 corpus. The same system setup was evaluated on T-POS and NPS English corpora achieving 90.1% and 92.7% accuracy respectively.

- Alonso et al. (2016) experimented POS tagging on ExtremeUGC dataset using Melt tagger (Denis and Sagot, 2009) with a set of normalization rules, achieving 84.72% accuracy.

6.2.3. Spanish, German and Italian

Rei et al. (2016) reported inter-Annotator Agreement per language on xLime dataset, 88% for German, 87% for Italian and 85% for Spanish.

6.3. Word Embedding

Words embeddings initialization is computed by a look-up table of each of pretrained model. All words are lowercased before passing through the look-up table for conversion to their corresponding vectors.

Multiple sets of published pre-trained vectors are publicly available for English. Experiments in (Meftah et al., 2017) showed that an initialization with a combination of several pre-trained embedding vectors (from different pre-trained models) improves significantly the performances. Therefore, for English experiments, we initialize word embedding with a concatenation of four pre-trained models:

1. Word2vec (Mikolov et al., 2013), trained on part of Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words.

2. FastText (Bojanowski et al., 2016), which is very similar to Word2vec (Using SkipGram) but it also uses sub-word information in the prediction model. FastText Facebook embedding is trained on Wikipedia for 294 languages and contains 300-dimensional words vectors.

3. GloVe (Pennington et al., 2014) is a model based on global word-word co-occurrence statistics. We use two Glove's models. The first, which we name "Glove", trained on 42 billions words from a web crawling, contains 300-dimensional vectors for 1.9M words. And the second, which we name "Glove-Twitter", trained on 2 billion tweets, contains 200-dimensional vectors for 1.2M words.

For experiments on French, Spanish, German and Italian, we use FastText 300-dimensional pre-trained embedding vectors trained on Wikipedia.

6.4. Transfer Learning Setup

The first scenario (cross-domain TL) is evaluated on English and French, following three main phases: (1) training the parent network on the source problem on rich out-of-domain data (WSJ for English and FTB for French) (2) transferring weights of the first set of parameters to the target problem. These weights are used to initialize the child model's first set of parameters, rather than starting from a random position[6]. And finally (3) fine-tinning the child network on low-resource in-domain data.

Since we have multiple target datasets for each language, a jointly training is performed in the step of child model's fine-tuning[7].

[6]The weights of the second set of parameters of the child model are randomly initialized.

[7]Using a smaller learning rate for weights that will be fine-tuned (first set of weights), in comparison to the randomly initialized weights (second set of weights) lead to slightly improvements.

Language	English			French		Spanish	German	Italian
Dataset	T-Pos	ARK	NPS	Fr2.0	UGC	xlime		
Acc. without transfer Learning (%)	89.13	91.33	92.9	91.14	87.89	90.87	90.1	89.41
Acc. with transfer learning (%)	**90.90**	**92.01**	**93.2**	**91.99**	**88.07**	**91.03**	**90.33**	**89.66**

Table 2: Our system accuracy (acc.) with and without transfer learning. Cross-domain transfer is performed for English and French and cross-task transfer for Spanish, German and Italian.

Method	**Acc. T-Pos (%)**	**Acc. ARK (%)**	**Acc. NPS (%)**
Derczynski et al. (2013)	88.69	–	–
Owoputi et al. (2013)	90.40	**93.20**	93.4
Nooralahzadeh et al. (2014)	90.1	–	92.7
Gui et al. (2017)	**90.92**	92.8	**94.1**
Our results	*90.90*	*92.01*	*93.2*

Table 3: Our system's performance on English social media datasets compared to state-of-the-art works.

The second scenario (cross-task TL) is evaluated on Spanish, German and Italian, we use TL approach by a jointly training of source and target tasks (NER and POS).

The training procedure for cross-task TL is as follows: At each epoch, we perform training on a batch from both datasets (source and target), and then, we perform the parameters optimization according to the loss function of the given task (The shared parameters are optimized to improve the performances of both tasks. However, each set of task's parameters is optimized only to improve the corresponding task). Training on NER is stopped before the POS tagging in order to preserve more specific features of the POS tagging task.

Mou et al. (2016) showed that the features represented by the lowest layers of neural networks are more general than topmost layers features in NLP applications. And the knowledge to transfer from the parent network to the child network depends on the relatedness of the source and the target tasks and data-sets. For this purpose, we followed the same experiments realized in (Meftah et al., 2017) to study the transferability of each layer of the neural network for each dataset, and to choose the set of layers to transfer from the parent problem to the child problem.

6.5. Training Settings

All experiments described in this section are implemented using the PyTorch library. The hyper-parameters have been chosen using cross-validation on the reported splits (In section 6.1.) for all the results reported in the following section. We use the Adam optimizer in all experiments. We set the character embedding dimension at 30, the dimension of hidden states of the GRUs layer at 100 and fully connected layer (FCL) dimension at 80. We use dropout training before the input to LSTM and FCL layers with a probability in order to avoid overfitting

7. Results and Discussion

7.1. Transfer Learning Performances

In this section, we compare in table 2 performances of the neural network model described in section 4. trained only on target dataset (without TL) against the neural network trained with TL. We can see that the TL method improves results on all languages.

Table 2 further shows that the improvements made by cross-domain TL (English and French results) are more important than improvements made by cross-task TL (Spanish, German and Italian results). This phenomenon can be explained by the fact that the underlying similarities between the source task and the target task are less transferable, hence the improvement is less substantial.

Additionally, an interesting note on French experiments, where the improvement brought by TL is more important on the French social media 2.0 (Fr2.0) dataset (+0.85%) compared to ExtremeUGC dataset (0.28%), that can be explained by the fact that the last dataset is more noisy (high divergence from the source dataset FTB) than Fr2.0.

We can also observe that the improvement brought by TL is more important on the T-Pos dataset (+1.77%) compared to Ark dataset (0.68%), that can be explained by the fact that T-POS dataset have similar tokenization and annotation scheme than the source dataset (PTB) in contrast to Ark dataset.

7.1.1. Cross-task TL performances

In order to understand how cross-task TL improves POS tagging performances on Spanish, German and Italian social media content. In particularly, which POS tags benefit more from transferring knowledge from NER task. Table 6 shows an important improvement on the accuracy of the POS tag "Noun" compared to the overall accuracy.

We provide an example in the table 7, where cross-task TL helps to assign the correct tag to the Spanish word *Internacional* (i.e International in English), tagged as an adjective by the model without TL. Although, the word *Internacional* is an adjective in most cases. However, in this case *Amnistía Internacional* is an organization, and the information brought by NER task helps to solve the ambiguity.

7.2. Comparison with State-of-the-art Results

In tables 3, 4 and 5, we show our system's performances compared to state-of-the-art results. We can see that our results are competitive compared to the state-of-the-art systems.

Method	Acc. Fr2.0 (%)	Acc. UGC (%)
Nooralahzadeh et al. (2014)	91.9	–
Alonso et al. (2016)	–	84.72
Our results	**91.99**	**88.07**

Table 4: Our system's performance on French social media datasets compared to state-of-the-art works.

Method	Acc. Spanish (%)	Acc. German (%)	Acc. Italian (%)
inter-Annotator Agreement	85	88	87
Our results	**91.03**	**90.33**	**89.66**

Table 5: Our system's performance on xLime datasets compared to inter-Annotator Agreement.

Language		Sp	Ger	IT
W/o TL	**Overall acc. (%)**	90.87	90.1	89.41
	Acc. on nouns (%)	*92.54*	*91.98*	*94.12*
W TL	**Overall acc. (%)**	91.03	90.33	89.66
	Acc. on nouns (%)	*97.65*	*98.02*	*98.2*

Table 6: Improvement of the accuracy of the tag "Noun" compared to the improvement of the overall accuracy after using cross-domain transfer learning, on Spanish (Sp), German (Ger) and Italian (IT).

W/o TL	... de/**ADP** Amnistía/**NOUN** Internacional/**ADJ** :/**.** #EEUU/**#** ...
W TL	... de/**ADP** Amnistía/**NOUN** Internacional/**NOUN** :/**.** #EEUU/**#** ...

Table 7: Our model POS tagging example of a Spanish tweet, without TL (W/o TL) in the first line and with TL (W TL) in the second line.

Tables 4 and 5 show that our model outperforms state-of-the-art systems on French, Spanish, German and Italian.
On table 3, we can see that our method outperforms state-of-the-art approaches (Derczynski et al., 2013) and (Owoputi et al., 2013) on T-POS experiments. However, it performs worse than (Owoputi et al., 2013) on ARK-dataset. Our model is end-to-end and the most of errors in our system were caused by hashtags and proper nouns. These issues were resolved in (Owoputi et al., 2013) by adding external knowledge (a list of named entities) and rules to detect hashtags, etc.
We can observe that (Gui et al., 2017) achieves better performances than our model in both datasets, an effective model (Adversarial Neural Network) was used in their work with huge amounts of unlabeled in-domain-data (More than 1.17 millions token) and 1 million token from WSJ. In addition, they used regular expressions to perfectly tag Twitter-specific tags: retweets, @usernames, hashtags, and urls, contrariwise our model which is end-to-end and does not use hand-crafted rules.

7.3. Improving Results Using Pre-processing

As discussed in the above section, most of errors in our system were caused by Twitter specific tokens (e.g. our model accuracy on T-POS dataset on hashtags is equal to 62%). In this section, we normalize Twitter specific tokens, we substitute every word starting with a # character by a special token. Similarly, all words starting with the prefix http are replaced by a url token.
Table 8 illustrates our model performances on English datasets after pre-processing rules for hashtags, urls, usernames and at-mentions. We can see a significant improvement on accuracy on all of English datasets, outperforming state-of-the-art work (Gui et al., 2017) on T-Pos dataset.

Method	Without prep	With prep
Acc. on T-Pos (%)	90.90	**91.03**
Acc. on ARK (%)	92.01	**92.6**
Acc. on NPS (%)	93.2	**93.41**

Table 8: Performances on English social media datasets before (Without prep) and after (With prep) the integration of pre-processing rules for hashtags, Urls, usernames and at-mentions.

8. Conclusion

This paper presented a neural network model using Transfer Learning (TL) for Part-of-speech (POS) tagging of social media texts. Two scenarios of TL were experimented. The first is cross-domain TL, where we leverage available huge amounts of POS-labeled standard English and French to improve English and French social media texts POS tagging. The second scenario is cross-task TL, where we use named entities labeled data to improve POS tagging of Spanish, German and Italian social media texts. Our experiments show that both scenarios of TL improve the performance of the POS tagging task. For future work, we plan to model the similarities and differences between the source and target languages in order to incorporate this external linguistic knowledge in the neural network model. In addition to that, we aim to conduct a study on morphologically rich and complex languages such as Arabic that is well known for its diverse dialects (22 dialects distributed over 5 regional categories) that we can find on social media.

9. Acknowledgments

This research work is supported by the ASGARD project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700381.

10. Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In *Treebanks*, pages 165–187. Springer.

Alonso, H. M., Seddah, D., and Sagot, B. (2016). From noisy questions to minecraft texts: Annotation challenges in extreme syntax scenarios. *WNUT 2016*, page 13.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Chiu, J. P. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Choi, J. D. (2016). Dynamic feature induction: The last gist to the state-of-the-art. In *HLT-NAACL*, pages 271–281.

Denis, P. and Sagot, B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, volume 1.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206.

Forsythand, E. N. and Martell, C. H. (2007). Lexical and discourse analysis of online chat dialog. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 19–26. IEEE.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.

Gui, T., Zhang, Q., Huang, H., Peng, M., and Huang, X. (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 2401–2410.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Lee, J. Y., Dernoncourt, F., and Szolovits, P. (2017). Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Meftah, S., Semmar, N., Zennaki, O., and Sadat, F. (2017). Using transfer learning in part-of-speech tagging of english tweets.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2016). How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.

Nooralahzadeh, F., Brun, C., and Roux, C. (2014). Part of speech tagging for french social media data. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1764–1772.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Rei, L., Mladenic, D., and Krek, S. (2016). A multilingual social media linguistic corpus. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana, Slovenia*.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Seddah, D., Sagot, B., Candito, M., Mouilleron, V., and Combet, V. (2012). The french social media bank: a treebank of noisy user generated content. In *COLING 2012-24th International Conference on Computational Linguistics*.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Yang, Z., Salakhutdinov, R., and Cohen, W. W. (2017). Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv preprint arXiv:1703.06345*.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.