applied sciences

MDPI

Article

A Multitask-Based Neural Machine Translation Model with Part-of-Speech Tags Integration for Arabic Dialects

Laith H. Baniata [1], Seyoung Park [1,*] and Seong-Bae Park [2]

[1] School of Computer Science and Engineering, Kyungpook National University, 80 Daehakro, Buk-gu 41566, Daegu, Korea; laith@knu.ac.kr

[2] Department of Computer Science and Engineering, Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin 17104, Gyeonggi, Korea; sbpark71@khu.ac.kr

* Correspondence: seyoung@knu.ac.kr; Tel.: +82-10-7293-3162

Received: 26 October 2018; Accepted: 29 November 2018; Published: 5 December 2018

Abstract: The statistical machine translation for the Arabic language integrates external linguistic resources such as part-of-speech tags. The current research presents a Bidirectional Long Short-Term Memory (Bi-LSTM) - Conditional Random Fields (CRF) segment-level Arabic Dialect POS tagger model, which will be integrated into the Multitask Neural Machine Translation (NMT) model. The proposed solution for NMT is based on the recurrent neural network encoder-decoder NMT model that has been introduced recently. The study has proposed and developed a unified Multitask NMT model that shares an encoder between the two tasks; Arabic Dialect (AD) to Modern Standard Arabic (MSA) translation task and the segment-level POS tagging tasks. A shared layer and an invariant layer are shared between the translation tasks. By training translation tasks and POS tagging task alternately, the proposed model can leverage the characteristic information and improve the translation quality from Arabic dialects to Modern Standard Arabic. The experiments are conducted from Levantine Arabic (LA) to MSA and Maghrebi Arabic (MA) to MSA translation tasks. As an additional linguistic resource, the segment-level part-of-speech tags for Arabic dialects were also exploited. Experiments suggest that translation quality and the performance of POS tagger were improved with the implementation of multitask learning approach.

Keywords: NMT; MTL; POS tagging; CRF; Bi-LSTM; Arabic dialects; encoder; decoder; MSA

1. Introduction

Approaches of neural networks have gained significant importance in different tasks and applications. In speech recognition, Hung et al., [1] proposed to exploit robust principle component analysis (RPCA) to extract the sparse partition of the FBANK/MFCC features of noise-corrupted utterances, which corresponds to the speech-dominant elements. The resulting RPCA-wise features show significantly improved recognition performance relative to the original FBANK/MFCC features under the AURORA-4 median-vocabulary recognition task with the renowned deep neural network (DNN) as the acoustic models. Furthermore, the newly proposed features reveal good addition to the popular normalization methods, mean normalization (MN) and relative spectral (RASTA), as the respective integration provides further improvement in recognition accuracy compared with the individual component method. In electronics, Chowdhury et al. [2] propose MB-CNN, a memristive accelerator for binary convolutional neural networks that perform XNOR convolution in-situ novel 2R memristive data blocks to improve power, performance and memory requirements of embedded mobile devices. In energy disaggregation, Salerno et al. [3] proposed to apply extreme learning machine (ELMs) to Non-intrusive appliance load monitoring (NILM). ELMs, both in their shallow

and hierarchical configurations, act as a non-linear signal enhancement system allowing them to recover the power load of the target appliance from the aggregated load. The authors proposed to extend ELMs to the energy disaggregation problem and they have reported top performance on the UK-DALE dataset. Recent work has shown that exploiting a neural network can significantly improve the performance of Machine Translation (MT) and give a better translation quality. Neural Machine Translation (NMT), structured in a refined end-to-end architecture is incorporated of the capacity to raise the translation equality. One of the common issues prevalent in the NMT systems for Arabic dialects is the system's inability to translate parts of the source sentence and often, the parts of source sentences are translated twice as it was noticed in the translation models for the Arabic Dialects (AD) to the Modern Standard Arabic (MSA). NMT deals with the morphological, semantic, syntactic and different types of linguistic complexities that often pose a difficulty in capturing, without exploiting the linguistic resources such as part-of-speech (POS) tag. POS can be accorded as the process that assigns the part of speech to each word for a sequence of words. For the NMT system, the accuracy of POS tagging is an essential pre-processing step for high-quality translation. Similarly, employing POS information through POS tags as an additional feature, aids in modeling re-ordering between languages and improving the translation quality from AD to MSA. Consequently POS is considered to be advantageous because, it provides large amount of information on words and its neighbors. Knowing whether a word is a noun or a verb provides relevant information on the neighboring words (nouns are preceded by determiners and adjectives, verbs by nouns) as well as on the syntactic structure around the word (nouns are generally part of noun phrases), which makes part-of-speech tagging an essential component of machine translation task for Arabic dialects. Conventional approaches such as Statistical Machine Translation (SMT) requires high computational resources. SMT is not equipped to handle the word ordering problem, one of the major syntactic issues in Arabic dialects. Analysis of the word order involves figuring out the occurrence of subject, object and verb in the sentences. On the basis of the analysis, the different languages could be classed as SVO (English), SOV (Hindi) and VSO (Arabic) and some languages, such as Arabic Dialects allow a free word order. Videlicet, the word order does not convey any information about the subject and the object but possible different information (old & new). These profound differences pose challenges to SMT because as the sentences get longer, it is no longer simple enough to contain a subject, object and a verb but are complex constructions made up of several sentential components.

Likewise, another challenge on developing a neural machine translation model for Arabic dialect to MSA is the absence of the standardized orthographies for all Arabic dialects. The absence includes morphological differences which are evident in the use of clitics and affixes that do not exist in Modern Standard Arabic. Training NMT models usually require a large amount of annotated data, which is often unavailable in the case of low-resource languages such as Arabic dialects. Furthermore, the quality of translation decreases in parallel to the reduction in the amount of training data for low-resource languages. One of the main issues for developing AD-to-MSA neural machine translation system is the rareness of training data. Arabic Dialects is considered among the languages that has limited access and availability of resources. In the event of sparse parallel data, exploiting other knowledge resources such as POS tags can be crucial for the performance of translation from AD to MSA. The plausible techniques that address Dialectal Arabic machine translation challenges are still under study and no previous research work has focused on employing POS tags as additional linguistic features into NMT model for Arabic dialects. The goal of the current research work is to present a Bi-LSTM-CRF segment-level Arabic Dialect POS tagger model that can be integrated into the NMT model to achieve better performance. In order to carry out the integration, a unified Multitask NMT that trains several machine translations and POS tasks were proposed and developed. In the proposed model, the AD encoder is shared between AD-MSA translation task and the segment-level Bi-LSTM-Conditional Random Filed (CRF) POS tagging task. In addition to, only a shared layer and an invariant layer are shared between the translation tasks.

The current study will make use of sequence-to-sequence model for all the tasks. Experiments reveal that the performance on the Levantine Arabic (LA) to Modern Standard Arabic (MSA) machine translation task and the Maghrebi Arabic (MA) to MSA machine translation task measured on the

BLEU score can be improved. A high-test accuracy on the POS tagging task for both Levantine Arabic and Maghrebi Arabic was also obtained. Furthermore, the study analyzed three crucial points while designing the proposed model. First, the study investigated the influence of segment-level POS tagging task on Arabic dialect translation. Along with it, the impact of training the models without the POS tagging task with different pre-trained word embedding and with different dimensions was also examined. Lastly, the influence of training the proposed model with FastText pre-trained word embedding with segment-level POS tagging task was analyzed. The research experiments revealed the effectiveness of the chosen approach and pointed out that the approach was effective in leveraging common information to enhance the performance and quality of Arabic dialects translation, exclusively.

2. Related Work

Studies have revealed that the existing work on machine translation for Arabic dialects is limited. Most of the recent research work has focused on the statistical and rule-based approaches. Bakr [4] introduced a general method to convert sentences from the Egyptian Dialect into vocalized renditions of MSA sentences. To automatically tokenize and tag Arabic sentences, the researchers used the statistical approach. A method specifically based on significant rules was adopted with the aim of creating diacritics for target sentences in Modern Standard Arabic. The work was evaluated on a dataset of 1K of Egyptian dialect sentences (including training and test 800 and 200, respectively). For converting dialect words to MSA words, the system achieved an accuracy of 88%, whereas for producing the words into its correct order, the system performed at an accuracy of 78%.

On the other hand, Meftouh [5] presented PADIC, a multi-dialect Arabic corpus that covers the Modern Stranded Arabic, the Maghrebi dialects (Tunisian and Algerian) and the Levantine dialects (Syrian and Palestinian). In contrary to the recent works in the area, some experiments were conducted on several statistical machine translation systems that ran through a variety of possible pairs of languages (Modern Standard Arabic and dialects). The authors investigated the importance of using the proposed language model on machine translation, by employing smoothing techniques along with including the techniques within a broader framework. The study achieved satisfactory results when translating among the various dialects within Algeria, primarily due to the shared vocabulary. It was remarked that the statistical machine translation performed significantly well when the translation was carried out on the Palestinian and the Syrian dialects. The ease of practicality and effectiveness can be accorded to the linguistic proximity of the two vernaculars, concerning translations into Modern Standard Arabic, remarkable results were obtained with the Palestinian dialects.

The rule-based approach was proposed by Al-Gaphari [6] to transform the Sanaani dialect to Modern Standard Arabic, which is sued in the capital city of Yemen. The system designed generated 77.32% of accuracy on being tested on the Sanaani corpus of 9386 words. Tachicart & Bouzoubaa [7] suggested a rule-based approach that relies on the language models to translate the Moroccan dialect to MSA. The method based on the morphological analysis through the use of the Alkhalil morphological analyzer was proposed. Salloum & Habash [8] presented Elissa, which is a translation system constructed by rules designed and developed for the conversion of Arabic vernaculars to the standard form of Arabic. The described system works with Levantine (Jordanian, Syrian and Palestinian), Egyptian, Iraqi and Gulf Arabic dialects. Sadat [9] presented a model for the translation of the Tunisian Arabic Dialect to the standardized modern form of Arabic. The model is based on a bilingual lexicon which was designed for the particular context of the translation exercise and uses a set of grammatical mapping rules with an additional step for disambiguation. The grammatical mapping rules are based on a language model of Modern Standard Arabic which would aid in choosing the best possible translated target phrase. And, it is a word-based translation system. The model secured a BLEU score [10] of 14.32, in a test set consisting of 50 sentences from the Tunisian dialect.

Most of the methods mentioned above focused on the statistical and rule-based approaches. The Rule-Based Machine translation system have a significant drawback: the construction of such systems

demands a considerable amount of time. To improve the quality of the Rule-Based Machine Translation RBMT, it is necessary to modify the rules, which requires elevated level of linguistic knowledge. Moreover, the statistical approaches need high computational resources and cannot handle one of the syntactic issues in Arabic dialects; which is the word ordering problem. Multitasking is considered to be an effective methodology in improving the performance of an Arabic dialect translation task, with the help of other NLP related tasks such as the POS tagging. The only recent research work that has focused on neural machine translation for Arabic dialect was proposed by Laith H. Baniata et al., [11]. The researchers introduced a multi-task NMT model to translate from AD to MSA. The model is based on multi-task learning (MTL), which shares one decoder among language pairs and each source language has a separate encoder. Experiments show that, given a small parallel training data, the multi-task neural machine translation model is sufficient to generate the correct sequence and produce translations of high quality while learning the predictive structure of multiple targets.

Amongst the several approaches for the translation of Arabic dialect, external information into neural network-based models for the English language is considered as one of the most relevant. Collobert [12] proposed a unified neural network architecture and learning algorithm that can be applied to various natural language processing tasks including part-of-speech tagging, chunking, named entity recognition and semantic role labelling. The basic multi-task architecture of the model is carrying out the sharing of the lower layers to determine the characteristic features. After the shared layers, the remaining layers are split into the multiple specific tasks. Instead of exploiting man-made input features carefully optimized for each work, the model learns internal representations by vast amounts of mostly unlabeled training data. Additionally, the CNN (Convolutional Neural Network) model was used in the work. Pengfei Liu [13] presented a multitask learning to learn across multiple tasks jointly. By using the recurrent neural network, three different configurations of information sharing were applied to the model text, with task-specific and shared layers. The whole network is jointly trained on all the tasks. The experiments on four benchmark text classification tasks revealed that the introduced models can improve the performance of a task with the help of other related tasks. There have been few papers published on translating texts between closely-related languages .One example is the work by Costa-jussa [14] which showed a comparison between rule-based system, phrase-based system and neural machine translation (MT) systems in the Catalan-Spanish language pairs . When using in-domain test set, the performance of neural MT is better than other approaches. The experiments on the out-of-domain test set revealed that the rule-based system (Spanish-to-Catalan, in BLEU) and the phrase-based system (Catalan-to-Spanish) achieved better performance Hassani [15] proposed and implemented an Intralingual Machine Translation for translating texts in Kurmani to Sorani. The author used word-for-word translation (literal or direct translation) among the dialects. The results were tested using human raters method. According to different human raters, the experiments showed that this naive approach could provide a significantly intelligible results. The experiments also showed that this approach is might be able to be considered as an immediate remedy for the current lack of corpus issue. Costa-jussa et al. [16] introduced the first NMT system trained to translate between national language varieties. The researchers used the language variety pair Brazilian and European Portuguese as an example and parallel corpus of subtitles to train the NMT system. Compared to a Statistical Machine Translation (SMT) system trained on the same data, they reported a performance of 0.9 BLEU points in translating from European Portuguese to Brazilian Portuguese and 0.2 BLEU points when translating in the opposite direction. The results indicate that the NMT system produces better translation than the SMT systems not only in term of BLEU scores but also according to the judgment of seven native speakers. In the neural machine translation, application of additional word factors like POS tags has shown to be beneficial, as shown in Reference [17]. Jan Niehues & Eunah Cho [18] explained that multitask learning approach is successful and introduced additional knowledge into an end-to-end neural attention model. By jointly training various natural language processing (NLP) tasks in one system, the model can leverage common information and improve the performance of the individual task. The experiments are conducted for German to English translation task. As additional linguistic resources, POS information and named-

entities (NE) were exploited. Tests show that up to 1.5 BLEU points can improve the translation quality under the low-resource condition.

3. Neural Machine Translation

Neural machine translation (NMT) has lately been introduced as a promising methodology with the potential of handling the shortcoming of traditional statistical machine translation systems. The power of NMT models lies in its ability to learn directly, end-to-end fashion, the mapping from input text to associated output text. Neural models are not new in the field as Neco and Forcada [19] proposed such a solution many years ago. Other models by Schwenk et al. [20] and Kalchbrenner and Blunsom [21] have also been proposed, but for the first time Cho [22] and Sutskever [23] were able to design a powerful architecture for machine translation.

In the encoder-decoder architecture which was discussed by Peyman [24], two recurrent neural networks (RNNs) are trained together to maximize the conditional probability of a target sequence (candidate translation) $y = y_1, \ldots. y_m$, given a source sentence $x = x_1, \ldots. x_n$. Input words are sequentially processed consecutively until the end of the input string is reached. An encoder reads words and maps the input sequence into a representation with a fixed-length. At each time in step t, an input word is taken and the hidden state is further updated. This process can be expressed as in Equation (1):

$$h_t = f(E_x[x_t], h_{t-1}) \quad (1)$$

where $h_t \epsilon R^d$ the hidden state (a vector) is at the time step t and $f(.)$ is a recurrent function such as long short-term memory (LSTM) [25] or gated recurrent unit (GRU). $f(.)$ is responsible for updating the hidden state of the layer and other associated units (if there are any, such as memoryunits, etc.) $E_x \in R^{|V_x| \times d}$ is an embedding matrix for source symbols (d is the embedding size). The embedding matrix is a Look-up table (LUT) whose cells are treated as network parameters and updated during training. The embedding (numerical vector) for the v th word in v_x (vocabulary) resides in the vth row of the table. In the next step, the model undertakes processing for all words in the source sequence; h_n is a summary of the input sequence which is referred to as the context vector (c). Another RNN is initialized by c and seeks to produce a target translation. There is one word sampled from a target vocabulary v_y at each step of the process. The decoder conditions the probability of picking a target word $y_t$on the context vector, the last predicted target symbol and the decoder's state. This can be expressed in Equations (2):

$$y_t = g(E_y[y_{t-1}], S_{t,}c) \quad (2)$$

$S_t = f(E_y[y_{t-1}], S_{t-1}, c)$ where S_t is the decoder's hidden state. Since we compute the probability of selecting y_t as the target word, $g(.)$ should give a value in the range $[0,1]$. The most common function for $g(.)$ is Softmax. The encoder and decoder RNNs are trained together to maximize the log probability of generating a target translation and are given an input sequence x, so the training standards can be defined as in Equation (3):

$$\max_{\theta} \frac{1}{K} \sum_{k=1}^{k} \log(y_k | x_k) \quad (3)$$

where θ is a collection of network parameters and k designates the size of the training set. As mentioned before, recurrent functions in encoder-decoder models are not usual mathematical functions. RNNs are not powerful enough to capture all features about sequences, so more powerful choices, such as LSTM RNNs are needed.

4. Segment Level Bi-LSTM—CRF Model for Arabic Dialect POS Tagging

A conditional random field (CRF) is an undirected graphical model that was successfully experimented in various sequence labeling tasks such as POS tagging, word segmentation and named entity recognition (NER). The CRF model can prevent the limited feature selection in Hidden Markov models and MEM by considering the correlation between labels in neighborhoods. Moreover, it can

acquire a global optimum via a process of global feature normalization. For a given sequence $s = \{s_1, s_2, \ldots, s_n\}$, Where s_i is the vector of the ith Arabic dialect segment and let $L = \{l_1, l_2, .., l_n\}$ be a sequence of labels for S, where l_i is the label of the ith Arabic dialect segment. The following equation can express the linear-chain CRF model:

$$p\,(L|S;W,b) = \frac{\prod_{i=1}^{n} \varphi_i(l_{i-1}, l_i, S)}{\sum_{Q\,\in\varphi(S)} \prod_{i=1}^{n} \varphi_i(Q_{i-1}, Q_i, S)} \tag{4}$$

where $\varphi_i(Q_{i-1}, Q_i, S) = \exp\left(W_{Ql}^{T} S_i + b_{Q,l}\right)$ is the potential function corresponding to a label pair $r(Q, l)$, W_T is the weight vector, b is bias and $\varphi\,(S)$ denotes the set of possible label sequence for S. The current study had developed a segment-level Bi-LSTM-CRF model to do POS tagging task for Arabic dialect as shown in Figure 1. First, the Arabic Dialect segments with its labels (tags) were retrieved and a vocabulary dictionary for the segments and its labels were created. Then, a reverse vocabulary dictionary was also created to map the segments and its tags to a sequence of numbers and then, the sequences were padded. The study used an equal-length input, so the sentences were padded to a length of 25 for both dialects (Levantine and Maghrebi). The model consists of a forward LSTM, backward LSTM and CRF layer. The key point is to use forward LSTM to capture the previous input feature, backward LSTM to capture future input feature and CRF layer to obtain sentence level tag information. Let's assume $X = \{x_1, x_2, .. x_n\}$ represents a general input sequence, $y = \{y_1, y_2, \ldots y_n\}$ represents the tag sequence for X and P_{n*k} denotes to the probability matrix, where k is the number of tag types. The optimal tag sequence can be obtained by maximizing the target function.

$$\mathrm{s(X, y)} = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=0}^{n} P_{i, y_i} \tag{5}$$

where $P_{i,j}$ is the probability that the ith segment is tagged as the jth tag and A is the state-transition matrix, where $A_{i,j}$ is the probability of transferring from the ith tag to jth tag.

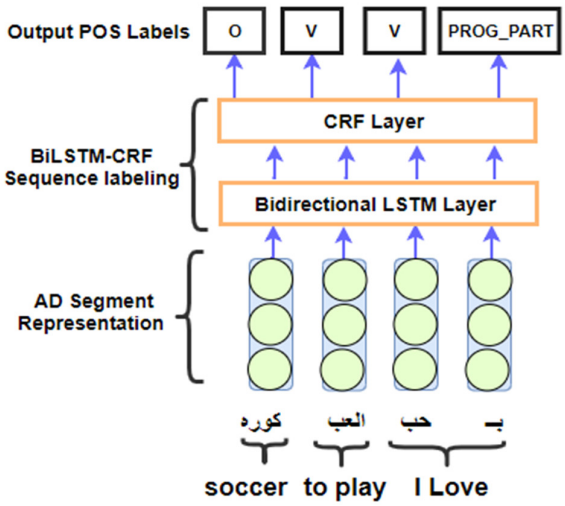

Figure 1. Arabic Dialect Segment-Level Bi-LSTM-CRF part-pf-speech tagging model.

5. Methodology

In natural language processing pipeline, information extraction, parsing and neural machine translation employs the POS tags as an additional linguistic feature. Recent research on the topic of deep learning methodologies such as CNN or RNN models has accorded to the possible application of similar techniques through different NLP tasks. That is, instead of using the output of a model as an input for another model, a single one unified model to do translation tasks from Arabic Dialect (AD) to Modern Standard Arabic (MSA) and from MSA to English (ENG) as well as performing

Dialectal Arabic POS tagging task can be developed. The Multitask Learning (MTL) model automatically learns and shares relevant and necessary information across the tasks. MTL as an approach induces transfers that improve the generalization by using the domain information contained in the training signals of related tasks as an inductive bias. MTL utilizes the correlation and the shared representation between related translation tasks, such as AD-MSA and MSA-ENG and exploiting the Arabic dialect linguistic features from the POS tagging task to improve translation quality by learning the tasks in parallel. The information learnt for each task has the possibilities of aiding other tasks to effectuate effective learning. The primary purpose of the inductive transfer is to leverage additional sources of information to improve the performance of learning on the current task. Inductive transfer can be used to enhance generalization accuracy, the speed of learning and the intelligibility of learned models. A learner that learns many related tasks at the same time can use these tasks as inductive bias for each other and thus better learn the domain's regularities. This can make learning more precise and allow translation tasks with a small amount of training data to be learned better. The architecture of the designed model is a Recurrent Neural Network (RNN) based-encoder decoder architecture with two translation tasks and one part of speech (POS) tagging task. The central point of the proposed multitask NMT model for Arabic dialects is, the sharing scheme in latent feature space. In deep learning models, the latent features can be considered as the hidden states of the Bi-LSTM at the end of the sentence. Here, the study has incorporated a shared-private scheme with multitasking NMT model. All the translations tasks use a separate translation decoder across all the various language pairs such as DA-MSA and MSA-ENG and a separate encoder encodes each source language due to the peculiarities the language. All the translation tasks share a private layer and invariant layer and the Arabic Dialect segment level POS tagging task shares the same encoder with AD machine translation task. Sharing more information across the tasks is preferred and the model details are described in the following section.

6. Proposed Model

The proposed multitask NMT model relies on recent machine translation architectures for multiple languages [26–28]. The overall structure of multitask NMT model consist of four parts as shown in Figure 2, shared encoder E, decoder D, shared-private layer PL, shared-language invariant layer IL and conditional random field layer (CRF). The model is designed to carry out the translation tasks and POS tagging task in parallel. The first task is the translation from Arabic Dialect (AD) to the Modern Standard Arabic (MSA), the second task is the translation from the Modern Standard Arabic (MSA) to English (ENG) and the POS tagging task for Arabic Dialect. Therefore, all parts (two encoders, one shared private layer, one shared language invariant layer, two decoders and a conditional random field layer (CRF)) are applicable to all the tasks in general. Hence, there are seven components E1_AD, E1_Segment, E2_MSA, PL_AD, PL_MSA, IL_AD, IL_MSA, D1_MSA, D2_ENG and CRF_POS. The key factor relevant for designing the Multitask learning model is the level of sharing across of the tasks. The Multitask NMT model is able is to learn the translations problem together with other related issues, such as POS tagging at the same time by using a shared representation. In the case of task commonality and limited training data, model shows a better performance than a model singularly trained on a single dataset. Moreover, the model can capture more morphological, semantic, syntactic and lexical features of Arabic dialects. Part of speech tagging (POS) is one of the main building blocks in many natural language processing applications. The POS of Arabic dialects is challenging due to the lack of spelling standards and the pervasiveness of transformative morphological operations such as letters substitution or deletion and word merging and lexical borrowing from foreign languages. Research work on Arabic dialects POS tagging has focused on developing resources and tools for each dialect separately [29,30], because each dialect has different affixes, different lexical and word ordering choices and many foreign languages influence each dialect.

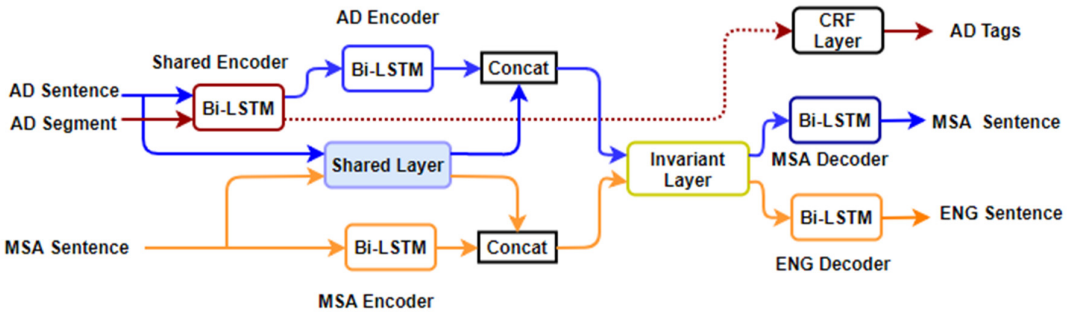

Figure 2. The Architecture of the proposed model.

Word segmentation and POS tagging are core steps such as the neural machine translation, for high-level natural language processing tasks. As a reason for exploiting external linguistic features in proposed Multitask NMT model, the current study presents a segment-level part of speech tagger for Arabic Dialect using Bi-LSTM–CRF architecture. The POS dialectal Arabic dataset described by Kareem Darwish [31] was used. Segmentation and POS tagging, as proposed by Eldesouki [32], to overcome the need for the standardization of different dialectal writings, were applied to the original raw text without any modification and correction. For example, the word بحب اسمع (baheb asma) "I love to listen" is segmented as اسمع + حب + ب (ba + heb + asma) and tagged as PROG PART+V+V. Words are segmented into tokens (clitics) and POS is given for each token. In the dataset, tokens, words and sentences are separated by token boundary tag (TB), word boundary tag (WB) and end of sentence tag (EOS) respectively. Boundaries are potentially useful, as it can be used to learn the symbols to be attended to and when to forget the information in the recurrent layer. Using the Arabic Dialects segments as an input feature to the shared encoder between the AD-MSA and POS tasks, it guarantees the process of information sharing between the word forms and the segment form. Neural models can learn that inflectional variants are semantically related and can be represented as a similar point in the vector space. The researchers believe that the segment representation increases the data efficiency; low-frequency word or variants might be unknown to the word level model. By sharing the encoder between the segment-level POS tagging task and AD-MSA translation task, the model will learn the similarity between Arabic Dialect words that share the same segments and capture more syntactic and semantic features. Consider the following two Arabic words هيلعب بالكوره "he will play soccer" that has the segments: ه + كور + ال + ب + يلعب + هـ and the words بحب الكوره "I love soccer", these words share four segments ه , كور , ال , ب . Sharing more information between the segment-level AD POS tagging task and the AD-MSA translation task is beneficial, practical and improve the translation quality and generate the correct target sequence. More details pertaining to the process and the application are described in the following sections of the paper, along with a discussion on the training schedule for each task.

6.1. Arabic Dialect Encoding with Bi-LSTM

Long-Short Term Memory network is a type of recurrent neural network RNN that specifically address the issue of learning long-term dependencies by boosting the RNN with memory vector $m_t \in R^d$. In the proposed Multitask NMT architecture, we used Bi-LSTM as an encoder for the translation tasks and part of speech POS tagging task. An LSTM unit takes x_t, h_{t-1} and m_{t-1} as its input and produce h_t and m_t by computing the following Equation (6):

$$
\begin{aligned}
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i), \\
f_t &= \sigma\big(W_f x_t + U_f h_{t-1} + b_f\big). \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o), \\
g_t &= \tanh\big(W_g x_t + U_g h_{t-1} + b_g\big), \\
m_t &= f_t \odot m_{t-1} + i_t \odot g_t, \\
h_t &= o_t \odot \tanh(m_t)
\end{aligned}
\tag{6}
$$

where i_t, f_t and o_t designate the input, forget and output gates respectively. These gates collectively determine how to update the current memory cell m_t and the current hidden state h_t. The parameter d is used to indicate the memory dimension in the LSTM were all the vectors in the defined architecture has the same dimension. $\sigma(.)$ is an element-wise sigmoid function with an output range between $[0,1]$. Subsequently, $tanh$ indicates the hyperbolic tangent function that has an output range between $[-1,1]$ and $\odot$ denotes the element-wise multiplication function and W_p, U_p and b_p, $p \in \{i, f, o, g\}$ are considered as network parameters. The function f_t is set to have a better understanding of mechanisms involved in the architecture and to control distinct type of information that is needed to be discarded from old memory cell. In addition, the use of i_t to control in the amount of information that is stored in the current memory cell and o_t to control the parameters required to be provided as an output based on the memory cell m_t. LSTM was designed to deal with vanishing and exploding gradient problem when training traditional recurrent neural network. LSTM structure is illustrated in Figure 3.

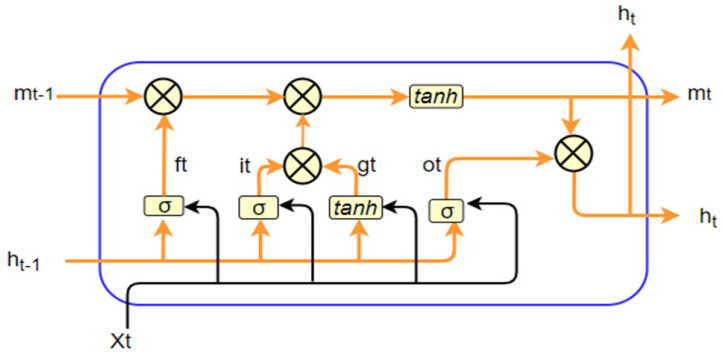

Figure 3. Long short-term memory (LSTM) Structure.

For a given sequence $(x_1, x_2, \ldots x_n)$ contains n words, each is represented as d dimensional vector, LSTM computes a representation $\overrightarrow{h_t}$ of the left context of the sentence at every word t. Gathering a representation of the right context $\overleftarrow{h_t}$ will add useful information. This can be done using second LSTM that reads the same sentence in reverse. We will refer to the former as the forward LSTM and the latter as the backward LSTM. These are two distinct networks with different parameters. This forward and backward LSTM pair is referred to as a bidirectional LSTM (Bi-LSTM). The representation of a word using this model is obtained by concatenating its left and right context representations $\left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$. These representations effectively include a representation of a word in context, which is useful for numerous tagging applications.

6.2. Shared-Private Scheme

The central idea of multi-task learning is the sharing scheme in latent feature space. Latent features can be regarded as the hidden states of Bi-LSTM encoders in the multi-task NMT model. So that, sharing schemes are different in how to group the shared features. Here, a shared-private scheme is used. The shared-private scheme has two feature spaces for each translation task: Bi-LSTM encoder is used to capture task-dependent features and a private Bi-LSTM layer is used to capture

the task-invariant features. As shown in Figure 4, each translation task for each Arabic dialect is assigned a private Bi-LSTM layer (LA encoder, MA encoder) and a shared Bi-LSTM layer (shared layer). Formally, for any Arabic dialect sentence in task k, its shared representation s_t^k can be captured along with task-specific representation h_t^k:

$$s_t^k = BiLSTM(x_t, s_{t-1}^k, \theta_s) \quad (7)$$

$$h_t^k = BiLSTM(x_t, h_{t-1}^m, \theta_k) \quad (8)$$

Consequently, the features from the private space and the shared space are concatenated. By exploiting the shared-private scheme, multitask NMT model can learn task-specific and task-invariant features non-redundantly, therefore capturing the shared-private separation on different translation tasks. Moreover, the model will improve the representation of the Arabic dialect sentence, capture more features and produce better translation quality for the Arabic dialects.

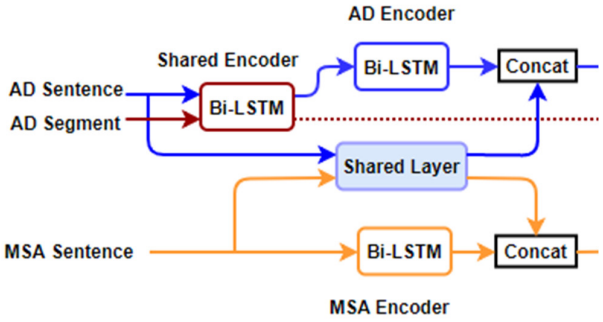

Figure 4. The Layout of shared-private scheme.

6.3. NMT Decoding for Arabic Dialects Sentence

Neural machine translation NMT maps all words of source sentence for Arabic dialect into a continuous space through an embedding layer. The word embeddings are processed by a bidirectional layer implemented with LSTM that will encode each source word together with its right and left context in a sequence of hidden states. Once the hidden states of the Bi-LSTM encoder are computed, the decoder starts generating the target string one word at a time. The decoder is implemented as a Bi-LSTM layer. As shown in Figure 2, a shared language-invariant layer and Bi-LSTM decoders are used for all the translation tasks. The model has two decoders: LA decoder and MA decoder that capture more information from Modern Standard Arabic (MSA) that benefits to the internal representation of Levantine Arabic (LA) and Maghrebi Arabic dialects. Each decoder is followed by a dense layer and output layer. Each output layer is composed of SoftMax layer. Figure 2 describes the shared layers depending on the architecture of the model. The hard parameter sharing was applied in the proposed multitask NMT model by sharing the shared-layer and the language-invariant layer across the related tasks.

6.4. Optimization

The Nadam optimization approach, which is a mini-batch stochastic gradient descent approach with the Nestrove momentum, was used for optimization. Mini batches of size 140 were learned with fixed sizes within a language pair (MSA-ENG) for some iterations and was continued on to the next language pair (AD-MSA). Also, the Nadam optimization algorithm is used in the Arabic dialect POS tagging task, as the weight updates are set using the Nadam optimization algorithm. The current paper adopts the same parameter settings and model structure for the translations tasks and different model architecture for the POS tagging task. The model has different weight for the individual tasks due to a default training schedule. First, the model is trained on the POS task for some iterations and once the POS task training finishes, the training is performed on MSA-ENG translation task and then on the AD-MSA translation task. Successively, the process of training is continued alternately.

7. Experiments and Results

We performed several experiments to evaluate the proposed multitask NMT model on different translation tasks and POS tagging task. The proposed model is reviewed on the translation task from MSA to AD and from MSA to English. The experiments were conducted on two different types of Arabic dialects: Levantine and Maghrebi, Levantine is a spoken variety of widely used in Jordan, Syria, Palestine and Lebanon. Maghrebi Arabic is a spoken dialect used in Morocco, Algeria and Tunisia. The Multi-task Learning Based-Neural Machine translation Models will be used to handle the low-resource languages. Moreover, 10,000 sentence pairs was used for MSA-ENG translation task and 20,000 sentence pairs was used for the ENG-GER translation task.

7.1. Data

For the translation tasks (see Supplementary Materials), the study used the same parallel corpus used by Laith Baniata et al. [11]. We had concatenated the Maghrebi Dialects (Moroccan Dialect, Algerian Dialect and Tunisian Dialect) together from the PADIC Corpus and MPCA Corpus [33] as well the Levantine Dialects (Jordanian Dialect, Syrian Dialect and Palestinian Dialect) are concatenated together from the same corpora. The Multitask NMT model is trained on 13,805 sentence pairs for Levantine Dialect (LD) and 17,736 sentence pairs for Maghrebi Dialect (MD). The text data was collected from social media, movies, TV shows. For the test set, the Multitask NMT model on 2000 sentence pairs for Levantine Dialect and 2000 sentence pair for Maghrebi Dialect were tested. For the POS tagging task, the dataset described by Kareem Darwish et al. [31]. were adopted. The adopted dataset is in-of-domain data for machine translation task and the test data used is also from the mentioned dataset. The POS dataset includes a set of 350 tweets for four major Arabic dialects (Egyptian, Levantine, Gulf and Maghrebi).

Three cross-validation methods, that is, independent dataset test, sub-sampling (or K-fold cross-validation) test and jackknife test, are often used to evaluate the anticipated success rate of a predictor. Among the three methods, however, the jackknife test is deemed the least arbitrary and most objective [34] and hence has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors [35–39]. However, this procedure is time- and source-consuming. Therefore, in this paper, we used K-fold cross-validation, where K set to 1 such that a single train/test split is created to evaluate our proposed model.

In the Bi-LSTM–CRF POS tagging model, the study used the Levantine Dialect that includes 7221 words and Maghrebi Dialect that consists of 6400 words. The dataset has unstructured information that might affect the performance of the model. Therefore, a pre-processing stage has been employed for Arabic dialect and Modern Standard Arabic sentences. Punctuations, diacritics, hashtags and non-Arabic characters were removed for both AD and MSA. Also, an orthographic normalization was performed. For example, the transformation of all إأآا characters to the ا character was applied. Stemming or stop word removal has not been employed. The sequence length is set to 55 for the translation tasks and 25 to POS tagging task. We tokenize both MSA and AD using Python tokenizer. The modern standard Arabic contains a variety of tokens than English and Arabic Dialects and its sentences are shorter than English and Arabic Dialects. Modern Standard Arabic has more probability in its infrequent words than English and the words in the long tail tend to be morphologically complex, alternately beginning with affixes like “AL” “ال” (the) or “wa“ "و" (and).

7.2. Training

The framework developed by Python and Keras was used to train the proposed model. The study employed 22,459-word vocabularies for Maghrebi Arabic (MA) and 19,327-word vocabularies for Levantine Arabic on AD-MSA translation task, 10,185-word vocabularies for Modern Slandered Arabic is used on MSA-ENG translation task. Moreover, 2606-Token and 23-tags for Levantine Arabic (LA) is used for on the POS tagging task. 2423 Token and 23 Tags for Maghrebi Arabic (MA) is used on POS tagging task. For translation tasks, the Nadam optimization algorithm is used with $\beta 1 = 0.9$ and $\beta 2 = 0.999$ and a learning rate value set to 0.002. We set a recurrent dropout value to 0.20 and 0.23

to AD and MSA Bi-LSTM encoders. A dropout layer with dropout value 0.40 is added after the embedding layer for each translation task. For POS tagging task on Arabic Dialect, the Nadam optimizer is also used with $\beta 1 = 0.9$ and $\beta 2 = 0.999$ and learning rate value 0.002. We passed the embedding of the POS tagging task to SpatialDropout1D layer with dropout value 0.1. Although the model is initially incorporating (6 M trainable parameters for MSA-ENG task, 11.9 M trainable parameters for LA-MSA task, 1.3 M trainable parameters for POS_LEV task, 13 M trainable parameters for MA-MSA and 1.2 M trainable parameters for POS_MA: word embedding and hidden size are shown in Table 1, this model size proved to require 1190 s per epoch for LA-MSA task, 2153 seconds per epoch for MA-MSA task, 36 seconds per epoch for MSA-ENG Task and 18 seconds for POS_LEV task and 15 s POS_MA task . We had trained the model alternatively on the MSA-ENG and the AD-MSA task. The training data will be randomly shuffled at each epoch for all tasks. The experiment made use of a batch size of 140 for each translation task and a batch size of 50 for the POS tagging task. The model is trained by minimizing the cross-entropy loss for each translation task and by maximizing the sparse categorical accuracy for POS tagging task.

Table 1. Results of Multi-Task neural machine translation (NMT) without part-of-speech (POS) in terms of BLEU score.

Pairs	Embedding Type	Embedding Size	Hidden Size	Epochs	BLEU
LA-MSA	Random	150	150	102	0.36
MA-MSA	Random	150	150	102	0.26
MSA-ENG	Random	150	150	49	0.29
ENG-GER	Random	150	150	45	0.34
LA-MSA	FastText	300	150	52	0.33
MA-MSA	FastText	300	150	52	0.32
LA-MSA	Polyglot	64	150	64	0.26
MA-MSA	Polyglot	64	150	64	0.25

7.3. Results

7.3.1. Automatic Metric

On conducting the experiment, it can be asserted that the researchers have carried out extensive experiments on the Multitask NMT model by utilizing POS and by not exploiting the POS linguistic resources. The model was also experimented using different pre-trained word embedding like polyglot and FastText [40,41]. The translation performances measured will be listed in the BLEU score automatic evaluation metric [10]. The results in Table 1 show the efficiency of Multitask NMT model without adding POS tagging using different pre-trained word embedding. It can be noticed from Table 1 that; the proposed model has obtained excellent results with the FastText pre-trained embedding for both LA-MSA and MA-MSA translations tasks. The results suggest success and satisfaction because the FastText pre-trained word embedding was trained on the Egyptian dialects.

Moreover, the FastText vectors in dimension 300 were obtained using the Skip-Gram model. Therefore, it can be summarized that the models that were trained using the FastText pre-trained word embedding outperforms the models that were trained using polyglot. It also outperforms the model that was trained using random embedding. It is clear from the Table 2 that in all cases, the proposed Multitask NMT model with POS tagging outperforms the Multitask NMT model (that is trained without POS tagging). More importantly, results in Table 2 shows that the performance of Multitask NMT model that exploits the POS tagging for the LA-MSA, MA-MSA and ENG-MSA task is higher when compared to the Multitask learning- based Arabic dialect NMT model, which was proposed by Laith Baniata et al. [11] as shown in Table 3. The result produces a definite understanding because the closeness of Levantine Arabic and Modern standard Arabic and both dialects share many vocabularies. The central reason behind the development of the result was the utilization of the segment- level POS tagging RNN-CRF model and the utilization of the shared-private scheme for both the dialects. Exploiting the POS tags for both dialects (Levantine and

Maghrebi) as an additional feature in Multi-task NMT model was beneficial in word ordering between languages and produced better translation quality. By utilizing the shared private scheme, the model was able to learn better representation and capture more features from source languages and handle the syntactic issue in Arabic dialect which is the word ordering problem. Also, it was noticed that Maghrebi Arabic which is a mixture of many different languages (Berber, Latin, (African Romance) old Arabic, Turkish, Spanish, Italian and Niger-Congo languages) and new word from English and French; the Multitask NMT model improved the performance of translation on the training dataset of Maghrebi Arabic-MSA. Moreover, the outstanding results on BLEU score were an inherent result of the inductive transfer, which can be utilized to leverage the additional source of information to improve generalization accuracy, the speed of learning and the intelligibility of learned models. Furthermore, the model has improved the performance of learning on the MSA-English and English-German translation tasks.

Table 2. Results of Multi-Task NMT with POS in terms of BLEU score.

Model	Pairs	Embedding Type	Epochs	BLEU	Accuracy
NMT + POS_LEV	LA-MSA	FastText	90	**0.43**	
NMT + POS_LEV	MSA-ENG	FastText	50	0.30	
NMT + POS_LEV	POS_LEV	Random	40		98%
NMT + POS_MA	MA-MSA	FastText	50	**0.34**	
NMT + POS_MA	MSA-ENG	FastText	30	0.29	
NMT + POS_MA	POS_MA	Random	20		99%

Table 3. Results of Multi-Task NMT with shared decoder in terms of BLEU score [11].

Model	Pairs	Embedding Type	Embedding Size	Epochs	BLEU
Single NMT	LA-MSA	Random	150	120	0.17
Multitask	LA-MSA	Random	150	170	0.41
Single NMT	MA-MSA	Random	180	120	0.16
Multitask	MA-MSA	Random	180	230	0.30
Single NMT	MSA-ENG	Random	160	120	0.10
Multitask	MSA-ENG	Random	150	170	0.27

7.3.2. Human Evaluation

We present human evaluation experiments to validate the results obtained with automatic evaluation metric. We used a pilot rating experiments that was used by Costa-jussa et al. [16] in which participants were asked to rate translations using a 1 to 7 Likert scale. We evaluated the quality of the LA-MSA translation direction asking a group of 7 speakers of MSA and they are familiar with the Levantine Arabic (LA) to rate sentences produced by the Multi-Task NMT with POS system and Multi-Task NMT without POS system. We presented to the native speakers a source segment in LA and two translations in MSA. We randomly selected 40 segments and create two sub-sets of 20 segments each. We provide each annotator with a sub-set and ask them to rate the translations taking both fluency and adequacy into account using a Likert scale of 1 to 7. The average results obtained by each system in the pilot rating experiment are presented in Tables 4 and 5.

Table 4. Human evaluation scores—Pilot rating experiment for LA-MSA.

Outcome	Average Score
Multitask NMT	1.4
Multitask NMT + POS	5.9

Table 5. Human evaluation scores—Pilot rating experiment for MA-MSA.

Outcome	Average Score
Multitask NMT	1.3
Multitask NMT + POS	4.4

The average results suggest that humans have a rather positive opinion about the translations produced by both systems. The average score for LA-MSA translation task obtained by the Multitask NMT+POS system was 5.9 whereas the Multitask NMT without POS obtained 1.4. Furthermore, the average score for MA-MSA translation task obtained by the Multitask NMT + POS system was 4 whereas the Multitask NMT without POS obtained 1.3. The outcomes of the pilot rating experiments confirm that the Multitask NMT+POS system produces a higher quality output than Multitask NMT without POS system for both tasks.

8. Model Analysis and Discussion

Intending to clarify the influence of other related tasks such as POS tagging and MSA-ENG tasks on AD-MSA translation task efficiency, the translation examples will be portrayed in Table 5. For the examples, the study has utilized the proposed Multitask NMT model, which has been trained on all the three tasks with a shared encoder between the Arabic Dialect POS tagging task and AD-MSA translation task. A common problem of many Arabic Dialect Neural machine translation systems is that they do not translate parts of the source sentence, or that parts of the source sentence are translated twice. Furthermore, the absence of the standardization in Arabic dialects which are evident in the use of clitics and affixes cannot be captured correctly without employing the POS tagging task on the segment level in the current proposed model. The baseline model (Multitask NMT Model without POS) is induced to the issue of lacking, as shown in the first two examples. The translation quality of the proposed Multitask NMT model with POS tags linguistic features has improved significantly on being compared to the baseline model (Multitask NMT Model without POS tags) in several aspects. In Table 6, the baseline model is not translating the Levantine Arabic Dialect word نحكي nehky 'to speak' into الكلام Alkalam 'talk' and not adding the preposition letter من min 'of' into the target sentence but is translating a different word اجل ajal 'yes' which is not related to the reference sentence. The Multitask NMT model with POS tagging translated the whole sentence correctly with same meaning even the preposition letter is not the same one in the reference but it is another preposition letter used in the Arabic language with a different way. In Table 7, the baseline model could not translate Maghrebi Arabic sentence correctly. The word ذهبت thahabto 'I went' is repeated twice and the two words على ود ala wed 'in order to' were translated into عند enda 'at', which is incorrect, while the proposed Multitask NMT model with POS tags translated 90% of the Maghrebi Arabic sentence correctly. The influence of segment level Arabic dialect POS tagging task on the translation quality is evident. The model can handle the free word order problem and generate a correct order and context of the target language sentence, as shown in Tables 6 and 7.

Likewise, the Multitask NMT model with POS tags can generate the correct sequences. Furthermore, the Multitask NMT model with POS tags was able to obtain a remarkable BLEU score for both Levantine and Maghrebi Arabic Dialect, when compared with other models NMT models as illustrated in Figures 5 and 6. It can be noticed that the representation of the source language that was learned from the multitask NMT model with POS tags is beneficial and improved the translation performance of all language pairs. The proposed model proved to achieve perfect translation quality on different language pairs as shown in Tables 1 and 2 for the MSA-ENG and ENG-GER tasks. Statistical Machine Translation (STM) systems were tested on the same dataset that were used in earlier experiments Meftouh et al. [5] consisting 2000 sentence pairs for Maghrebi Arabic and 2000 sentence pairs for Levantine Arabic (Syrian Dialect). In comparison to a Statistical Machine Translation (SMT) system trained on the same data, the proposed Multitask NMT model with POS tags reported a performance of 0.43 BLEU points in translating from Levantine Arabic to MSA and 0.34 BLEU points when translating from Maghrebi Arabic to MSA. The results indicate that the NMT system produces better translation than the SMT systems not only in term of BLEU scores but also

according to the human evaluation experiments. In general, the proposed model can generate a correct sentence of the target language and convey information about the verb, subject and object for a free word order language like Arabic dialects.

Table 6. Translation examples for Levantine.

Levantine Arabic	حتى نقدر نحكي
Reference—MSA	حتى نتمكن من الكلام
Multi-Task without POS	حتى نتمكن اجل
Multi-Task + POS	حتى نتمكن في الكلام
English Translation	So we can speak

Table 7. Translation examples for Maghrebi.

Maghrebi Arabic	مشيت على ود التحاليل
Reference—MSA	ذهبت من اجل التحاليل
Multi-Task without POS	ذهبت عند ذهبت
Multi-Task + POS	ذهبت على شديد التحاليل
English Translation	I went for analysis

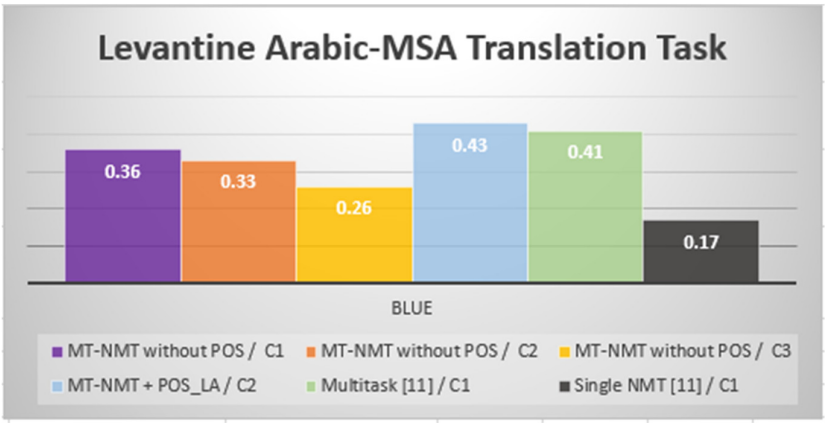

Figure 5. Levantine Arabic-MSA BLEU score with different models, where C1 is a randomly initialized embeddings, C2 a pre-trained/FastText and C3 a pre-trained/polyglot.

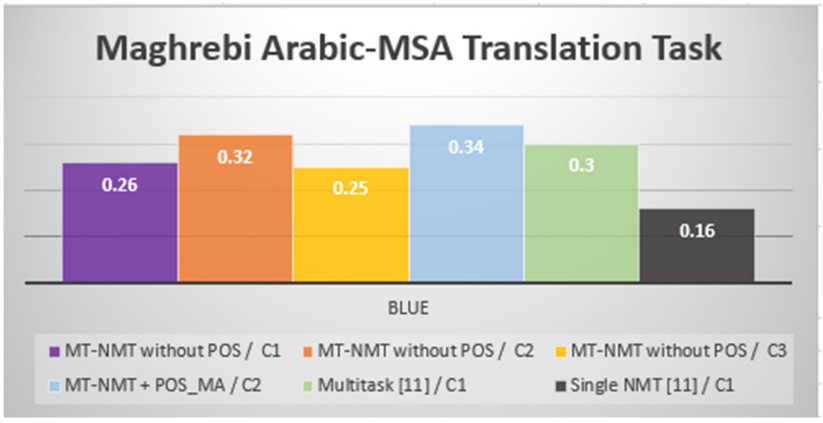

Figure 6. Maghrebi Arabic-MSA BLEU score with different models, where C1 is a randomly initialized embeddings, C2 a pre-trained/FastText and C3 a pre-trained/polyglot.

9. Conclusions

In the current research work, we proposed a Multitask Learning NMT model based on a recurrent neural network encoder-decoder model introduced in recent time. By training the proposed model, not only on the machine translation task but also on other NLP tasks such as POS tagging, the model achieved a definite improvement of the translation performance. The results of the study suggest that the proposed Multitask NMT with POS linguistic features improved the translation BLEU score for the LA-MSA, MA-MSA and MSA-ENG tasks. Moreover, we were able to obtain high accuracy on the test set for the POS tagging tasks, on both the Levantine Arabic and Maghrebi Arabic. The POS linguistic features proved to be a promising approach which is crucial for low-resource languages such as Arabic dialects. The additional linguistic resources such as the segment-level POS for Arabic dialects were beneficial for the translation quality form Arabic dialect to the Modern Standard Arabic. Additionally, the study gains a probable stand because the effective results were achieved through the application of the POS tagging corpus within the domain.

Furthermore, the POS corpus was smaller than the available parallel data. Effective performance on tasks was achieved with a model sharing in the shared layer and invariant layer for the translation tasks and encoder sharing between the AD translation task and POS task. In the current research work, the performance of machine translation tasks for Arabic dialects to MSA was significantly improved by adapting Multi-Task learning approach with segment-level POS tagging for the Arabic dialects. The current proposed Multi-Task NMT model can handle the issue of the rareness of the training data for Arabic dialects. Furthermore, the proposed model proved to address the syntactic issue in Arabic dialect - which is the word ordering problem. The proposed Multitask NMT model with POS tagging is practical and efficient for both low-resource and rich resource languages under the Multi-Task learning framework.

Supplementary Materials: The datasets generated during the current study are available in [AD_NMT] repository (https://github.com/laith85/AD_NMT).

Author Contributions: L.H.B. conceived and designed the methodology and experiments; L.H.B. performed the experiments; S.P. and S.-B.P. analyzed the data; L.H.B. wrote the paper; S.P. and S.-B.P. reviewed and edited the paper.

Funding: This research received no external funding.

Acknowledgments: This study was supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hung, J.-W.; Lin, J.-S.; Wu, P.-J. Employing Robust Principal Component Analysis for Noise-Robust Speech Feature Extraction in Automatic Speech Recognition with the Structure of a Deep Neural Network. *Appl. Syst. Innov.* **2018**, *1*, 28.
2. Pal Chowdhury, A.; Kulkarni, P.; Nazm Bojnordi, M. MB-CNN: Memristive Binary Convolutional Neural Networks for Embedded Mobile Devices. *J. Low Power Electron. Appl.* **2018**, *8*, 38.
3. Salerno, V.; Rabbeni, G. An extreme learning machine approach to effective energy disaggregation. *Electronics* **2018**, *7*, 235.
4. Abo Bakr, H.; Shaalan, K.; Ziedan, I. A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In Proceedings of the 6th International Conference on Informatics and Systems, Cairo, Egypt, 27–29 March 2008; Cairo University: Cairo, Egypt, 2008.
5. Meftouh, K.; Harrat, S.; Jamoussi, S.; Abbas, M.; Smaili, K. Machine translation experiments on padic: A parallel Arabic dialect corpus. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, Shanghai, China, 30 October–1 November 2015.
6. Al-Gaphari, G.H.; Al-Yadoumi, M. A method to convert Sana'ani accent to Modern Standard Arabic. *Int. J. Inf. Sci. Manag.* **2012**, *8*, 39–49.

7. Ridouane, T.; Bouzoubaa, K. A hybrid approach to translate Moroccan Arabic dialect. In Proceedings of the 9th International Conference on Intelligent Systems: Theories and Applications, Rabat, Morocco, 7–8 May 2014; pp. 1–5.
8. Salloum, W.; Habash, N. Elissa: A dialectal to standard Arabic machine translation system. In Proceedings of COLING 2012 24th International Conference on Computational Linguistics, Mumbai, India, 8–15 December 2012; pp. 385–392.
9. Sadat, F.; Mallek, F.; Boudabous, M.; Sellami, R.; Farzindar, A. Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Application—The case of Tunisian Arabic and the Social Media. In Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, Dublin, Ireland, 24 August 2014; pp. 102–110.
10. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
11. Baniata, L.H.; Park, S.Y.; Park, S.B. A Neural Machine Translation Model for Arabic Dialects That Utilises Multi-Task Learning (MTL). *Comput. Intell. Neurosci.* 2018 (in press).
12. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
13. Liu, P.; Qiu, X.; Huang, X. Adversarial multi-task learning for text classification. *arXiv* **2017**, arXiv:1704.05742.
14. Costa-Jussà, M.R. Why Catalan-Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Valencia, Spain, 3 April 2017; pp. 55–62.
15. Hassani, H. Kurdish interdialect machine translation. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Valencia, Spain, 3 April 2017; pp. 63–72.
16. Costa-jussà, M.R.; Zampieri, M.; Pal, S. A Neural Approach to Language Variety Translation. In Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Santa Fe, NM, USA, 20–21 August 2018; pp. 275–282.
17. Sennrich, R.; Haddow, B. Linguistic input features improve neural machine translation. *arXiv* **2016**, arXiv:1606.02892.
18. Niehues, J.; Cho, E. Exploiting linguistic resources for neural machine translation using multi-task learning. *arXiv* **2017**, arXiv:1708.00993.
19. Neco, R.P.; Forcada, M.L. Asynchronous translations with recurrent neural nets. In Proceedings of the International Conference on Neural Networks, Houston, TX, USA, 9–12 June 1997; pp. 2535–2540.
20. Schwenk, H.; Dchelotte, D.; Gauvain, J.L. Continuous space language models for statistical machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL), Sydney, Australia, 17–18 July 2006; pp. 723–730.
21. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
22. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
23. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 3104–3112.
24. Passban, P.; Liu, Q.; Way, A. Translating low-resource languages by vocabulary adaptation from close counterparts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2017**, *16*, 29.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
26. Dong, D.; Wu, H.; He, W.; Yu, D.; Wang, H. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 1723–1732.

27. Firat, O.; Cho, K.; Bengio, Y. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv* **2016**, arXiv:1601.01073.
28. Ha, TL.; Niehues, J.; Waibel, A. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv* **2016**, arXiv:1611.04798.
29. Duh, K.; Kirchhoff, K. POS tagging of dialectal Arabic: A minimally supervised approach. In Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor, MI, USA, 29 June 2005; pp. 55–62.
30. Habash, N.; Roth, R.; Rambow, O.; Eskander, R.; Tomeh, N. Morphological analysis and disambiguation for dialectal Arabic. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 426–432.
31. Darwish, K.; Mubarak, H.; Abdelali, A.; Eldesouki, M.; Samih, Y.; Alharbi, R.; Attia, M.; Magdy, W.; Kallmeyer, L. Multi-Dialect Arabic POS Tagging: A CRF Approach. In Proceedings of the Language Resources and Evaluation Conference (LREC), Miyazaki, Japan, 7–12 May 2018.
32. Eldesouki, M.; Samih, Y.; Abdelali, A.; Attia, M.; Mubarak, H.; Darwish, K.; Laura, K. Arabic Multi-Dialect Segmentation: Bi-LSTM-CRF vs. SVM. *arXiv* **2017**, arXiv:1708.05891.
33. Bouamor, H.; Habash, N.; Oflazer, K. A Multidialectal Parallel Corpus of Arabic. In Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 26–31 May 2014; pp. 1240–1245.
34. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247.
35. Manavalan, B.; Shin, T.H.; Lee, G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* **2018**, *9*, 476.
36. Wei, L.; Zhou, C.; Chen, H.; Song, J.; Su, R. ACPred-FL: A sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* **2018**, *34*, 4007–4016.
37. Basith, S.; Manavalan, B.; Shin, T.H.; Lee, G. iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 412–420.
38. Qiang, X.; Zhou, C.; Ye, X.; Du, P.F.; Su, R.; Wei, L. CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief. Bioinform.* **2018**, doi:10.1093/bib/bby091.
39. Manavalan, B.; Govindaraj, R.G.; Shin, T.H.; Kim, M.O.; Lee, G. iBCE-EL: A new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.* **2018**, *9*, 1695.
40. Al-Rfou, R.; Perozzi, B.; Skiena, S. Polyglot: Distributed word representations for multilingual NLP. *arXiv* **2013**, arXiv:1307.1662.
41. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *arXiv* **2016**, arXiv:1607.04606.