Evaluating LSTM Networks, HMM and WFST in Malay Part-of-Speech Tagging

Tien-Ping Tan[1], Bali Ranaivo-Malançon[2], Laurent Besacier[3], Yin-Lai Yeong[1], Keng Hoon Gan[1], and Enya Kong Tang[1]

[1]*School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia.*

[2]*Faculty of Computer Science & Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia.*

[3]*LIG, Université Grenoble Alpes, CNRS, Grenoble, France.*

tienping@usm.my

***Abstract*—Long short term memory (LSTM) networks have been gaining popularity in modeling sequential data such as phoneme recognition, speech translation, language modeling, speech synthesis, chatbot-like dialog systems and others. This paper investigates the attention-based encoder-decoder LSTM networks in Malay part-of-speech (POS) tagging when it is compared to weighted finite state transducer (WFST) and hidden Markov model (HMM). The attractiveness of LSTM networks is its strength in modeling long distance dependencies. Malay POS tagging is examined from two different conditions: with and without morphological information. The experiment results show that LSTM networks that are trained without any explicit morphological knowledge perform nearly equally with WFST but better than HMM approach that is trained with morphological information.**

***Index Terms*—Malay Part-Of-Speech Tagging; Recurrence Neural Network (RNN); Long Short Term Memory (LSTM) Networks, Sequence-To-Sequence Learning.**

I. INTRODUCTION

Recently, neural networks have been gaining popularity in the field of artificial intelligence. The advancements are due to the breakthrough in the algorithms that learn and recognize very complex patterns using deep layers of neural networks or commonly known as the deep neural networks (DNN) [1], and the introduction of different types of neural network such as convolutional neural network and recurrent neural network (RNN). For instance, convolutional neural networks, which are special type of feed-forward neural networks with two-dimensions networks, have shown tremendous accuracy in classifying images through local receptive fields, shared weights, pooling, from simple handwritten digit recognition to more complex face recognition. In the modeling of sequential patterns, such as phoneme recognition [2], automatic speech recognition [3][4], speech synthesis [5], speech translation [6], chatbot and many others, RNN or the more specialized type of RNN, the long short term memory (LSTM) networks have shown to be better than many of the traditional approaches.

This paper presents a comparative study of three methods to solve the problem of Malay part-of-speech (POS) tagging. These methods are LSTM networks, weighted finite state transducer (WFST) and hidden Markov model (HMM). The objective is to examine the performance of the current state of the art attention-based encoder-decoder LSTM networks while compared to WFST and HMM in POS tagging. POS tagging is a language processing task that assigned a POS tag (e.g., noun, verb, adjective, etc.) to each word in a sentence. Taking a different approach, in this study, the pairs of word/POS tag are not provided. Instead, the proposed model will learn the sequence-to-sequence mapping from the sequential data provided. The benefit of this approach is that, for certain languages without clear word boundary, the implicit word boundary knowledge is learnt from the data. The main challenge for the algorithm is to find the word alignment information from the data provided as illustrated in the examples in Table 1.

Table 1
Example Sentences and their POS

No	Malay Sentence	Meaning (English)	POS Annotation
1.	*pasaran buruh*	labor market	N N
2.	*kedua - dua benua*	both continents	NUM_CART N
3.	*cintaku*	my love	GEN_PRO N
4..	*kuala lumpur*	Kuala Lumpur	N

In addition, the examined approach must find the alignment between the word and its POS tag from the data, with the possibility that a word (a string separated by space) may map to more than one POS tag (example 3 in Table 1), or more than one word may map to a single POS tag (example 4 in Table 1).

II. MALAY AND POS TAGGING

Malay is the official language used in Malaysia, Indonesia, Singapore, and Brunei. Malay is an agglutinative language. As such, new words can be created by adding one or several – less than three – affixes to a base word. The affixed can be the host of proclitic, enclitic and particle. Figure 1 shows the morphological structure of a Malay word [7].

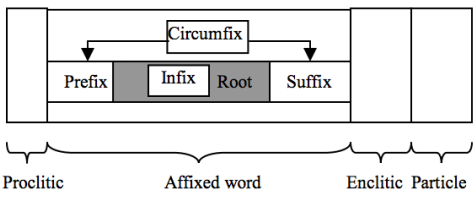

Figure 1: Morphological structure of Malay word [7]

The two proclitic (*ku-* 'I' and *kau-* 'you') and four enclitics (*-ku* 'me, my', *-kau* 'you, your', *-mu* 'you, your' and *-nya*