HelaNER 2.0: A Novel Deep Neural Model for Named Entity Boundary Detection

Y.H.P.P Priyadarshana, L Ranathunga
Department of Information Technology, University of Moratuwa, Moratuwa, Sri Lanka

***Abstract*—Named entity recognition (NER) is a sequential labelling task in categorizing textual nuggets into specific types. Named entity boundary detection can be recognized as a prominent research area under the NER domain which has been heavily adapted for information extraction, event extraction, information retrieval, sentiment analysis etc. Named entities (NE) can be identified as per flat NEs and nested NEs in nature and limited research attempts have been made for nested NE boundary detection. NER in low resource settings has been identified as a current trend. This research work has been scoped down to unveil the uniqueness in NE boundary detection based on Sinhala related contents which have been extracted from social media. The prime objective of this research attempt is to enhance the approach of named entity boundary detection. Considering the low resource settings, as the initial step, the linguistic patterns, complexity matrices and structures of the extracted social media statements have been analyzed further. A dedicated corpus of more than 100,000 tuples of Sinhala related social media content has been annotated by an expert panel. As per the scientific novelties, NE head word detection loss function, which was introduced in HelaNER 1.0, has been further improved and the NE boundary detection has been further enhanced through tuning up the stack pointer networks. Additionally, NE linking has been improved as a by-product of the previously mentioned enhancements. Various experimentations have been conducted, evaluated and the outcome has revealed that our enhancements have achieved the state-of-art performance over the existing baselines.**

Keywords*—Computational linguistics; deep neural networks; natural language processing; named entity boundary detection; named entity recognition

I. Introduction

Named entity recognition (NER) is an active research area and a prominent computational task under the key area of natural language processing (NLP) [1]. Further, NER can be recognized as a semantic level sequential labelling task in terms of identifying textual spans into the specified categories such as Person, Location and Organization. Several textual spans can be categorized under the above-mentioned predefined categories which are known as named entities (NE). Named entities can be categorized into two main different aspects such as named entity type and the named entity boundary. When it comes for the NE type aspect, two variations can be established such as fine-grained NERs and coarse-grained NERs. Fine-grained NERs typically consider much larger scopes and its contribution towards the NE boundary aspect would be slightly low. When it comes for coarse-grained NERs, which is considered as high potential capabilities in demonstrating the NE boundary related aspects on the basis of specific niche NE types. The specific niche NE types have already been taken into the consideration under the flat NE boundary aspects [2]. Considering the NE boundary and type aspects still there is a potential vacuum under NE boundary avenues compared to the other variation, NE type component.

A novel approach has been invented through a recent prominent research activity to detect the respective boundaries of named entities [2]. The solution has been focused on the domain of Sinhala statements which have been extracted from social media such as Facebook, YouTube, and Twitter. This solution has been further enhanced from HelaNER 2.0 where an enhanced algorithm has been demonstrated to determine the respective named entity mention boundaries. These novel enhancements would be worthwhile to showcase the importance of addressing the social linguistic issues specially for low resource language settings. Named entity boundary detection still can be recognized as an unknown area considering Sinhala language. In terms of manipulating the social media contents which have been published in Sinhala, it is really worthwhile to undercover these types of niche domains. Compared to the other available benchmarks for NE boundary detection, several novel avenues and enhancements have been identified and showcased from this research attempt. The main contributions of this paper consist of the following:

1) The size of the combinational multi-layer classifier which is dedicated for NE head word detection along with the NE type detection has been increased. Additionally, the NE head word detection loss function has been further improved. Due to these enhancements in HelaNER 2.0, the performance of detecting the NE head words has been increased.

2) The stack pointer network which has been used for NE boundary detection under [2], has been further tuned up and improved. Specially, the 2nd and 3rd hidden layers which are dedicated for deriving character level representations have been further enhanced. These enhancements have been critically demonstrated and evaluated in the respective sections of this paper. Due to this enhancement, the overall accuracy for NE boundary detection process has been increased in a considerable margin compared with the other existing benchmarks.

3) As a by-product, the overall NE linking process has been further improved due to the above-mentioned enhancements. These novel enhancements have been critically analyzed and showcased under the methodological section.

4) Those main contributions have been critically evaluated considering the state-of-the-art benchmarks. The corpus has been increased compared to our previous research attempt [2]. The results have been demonstrated under the experimental results section.

The ultimate motivation of this research attempt is to develop a novel framework for named entity boundary detection for social media analysis. Detecting both NE boundary and NE type for named entities as an aggregate mechanism will eventually tune up the accuracy and performance of NE linking to knowledge bases. This proposed novel framework allows to capture the named entity boundary detection along with their respective named entity types by filling out the research gap which has been identified where the overall solution is capable of handling multiple domains including Sinhala. This framework can be further introduced as a specialized niche version for the domain Sinhala and is capable of enhancing up to a more generalized version considering multiple domains.

The rest of this paper is organized as follows. Related work about named entity (NE) boundary detection along with the specific NE type is given in Section II. Section III presents the scientific methodology of the overall approach in detail manner. Section IV introduces the experimental results of this entire solution along with the comprehensive evaluation procedure which has been conducted considering the existing prominent baselines. The discussion and future avenues of this novel mechanism have been discussed in Section V.

II. Related Work

Multiple deep neural and non-neural based computational systems have been established to showcase the usage of detecting NER in social media analysis in the recent past. Some of such systems have shown promising results over the existing benchmarks. The overall literature review would be separated into two sections such as analyzing the existing NE boundary detection systems along with the respective algorithms and critically analyze the statements which have been circulated in the social media platforms.

A. Analyzing NE Boundary Detection

Due to the low language resources settings, deep transfer learning has been adapted to determine the starting and ending indices of name entities [3]. Such a system has been introduced to overcome certain issues such as removing the noisy data and avoiding hand-crafted feature settings. Due to the noisiness of the corpus, the system has failed in determining the respective NE types along with detecting the boundaries. In terms of enhancing NER, multiple NE linking mechanisms of detecting boundaries have been introduced considering the lexical and morphosyntactic features [4]. Classifying textual segmentations into text span identification has been identified as a significant drawback of this approach. A novel stack pointer networks-based approach along with deep adversarial transfer learning which is capable of detecting the NE boundaries has been demonstrated [5]. This system has a unique capability of deriving both starting and ending NE boundary tags but still has failed in detecting the respective NE type. In terms of addressing the issue of [6], a context encoding neural model has been invented [7]. A combination of Bi-LSTM (Long Short-Term Memory) model with a CNN (Convolutional Neural Network) to capture character-level features relates to NE type and boundary aspects can be identified as the main contribution of the above-mentioned approach. Even though this model has performed well in capturing NE boundaries still the NE type capturing has to be improved further.

BDRYBOT, a neural based model for detecting NE boundaries has been demonstrated considering the NE linking procedures [1]. The model has been consisted with a CNN in terms of enhancing the character level contextual representations. The encoding phase has been enriched with a Bi-GRU (Gated Recurrent Units) model instead of Bi-LSTM considering low computational consumption. When it comes for the NE type detection along with NE boundary detection, still the system has not been improved to achieve up to that level and can be stated as a future avenue. A novel multitasking learning neural based boundary detection model has been introduced considering the nature of inner and outer layers of nested entity mentions [8]. This approach has been enriched with sequential classification tasks in terms of reducing the computational costs hence it has demonstrated some promising results in nested NER over the existing benchmarks. Even though this model has shown some promising results, still it has been limited for extracting implicit NE mention regions. Another hybrid NE tagging architecture has been showcased considering dual neural models, Bi-LSTM, and stack LSTM [9]. Here, a special NE tagging pattern, IOBES scheme has been adapted [10]. Though the NE identification performance has been boosted up due to this novel NE tagging pattern still this has been limited to NE explicit variations.

A novel deep neural network based exhaustive approach has been invented for nested NE mention recognition [11]. Here, both NE boundary related information and NE type related information have been considered. Additionally, inner NE feature representation also has been considered under this approach. Bi-LSTM model has been developed to fulfill the objectives by visualizing the character-based feature representations on top of contextual word vectors. Compared to the other well-known approaches which have been dedicated for nested NE tasks [12] [13] [14], the exhaustive model has shown promising outcomes under the time related complexities as well. Still there is a vacuum for entity mention outside feature representations apart from only limiting to inside feature mappings. A recent deep neural based approach, anchor region networks, has been demonstrated to showcase the value and impact of nested NEs [15]. The primary assumption has been set off as the head or main words would govern the whole structure of entity mention nuggets considering the respective boundary aspects. These fundamentals have been used to determine the benchmarking models with some major enhancements. Span based models [16] [17] [18] have been considered as a revolutionary approach for determining NER. Classification of sentence nuggets into the respective subsequences using span-based architecture [19] has been exhibited recently. Both multitask learning and BERT based classifications have been used to

play with the boundary level aspects of NEs. Even though this has been accepted as a novel methodology, still the conducted experiments have been revealed that the performance of detecting NE boundaries through span-based models is poor.

An enhanced Machine Reading Comprehensive (MRC) model has been introduced in terms of addressing both flat NERs and nested NERs scenarios [20]. This novel model has focused on the theoretical approaches under the sequential labelling domain. In nature, MRC models are dedicated for extracting NE mention spans regardless of flat or nested nature [21]. The text classification objective has been accomplished using BERT based models [22]. Even though this model has been performed well enough for coarse-grained NER types, still there is a demand for such an advancement for considering fine-grained NER types as well. There is a trend of adapting object detection and mapping techniques in the theory of computer vision for extracting NE mentions. One of such advancements has been exhibited as NE boundary regression model [23]. In this approach, Nested NE mention detection has been conducted considering the overlapping of each NE mentions in the context by applying a concept called NE bounding boxes. A deep neural network model called convolutional feature maps has been invented for NE boundary detection. The same technique has been used for textual information extraction and information retrieval from movie reviews recently [24] [25]. Even though this model has showcased some promising results over the existing benchmarks still NE type detection aspect has been missed.

B. Analyzing Social Media Statements

Even though Sinhala is a morphological rich language, very limited amount of NLP resources has been introduced considering both micro level and macro level tasks in computational linguistics. Hence, very minimum facilities can be observed for accomplishing major activities like named entity boundary detection for Sinhala context. The paradigms of Sinhala social media statements analysis can be listed down as lexicon-based approaches and machine learning approaches [26]. The lexicon-based category has been specialized in applying for scenarios such as catering online forums, online blog posts and comments which have been extracted from online channels [27]. A specific speech lexicon has been adapted in terms of capturing verbs and nouns in the context [28]. Here, a NE recognizer has been used to detect the related nouns in the context. Since the whole procedure has been dominated under a particular lexicon, the approach can be recognized as a restrictive limited approach. In the recent past, an overwhelming demand has been experienced for the adaption of machine learning based models in capturing hate speech related contents. A novel approach has been introduced to pick out hate speech related content from German corpus [29]. This approach has been enriched with the adaption of transfer learning, web scrapping mechanisms and usage of Bag-Of-Words (BOW). Another similar approach has been conducted to extract speech related content from Indonesian language [30]. Here, a random forest classifier has been adapted to showcase the value of word n-gram feature representations in classifying social media statements.

When it comes for the avenues in deep learning for classifying social media statements, the linguistic level matrices under micro level and macro level should be examined. A significant corpus is essential for implementing a valid supervised type of deep neural based model for accomplishing the analysis of extracted statements. Crowd sourcing can be identified as a handy approach to fulfill the issue of corpus unavailability [31] [32]. Those extracted data would be used to determine the respective word embeddings since it has been identified as the state of the art which has been discovered under textual processing [32] [33]. A unique classifier which has been invented adapting ensemble fundamentals has shown some promising results in classifying sexist and racist corpora [34]. Another novel LSTM based model has been introduced in terms of classifying speech related content in Italian language [35]. Those attempts can be identified as the fundamentals in deep neural models which have been invented for classifying social media content.

III. Research Methodology

The overall methodological approach can be categorized into two main components such as discovering the Sinhala related posts, comments, and statements which have been spread out in social media context through mandatory features, complexity measurements and other related aspects and constructing a novel framework for capturing named entity boundary detection considering the respective named entity type.

A. Sinhala NE Boundary Detection

The ultimate goal is to turn up with a neural based methodology to discover and analyze the Sinhala related social media context analysis based on a supervised approach. As the initial phase, Sinhala social media related corpus has been constructed. A specific web crawling mechanism has been used to extract the social media statements from Facebook, Twitter, and YouTube platforms. Once the statements have been crawled, those have been collected to a pool to be distributed among multiple set of annotators who are responsible for conducting manual annotations. More than 100,000 social media statements have been crawled for this purpose. Once the annotated statements have been constructed, then the next sets of preprocessing tasks have been determined. Several important intermediate steps have been followed such as preprocessing, stemming, parts-of-speech (POS) tagging to filter the dataset in terms of using it in the next phases. Considering the Sinhala POS tagging, due to the unavailability of a standard Sinhala POS tag set, a special pseudocode has been implemented and demonstrated under the Fig. 1.

Then as per the final step under this section, the specific neural model has been designed and implemented. A recurrent neural network (RNN) LSTM has been used as the foundation for implementing the deep neural model since RNN has been accepted as the state-of-the-art in processing sequential representations. Some optimal hyper-parameter settings have been enriched for obtaining better results under the training procedure. The hyper-parameter settings can be identified as a vital component of the entire procedure where some of the critical evaluations have been conducted considering different value adjustments. The deep neural model which is dedicated for NE boundary detection can be showcased as per Fig. 2.

```
Algorithm 4.1: POS Tagging
SET chunker TO RegexpParser(r'''
    NP:
    {<NNPI>*<JJ>*<NNN>*<NNN><NNPA.*>*<NNPI.*>*<PRP.*>*}
    VP:
    {<JVB>*<NVB>*<V.*>*}
  ''')
    FOR subtree IN parsed_tree.subtrees(filter=lambda t: t.label() EQUALS 'NP'):
      SET first_so TO []
      SET result TO subtree.leaves()
      OUTPUT(result)
      FOR word, tag IN result:
        first_so.append(word)
        SET so TO ' '.join(first_so)
      so_list.append(so)
      OUTPUT('SO List:' + str(so_list))
    IF len(so_list) != 2:
      OUTPUT('System has not taken the Subject-Object correctly')
    ELSE:
      SET sub TO so_list[0]
      SET ob TO so_list[1]
  RETURN triple
```

Fig. 1. Pseudocode for Parts of Speech (POS) Tagging.

```
Algorithm 4.1: NE Identification
FOR max_len IN max_len_arr:
  SET startTime TO time.time()
  SET tok TO Tokenizer(num_words=max_words)
  tok.fit_on_texts(X_train)
  SET train_sequences TO tok.texts_to_sequences(X_train)
  SET train_sequences_matrix TO sequence.pad_sequences(train_sequences,
maxlen=max_len)
  DEFINE FUNCTION RNN():
    SET INPUTs TO Input(name='INPUTs', shape=[max_len])
    SET layer TO Embedding(max_words, 50,
INPUT_length=max_len)(INPUTs)
    SET layer TO LSTM(10)(layer)
    SET layer TO Dense(256, name='FC1')(layer)
    SET layer TO Activation('relu')(layer)
    SET layer TO Dropout(0.5)(layer)
    SET layer TO Dense(1, name='out_layer')(layer)
```

Fig. 2. Pseudocode for LSTM based Deep Neural based Model for Sinhala NE Boundary Detection.

B. Determining NE Boundary Detection

Once the initial methodological step is ready by identifying and classifying the named entity identification for Sinhala, then the second phase which is determining the NE boundary detection can be initiated. The proposed architecture of the entire procedure for deriving named entity boundary detection can be showcased as Fig. 3. As per the word representation, word level embedding, and character level embedding have been followed. Since the embedding process is at a basic stage for Sinhala, Glove word embedding which has been developed based on English corpus will be used as the platform. The obtained embedding values have been fed to a Bi-LSTM layer in terms of generating the backward and forward value sets.

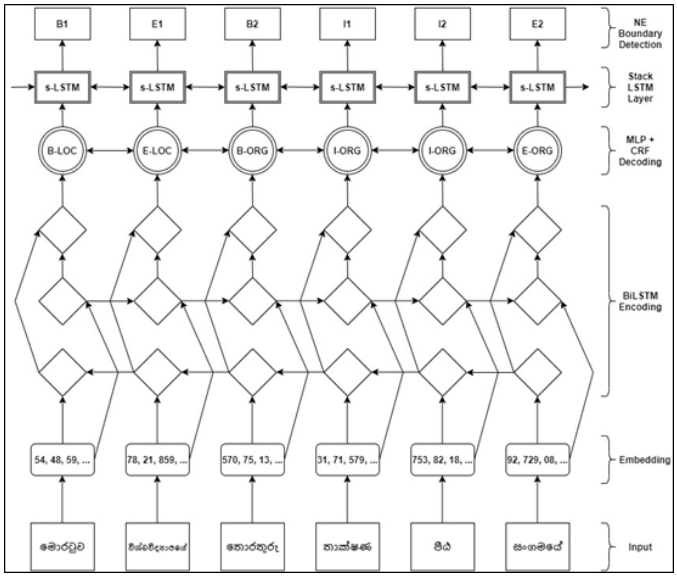

Fig. 3. Proposed Architecture for Named Entity Boundary Detection.

Special LSTM based deep neural model has been developed as the overall NE detection end to end framework. A Bi-LSTM layer would be used to derive the word level classifiers to obtain the respective head word which acts as the dominant words of the pre-defined context. As an example, considering English context, “Dilan Perera has been elected as the Secretary of the Colombo Sports Club”, Dilan Perera acts as a PER mention type while Colombo Sports Club takes ORG mention type. Additionally, Colombo could be recognized a LOC type. Considering the whole nugget, even though there are three main NE mentions, PER mention can be considered as the main or the head NE which governs the overall semantic contextual representation. So, the first step is to determine the head word of each unique sentence nugget. Once the head words have been obtained, the relevant contextual level representations have been fed into a structured classifier called multi-layer perceptron (MLP) along with the conditional random fields (CRF). MLP is capable in nature to derive the inner representations of the word level contextual representations. A combined classifier has been used to improve the level of accuracy and this combined mechanism can be mentioned as a unique approach which has not been used for any of the previous NE boundary detection frameworks. The whole process can be visualized as Fig. 4 as follows.

Once the head words have been obtained, the next step is to locate the specific entity mention nuggets per each dominant word. Here a novel theory called boundary bubbles has been introduced. Boundary bubbles (BB) are abstractive level contextual representations which are used to visualize the NE mention nuggets along with their respective identified NE head words. Fig. 5 shows the demonstration of boundary bubbles for the previously given example (Dilan Perera has been elected as the Secretary of the Colombo Sports Club).

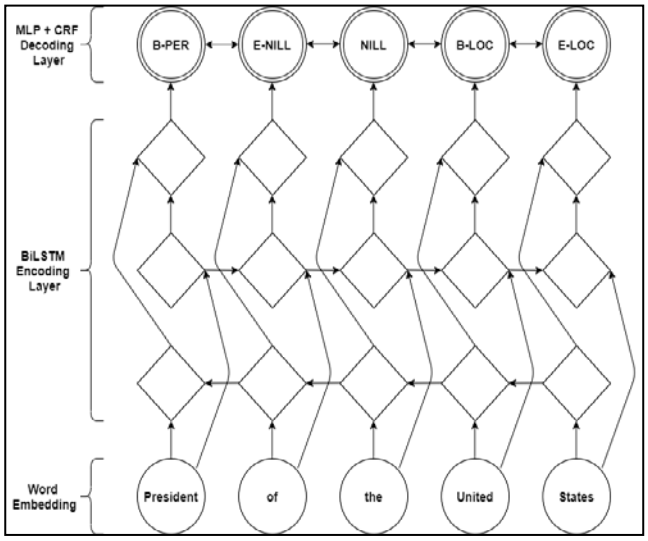

Fig. 4. Main System Architecture of NE Type Detection.

$$[[\textbf{Dilan Perera}]_{PER}\ has\ \dots\ the\ [\textbf{Colombo Sports Club}]_{ORG}]_{PER}$$

Fig. 5. Abstract Level Representation of Boundary Bubbles.

In terms of capturing mention nuggets, a special algorithm has been developed. Several parameters and assumptions have to be considered when declaring the specific algorithm. Two main such assumptions would be like, at least one word must be declared as the head word for that specific mention nugget and such a word should be found through the specific boundary bubble. The left boundary region and the right boundary region of the respective bubble should be determined for identifying the mention nuggets. Considering left and right boundary related values, two value tuples have been defined. A unique loss function has to be determined since it's mandatory to define the loss value under head word detection procedure. The head word detection loss function can be obtained as: $-logP(c_i|x_i)$.The collective loss function would be determined as a weighted average loss value considering the respective scenarios including the head word detector would be resulting to NILL. The NE mention nuggets identifier can be recognized as per (1) as follows.

$$\tau(x_i;\emptyset) = v_i.[-logP(c_i|x_i) + L^R(x_i;\emptyset)] + (1 - v_i).[-logP(NIL|x_i)] \quad (1)$$

$$L^R(x_i;\emptyset) = L^{left}(x_i;\emptyset) + L^{right}(x_i;\emptyset) \quad (2)$$

The respective above stated (2) would determine the identical structure for NE mention nuggets along with the respective starting and ending boundary regions. Additionally, v_i under (1), states the respective correlational value of the underline boundary bubble, in which the higher the value determines the stronger association considering the type of the bubble. The whole process would be complex if multiple head words per mention nuggets have been established. With all of these assumptions and decisions, v_i can be further demonstrated as:

$$v_i = \left[\frac{P(c_i|x_i)}{max_{x_t \in B_i} P(c_i|x_t)}\right] \alpha^v \quad (3)$$

Here α can be showcased as a specific hyper-parameter. Different values can be assigned with the intention of analyzing the outcome behavior of the overall procedure. As the basic value assignment, $\alpha = 0$ could generate a scenario where the determined NE mentions would be annotated with the respective boundary bubble type. When it comes for extracting the inner most head words, B_i determines the region of the extracted mention nuggets.

The final segment of the whole process of named entity boundary detection and type determination will be the boundary region classification. From the previous step, the respective NE mention nuggets have been derived and the next step is to determine the boundary detection using the obtained mention nuggets. In terms of deriving the NE boundary detection, a novel NE scope deriving mechanism which is called as IOBE (Inside, Outside, Beginning, End) pattern has been used. Even though IOB (Inside, Outside, Beginning) pattern has been adapted for NE boundary detection, usage of IOBE pattern can be identified as the first-time approach of adapting IOBE pattern along with stack pointer networks in terms of deriving the boundary detection. Plenty of other different patterns have been used for accomplishing different avenues under the named entity recognition and named entity boundary detection domains. Specially, BEO (Beginning, End, Outside) pattern has been used for one of the previous approaches under NE boundary assembling mechanisms for named entity recognition in biomedical domain [36]. As per our understanding, this is the initial attempt of using IOBE pattern for named entity boundary detection purposes under the named entity recognition domain. Further, this can be mentioned as one of the scientific core novelties of this research work. The respective design architecture related to NE boundary detection can be exhibited as follows under Fig. 6.

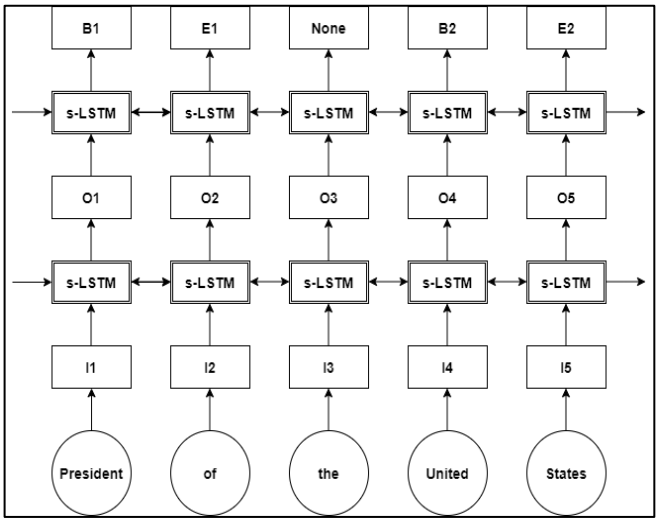

Fig. 6. Stack LSTM Architecture of NE Boundary Detection.

NE linking is a crucial task under information retrieval from the respective knowledge bases. Enhancing NE linking can be identified as a sub product of this overall procedure. Considering the three different components under the NE linking, our intention is to enhance the direct mapping through

the identification of NE boundaries. An explicit novelty will not be achieved from NE linking but a performance increment would only be expected due to the enhancements which have been exhibited under NE boundary detection. The respective enhanced algorithm which is dedicated for NE direct mapping purposes can be exhibited as (4). In the respective algorithm, $p(c_{ij}|e_i)$ determines the respective contextual level probabilistic distribution of each entity mention denoted as e_i. The Laplacian smoothing has been adapted to avoid the zero probability values.

$$p(c_{ij}|e_i) = \frac{t(e_i,c_{ij})+1}{\#e_i+|c_i|} \quad (4)$$

The overall implementation has been carried out in a well constructive manner. Python 3.7.4 has been used as the primary programming language to implement the overall system including all the components. Frameworks such as TensorFlow 1.15 and Flask 2.1 have been adapted to deliver the whole application in a much standard way. The mandatory natural language processing and computational linguistics libraries such as Torch 1.7, Spacy 2.1.9, Keras 2.2.4, NLTK 3.4.5 have been adapted to fulfill the overall implementation procedure in a smooth manner. After the system has been implemented, the experimentation can be carried out, and the related experimental results can be described in detail manner as follows.

IV. Experimental Results

The overall testing procedure has been conducted considering all the required aspects. An extensive experimental process has been carried out considering all the implemented components which have been mentioned earlier. Considering the overall experimental settings, several existing baseline mechanisms have been considered such as regional based models, conventional CRF models and hypergraph based computational models. A unique Sinhala NER corpus has been obtained which consists with 120,000 tuples as per training purposes, 100,000 tuples as per validation purposes and another 80,000 set as per accomplishing testing purposes. As per the hardware requirements, a high-performance graphics processing unit (GPU) machine has been acquired. Major sets of hyper-parameters have been set off for obtaining the competitive edge. Experimentations and the respective evaluation process have been focused with several sub-components such as evaluation on Sinhala NE identification, evaluation on word embedding, evaluation on NE head word detection, evaluation on NE mention boundary detection and evaluation on NE linking.

A. Evaluation on Sinhala NE Identification

Several annotators have been employed for constructing the main corpora such as classifying social media statements and grouping the identified named entities into NE categories. Hence, the Fleiss' Kappa [37] values have to be measured to evaluate the reliability of the corpus which has been used in the entire system. Once the annotation process has been accomplished, the inter annotator agreement has been processed to evaluate the relevant Kappa statistics. These respective individual Kappa statistics can be showcased under the Table I as follows.

TABLE I. Kappa Values For NE Categories

Kappa Values for Individual NE Categories					
NE Class	**Conditional Probability**	**Kappa**	**Standard Error**	**Z Value**	**P Value**
PER	0.84	0.76	0.12	5.12	0.04
LOC	0.76	0.64	0.12	4.87	0.04
ORG	0.67	0.62	0.12	4.28	0.03
DES	0.48	0.57	0.12	3.72	0.02
PRO	0.43	0.52	0.12	3.24	0.02

Once the Kappa values have been obtained for the respective individual NE categories, then the overall Kappa statistics can be obtained and demonstrated as follows.

According to the values of Table II, the overall Kappa value for NE identification and categorization can be recognized as 0.72. Considering the overall value and according to the classification of the Fleiss' Kappa, it can be concluded that the generated Kappa value has represented a good strength in the inter annotator agreements.

TABLE II. Overall Kappa Values for NE Categories

Overall Kappa					
NE Class	**Conditional Probability**	**Kappa**	**Standard Error**	**Z Value**	**P Value**
Overall	0.78	0.72	0.11	4.86	0.02

In terms of evaluating the model on NE type classification, set of main parameters have been defined and discussed in the previous chapter. The obtained results have been critically evaluated with the available benchmarks as per the Table III. The newly introduced unique model has been demonstrated as BB2022 (boundary bubbles).

TABLE III. Evaluation on NE Type Classification

Model	**Precision (%)**	**Recall (%)**	**F1 (%)**
SH2016[1]	75.9	70.1	72.8
KBP2018	72.6	73.0	72.8
ARN2019	75.2	72.5	73.9
BB2022	***76.2***	***73.6***	***74.9***

As per the exhibited comparison outcomes, it can be concluded that the proposed novel model performs better than the existing NE type detection benchmarks in a competitive edge.

B. Evaluation on Word Embedding

Due to the low level of language resources settings, a transfer learning mechanism must be adapted to derive the Sinhala related word embeddings. Specific four main hyper-parameters have been used in terms of deriving the training procedure such as dev detect, test detect, dev all and test all. Once the overall testing process has been designed, a unique set of 500 tuples have been used for testing procedure

[1] Existing benchmarks which have been implemented and available for NE type identification purposes.

considering the above-mentioned testing related hyper-parameter settings. As per the Table IV, the respective highest performance has been demonstrated once the value settings have been set to 0.6.

TABLE IV. RESULTS ON WORD EMBEDDING

Set	Dev Detect	Test Detect	Dev All	Test All
500	84.68	84.27	84.39	84.76

Once the word embeddings have been obtained, respective evaluation has been conducted considering the most prominent word embedding techniques such as Glove, Word2Vec, ELMo and BERT. The respective comparison can be exhibited under Fig. 7 as follows. As per the comparison when the k value gets 1.4 a performance increment can be observed considering all models except ELMo. When Word2Vec, Glove and BERT models have been considered, the highest outcome has been demonstrated by our approach which is the Glove model. Even though BERT has shown some promising results in most of the previous scenarios as per the constructive literature survey, the Glove model can be mentioned as an optimal alternative approach, especially for low resource language settings.

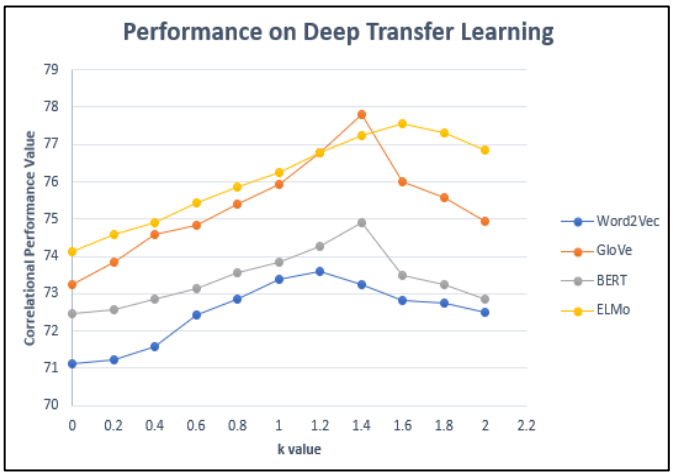

Fig. 7. Performance Comparison on Word Embedding through Transfer Learning.

Later, with the introduction of cross-lingual word embeddings, some state-of-the-art word embedding models have been invented. One of such prominent mechanisms would be the usage of XLM-R for obtaining word embeddings considering major text classification tasks [38] [39]. Hence, another evaluation approach has been applied to compare the proposed model and the XLM-R model where the results can be demonstrated under the Fig. 8 as follows.

As per the comparison between the GloVe and XLM-R models, it can be concluded that the maximum outcome has been showcased by GloVe even though XLM-R has performed well in most of the sections in the graph distribution.

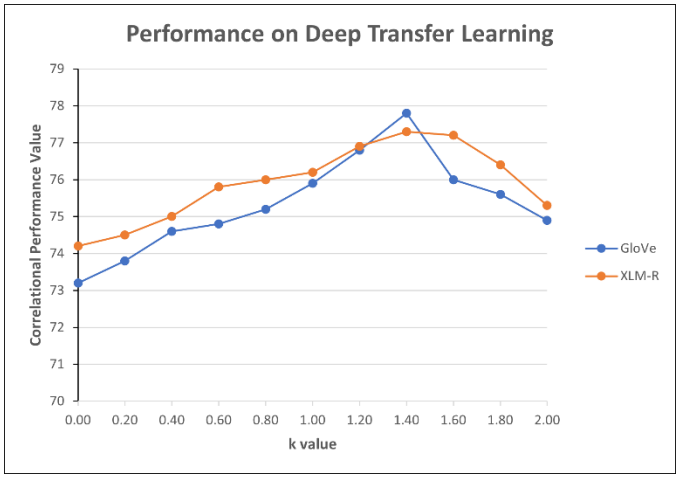

Fig. 8. GloVe and XLM-R Performance Comparison on Transfer Learning.

C. Evaluation on NE Head Word Detection

NE head word detection has been identified as one of the scientific core novelties of this entire research attempt. Considering the major evaluation matrices which have been stated under the word embedding and the head word loss value, the testing procedure can be determined. Considering the key factors such as NE mention identifier and the head word detection loss value, the respective experimental results can be obtained and listed under Table V as follows.

As per the evaluation purposes, the existing NE head word detection mechanisms related to the flat NEs have been considered. Those existing benchmarks models have already been critically analyzed and reviewed under the comprehensive literature survey and the evaluation procedure has been conducted respectively. The respective evaluation results for the NE head word detection process can be showcased as per Table VI. Our novel model has been denoted as BB2022 (Boundary Bubbles). As per the competitive analysis, the most prominent deep neural based NE detection models such as Sohrab et al. 2018 [11], Ju et al. 2019 [16] and ARN2019 [15] have been used. These models have been recompiled and re-generated using the openly available source code in the respective code repositories. Considering the evaluation results it can be concluded that our novel model has performed well against all the prominent baselines which are available for NE head word detection.

TABLE V. RESULTS ON NE HEAD WORD DETECTION

Metric	Precision (%)	Recall (%)	F1 (%)
Dev Detect	85.973315	78.53125	82.083945
Test Detect	78.783828	81.76144	
Dev All	0.2187500	0.239808	
Test All	0.2338790	0.253577	
Dev Gap	85.707962	78.31259	81.844137
Test Gap	84.696074	78.54994	81.507865

TABLE VI. EVALUATION ON NE HEAD WORD DETECTION

Model	Set	Precision %	Recall %	F1 %
Sohrab2018	1	71.68	71.25	71.28
	2	71.59	70.48	71.24
	3	71.43	71.39	71.31
	4	72.69	71.59	71.84
	5	71.89	71.52	71.64
	6	72.98	71.42	71.55
Ju et al2019	1	69.11	68.87	68.91
	2	68.47	68.24	68.12
	3	68.35	68.29	68.23
	4	68.48	68.22	68.31
	5	68.94	68.47	68.51
	6	69.24	68.86	68.23
ARN2019	1	78.13	76.21	78.06
	2	78.21	76.47	78.17
	3	78.65	76.34	78.35
	4	77.98	75.27	77.76
	5	77.83	76.39	77.76
	6	78.11	77.53	78.10
BB2022	1	78.95	77.35	78.54
	2	78.24	76.97	78.13
	3	78.88	77.31	78.24
	4	78.65	78.03	78.59
	5	79.16	77.68	78.88
	6	79.14	77.21	78.92

D. Evaluation on NE Boundary Detection

This component has already been mentioned as the next scientific core novelty of this overall research procedure under the methodology section. In other words, this component would be the core driving force of the entire solution. The respective performance related matrices have been generated for the five most prominent NE categories and the respective results can be demonstrated under Table VII as follows.

TABLE VII. RESULTS ON NE BOUNDARY DETECTION

NE Type	Precision (%)	Recall (%)	F1 (%)
ORG	0.84	0.89	0.86
DESIG	0.85	0.85	0.85
LOC	0.96	0.96	0.96
PER	0.96	0.95	0.93
DATE	0.88	0.91	0.89

Additionally, the accuracy also has been calculated considering the major hyper-parameters. Several hyper-parameters have been set off such as word embedding rate to be 25, output dimension to be 4 and the dropout rate as 0.1. Once the results have been obtained, the evaluation has been conducted. For evaluation purposes, as per the benchmarks, the systems which are dedicated for detecting NE boundaries have been used. The obtained evaluation results can be tabulated under Table VIII as follows. Considering the evaluation-outcome it can be concluded that our novel model, BB2022 (Boundary Bubbles), has performed well than the existing baselines.

TABLE VIII. EVALUATION ON NE BOUNDARY DETECTION

Model	Precision %	Recall %	F1 %
Sohrab2018	76.61	69.20	72.71
Ju et al2019	79.90	67.08	71.36
ARN2019	81.34	68.20	72.83
BB2022	**82.18**	**71.34**	**76.54**

E. Evaluation on NE Linking

Conducting the evaluation on NE linking can be described as the final phase of the overall evaluation procedure. In terms of accomplishing the NE linking procedure, several major topics under the Sinhala speech knowledgebase have been used. As per the methodology, only the direct mapping technique of the NE linking has been used so that the direct mapping technique should be evaluated as the evaluation criteria on the overall NE linking. Considering the major NE linking benchmarks, several prominent systems such as DBpedia Spotlight[2], SOTA[3] and VCG[4] can be listed. Our model has been evaluated with those existing benchmarks and the evaluation results can be showcased under Table IX as follows. As per the overall comparison, it can be concluded that our novel approach, BB2022 (Boundary Bubbles), has performed better than the existing benchmarks.

TABLE IX. EVALUATION ON NE LINKING

Model	Precision %	Recall %	F1 %
DBpedia Spotlight	52.64	52.48	52.53
SOTA	57.26	51.24	55.24
VCG	61.25	58.26	60.18
BB2022	**62.47**	**60.15**	**61.27**

V. DISCUSSION AND CONCLUSION

NE boundary detection in Sinhala context can be introduced as the core of this entire research attempt. The usage of social media for various different kinds of purposes has been increased in an alarming rate [8] [9]. There is a high demand for a sustainable computational framework for identifying and extracting such kind of various inputs which have been spread out in social media, especially in low NLP resources-based contexts like Sinhala [2]. Considering the overall approach of HelaNER 2.0, various enhancements have been designed, implemented, and evaluated. Major enhancements have been applied under the methodological components of word embedding through deep transfer learning, NE boundary detection and NE linking. All the predefined objectives such as Sinhala social media analysis, obtaining word embedding through transfer learning, NE head word detection, NE boundary detection and NE linking have been achieved. The novelties have been achieved in NE head word detection and NE boundary detection components. Respective evaluations have been conducted and the outcome has revealed that the novel approaches have outperformed the existing baselines in the market. Several points can be summarized as follows.

[2] https://www.dbpedia-spotlight.org/api

[3] Exploring Neural Entity Representations for Semantic Information, A Runge, E Hovy - arXiv preprint arXiv:2011.08951, 2020 - arxiv.org

[4] https://paperswithcode.com/task/entity-linking

Compared to the performance dimensions of the previous approaches, several avenues have been identified as the key performance indexes (KPIs) for HelaNER 2.0. Firstly, word embedding through deep transfer learning has been tuned-up in terms of considering character level feature representations. Secondly, NE head word detection has been improved by enhancing the NE head word detection loss function. Also, multiple MLP and CRF layers have been adapted as another major improvement. Several experimentations have been conducted based on different hyper-parameter value adjustments. Thirdly, NE boundary detection has been further enhanced by introducing stack pointer networks which can be identified as the core scientific contribution for the domain of NE boundary detection. A constructive evaluation process has been followed taking the most prominent approaches which have been established as baselines. The overall evaluation procedure has been strengthened further through increasing the overall evaluation cycles. As per the evaluation outcome, it can be concluded that our novel avenues which have been introduced under HelaNER 2.0, have outperformed the existing benchmarks. As it's been mentioned earlier, even though this approach can be mentioned as a niche specialized solution for Sinhala domain, the generalized version of this would make a real impact for NE boundary detection in other domains.

Considering all these aspects, few potential avenues can be introduced as possible future enhancements under the whole process of detecting Sinhala related NE boundaries along with the considered NE type. It has been assumed as a particular head word would not be shared among more than one sentence nuggets under the section of dominant word detection considering MLP on top of CRF along with NE type detection in the constructive methodological approach. Therefore, sharing a given respective head word among several multiple sentence nuggets has been considered as a possible future enhancement under the overall methodological approach. Experimenting on the applicability of sharing a particular head word among multiple different sentence nuggets would be a major scientific research advancement of this entire domain.

ACKNOWLEDGMENT

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education Sri Lanka funded by the World Bank.

REFERENCES

[1] Jing Li, Aixin Sun and Yukun Ma, "Neural Named Entity Boundary Detection", IEEE Transactions on Knowledge and Data Engineering, 2015.

[2] Y.H.P.P Priyadarshana, L. Ranathunga, C.R.J Amalraj and I. Perera, "HelaNER: A Novel Approach for Nested Named Entity Boundary Detection," presented at the IEEE 19th International Conference on Smart Technologies (EUROCON), Lviv, Ukraine, Jul. 6-8, 2021.

[3] Abhishek Abhishek, Ashish Anand, and Amit Awekar, "Fine-Grained Entity Type Classification by Jointly Learning Representations and Label Embeddings", in Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017.

[4] Ioannis Partalas, Cédric Lopez, Nadia Derbas and Ruslan Kalitvianski, "Learning to Search for Recognizing Named Entities in Twitter", in Proc of the 2nd Workshop on Noisy User-generated Text (WNUT), 2016.

[5] Jing Li, Deheng Ye and Shuo Shang, "Adversarial Transfer for Named Entity Boundary Detection with Pointer Networks", Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}, 2019.

[6] Jason P.C. Chiu and Eric Nichols, "Named Entity Recognition with Bidirectional LSTM-CNNs", Transactions of the Association for Computational Linguistics, Volume 4, 2016.

[7] R. Collobert et al., "Natural language processing (almost) from scratch," The Journal of Machine Learning Research, 12:2493–2537, 2011.

[8] Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung and Guandong Xu, "A Boundary-aware Neural Model for Nested Named Entity Recognition", in Proc of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, "Neural Architectures for Named Entity Recognition", in Proc of NAACL-HCT 2016, pp. 260-270.

[10] Miguel Ballesteros, Chris Dyer, and Noah A. Smith, "Improved transition-based dependency parsing by modeling characters instead of words with LSTMs", In Proc of EMNLP, 2015.

[11] Mohammad Golam Sohrab and Makoto Miwa, "Deep Exhaustive Model for Nested Named Entity Recognition", in Proc of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2843-2849.

[12] Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan, "Recognizing Names in Biomedical Texts: A Machine Learning Approach, Bioinformatics," 20(7):, 2004, pp.1178–1190.

[13] Arzoo Katiyar and Claire Cardie, "Nested Named Entity Recognition Revisited," in Proc of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana. ACL, 2018, pp. 861–871.

[14] Nguyen Truong Son and Nguyen Le Minh, "Nested Named Entity Recognition Using Multilayer Recurrent Neural Networks," in Proc. of PACLING 2017, Sedona Hotel, Yangon, Myanmar, 2017, pp 16–18.

[15] Hongyu Lin, Yaojie Lu, Xianpei Han and Le Sun, "Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks," in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[16] Wang, B., and Lu, W, "Neural segmental hypergraphs for overlapping mention recognition," In Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 204–214.

[17] Wang, B., Lu, W., Wang, Y., and Jin, H, "A neural transition-based model for nested mention recognition," in Proc. of the Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1011–1017.

[18] Xu, M., Jiang, H., and Watcharawittayakul, S, "A local detection approach for named entity recognition and mention detection," in Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, volume 1, 2017, pp. 1237–1247.

[19] Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang and Fei Huang, "Boundary Enhanced Neural Span Classification for Nested Named Entity Recognition," The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020.

[20] Xiaoya Li et al., "A Unified MRC Framework for Named Entity Recognition", in Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[21] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer, "Zero-shot relation extraction via reading comprehension," in Proc. of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), 2017, pp. 333–342.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proc. of NAACL-HLT, 2019, pp. 4171–4186.

[23] Y. Chen et al, "A Boundary Regression Model for Nested Named Entity Recognition," citated at Computation and Language (cs.CL); Artificial Intelligence (cs.AI) as arXiv:2011.14330 [cs.CL], 2020.

[24] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos." In null, 2003, pp. 1470.

[25] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. "Multimodal intelligence: Representation learning, information fusion, and applications." arXiv preprint arXiv:1911.03977, 2019.

[26] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning," 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, doi: 10.1109/ICTer48817.2019.9023655, pp. 1-8.

[27] Gitari, N. D. et al., "A lexicon-based approach for hate speech detection," International Journal of Multimedia and Ubiquitous Engineering, 10(4),. doi: 10.14257/ijmue.2015.10.4.21, 2015, pp. 215–230.

[28] Cambria, E. et al., "SenticNet : A Publicly Available Semantic Resource for Opinion Mining," Artificial Intelligence, doi: 10.1038/leu.2012.122, 2010, pp. 14–18.

[29] Köffer, S. et al., "Discussing the Value of Automatic HateSpeech Detection in Online Debates," Multikonferenz Wirtschaftsinformatik, October 2018. doi: 10.1111/j.1365-2923.2008. 03277.x, pp. 83–94.

[30] Alfina, I. et al., "Hate speech detection in the Indonesian language: A dataset and preliminary study," International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017, 2018–October, doi: 10.1109/ICACSIS.2017.8355039, pp. 233–237.

[31] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. of the Second Workshop on Language in Social Media, LSM '12, 2012, pp. 19-26.

[32] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content," in Proc. of the 25th International Conference on World Wide Web, WWW'16, 2016, pp. 145-153.

[33] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proc. of NAACL-HLT, 2016, pp. 88-93.

[34] P. Badjatiya, S. Gupta, M. Gupta and V. Varma, "Deep learning for hate speech detection in tweets," in Proc. of the 26th International Conference on World Wide Web Companion, 2017, pp. 759-760.

[35] F. Vigna, A. Cimino, F. DellOrletta, M. Petrocchi and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in Proc. Of the First Italian Conference on Cybersecurity, 2017, pp. 86 – 95.

[36] Y. Chen et al., "A Boundary Assembling Method for Nested Biomedical Named Entity Recognition," in IEEE Access, vol. 8, pp. 214141-214152, 2020, doi: 10.1109/ACCESS.2020.3040182.

[37] Gwet, Kilem L, Large-Sample Variance of Fleiss Generalized Kappa, Educational and Psychological Measurement 81, 2021, pp. 781-790.

[38] A. Alcoforado et al., ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling, 15th International Conference on Computational Processing of Portuguese, 2017.

[39] Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown, Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages, In Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4693–4703.