2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3503-9796-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISCSLP57327.2022.10037867

BERT-based Chinese Medicine Named Entity Recognition Model Applied to Medication Reminder Dialogue System

Tsung-Hsien Yang[1], *Matus Pleva*[2], *Daniel Hládek*[2] *and Ming-Hsiang Su*[3]

[1]Chunghwa Telecom laboratories, Taoyuan, Taiwan
[2]Technical University of Košice, Košice, Slovakia
[3]Department of Data Science, Soochow University, Taipei, Taiwan

[1]yasamyang@cht.com.tw, {[2]matus.pleva, [2]daniel.hladek}@tuke.sk,
[3]huntfox.su@gmail.com

Abstract

The general public in Taiwan generally believes that traditional Chinese medicine (TCM) is mild and has no side effects, but they ignore the safety of traditional Chinese medicine. If the Chinese medicine name and disease name can be correctly identified in the human-machine dialogue, it can help the dialogue system to give correct medication reminders. In this study, a named entity recognition was constructed and applied to the identification of Chinese medicine names and disease names, and the results could be further used in the human-computer dialogue system to provide people with correct Chinese medicine medication reminders. First, this study uses a web crawler to organize network resources to become a TCM named entity corpus, collecting 1097 articles, 1412 disease names and 38714 TCM names. Then we use the Chinese medicine name and BIO labeling method to label each article. Finally, this study trains and evaluates BERT, ALBERT, RoBERTa and GPT2 with biLSTM and CRF. The experimental results show that the NER system of RoBERTa combined with biLSTM and CRF achieves the best system performance, where the Precision is 0.96, the Recall is 0.96 and the F1 -score is 0.96.

Index Terms: named entity recognition, human-computer dialogue, Chinese medicine

1. Introduction

Traditional Chinese medicine (TCM) is an alternative medical practice drawn from traditional medicine in China, and most of its treatments have no logical mechanism of action [1]. Taiwan has a diverse medical environment that integrates traditional Chinese and Western medicine. In addition, the effect of TCM on health care and chronic diseases has gradually been paid more and more attention by the public recently, and the use of TCM has been widely popularized [2-3]. The general public generally believes that TCM is mild and has no side effects. Therefore, people often go to TCM stores to buy TCM without a doctor's prescription, or listen to underground radio stations to exaggerate the efficacy and buy related TCM, while ignoring the safety of TCM [4]. According to statistics, about 88.2% of the public have had the experience of purchasing and taking TCM in the past year [2]. However, there are all kinds of illegal advertisements for drugs and substandard products in the market. The Chinese people's drug habits are incorrect, which often increases the burden on the liver and kidneys, or even misuse, mixed use, and interaction of Chinese and Western medicines, resulting in changes in the efficacy or toxicity of drugs. The health effects cannot be underestimated. Wang et al. [5] pointed out that silver-haired patients suffer from complex illnesses and chronic diseases. They also believe that among the TCM, silver-haired patients have the most serious health impact. Among the 520 cases of medication errors, usage and dosage errors were the highest, accounting for 48.5%. During the chat, how to provide people with the correct way of using TCM and medication information, so as to avoid the deterioration of health caused by wrong medication, is a topic worthy of research in the dialogue system.

Deep learning is currently the most forward-looking method in the field of machine learning, such as dialogue understanding [6], dialogue response generation [7], image classification [8], speech translation [9] or named entity recognition (NER) [10] and other successful cases. Machines, trained by deep learning, can figure out potential abstract rules on their own from a vast array of data, without the need for guidance from others. In view of the excellent recognition ability of deep learning in recent years, this research will deepen deep learning technology and apply it to Chinese medicine NER, to provide dialogue system and human interaction, and remind to avoid medication errors. At present, the corpus collection in the field of Chinese medicine is not as diverse and rich as the collection of Chinese or English corpus, which greatly increases the difficulty of Chinese medicine NER. If the problem of Chinese medicine NER can be overcome, it will be possible to correctly identify the name of the medicine mentioned by people in the human-machine dialogue, which will be of great help to the generation of subsequent medication reminder responses. For the field of TCM, there is currently no public knowledge base of TCM suitable for use. If the network resources can be organized by web crawlers to become a knowledge base of TCM, including various types of TCM and the prescriptions and efficacy of TCM prescriptions, it will be of great help to the dialogue system and question-and-answer system.

Nowadays, in the task of NER, the method of Recurrent Neural Network-based (RNN) [11-12] is mainly used as the model of sequence labeling, and supplemented by the word vector of the character level or there are other text features. Chiu and Nichols [11] used a bidirectional long short-term memory recurrent neural network (biLSTM) in the sequence labeling model. After the output of the recurrent neural network, an artificial network was used to determine which current should be labeled. Named entities, the character-level model used by the author uses a Convolution Neural Network

(CNN) with some character characteristics, such as whether the first letter is capitalized.

While Lample et al. [12] used a layer of Conditional Random Field (CRF) to judge the current labeling result. They use the characteristics of CRF to let the labeling result of the previous time point affect the current labeling, thereby improving the accuracy rate. On the other hand, the encoding at the character level is also changed, they use a pre-trained word vector with another bidirectional long and short-term recurrent neural network at the character level. Finally, they concatenate the original word vector with the character-level vector as the representative vector for the word. In addition to using RNN to judge named entities, Strubell et al. [13] used an Iterated Dilated Convolution Neural Network (IDCNN) to process named entities to achieve the purpose of acceleration. They solved the disadvantage that convolutional networks are not suitable for sequence labeling through a special structure of convolutional networks.

NER systems are usually evaluated by comparing their output to human annotations, which can be quantified by exact matches. NER involves Entity Boundaries and Entity Types, and through Exact-Match Evaluation, the named entity is considered correct only when both Entity Boundaries and Entity Types match the ground truth [14-16]. Since most NER systems involve multiple entity types, it is often necessary to evaluate the performance of all entity types. Two methods are commonly used for this: Macro-averaged F-score and Micro-averaged F-score. Macro-averaged F-scores are calculated independently for each entity type and then averaged. The Micro-averaged F-score aggregates entity contributions across all categories to calculate an average. The latter can be severely affected by the quality of the large-class recognized entities in the corpus.

2. Dataset Collection

In Taiwan, although people often use TCM to maintain their bodies, few people organize TCM dataset for training NER models. In this study, the crawler program was used to retrieve the traditional Chinese medicine data from the KingNet website [17] and CloudTCM website [18]. For the types of traditional Chinese medicines and prescriptions, according to their names, efficacy, usage, contraindications, etc., they are automatically organized into a TCM dataset suitable for the NER model. On the KingNet website, a total of 730 articles containing 678,846 words were collected; while on the CloudTCM website, a total of 367 articles containing 1,219,168 words were collected. Examples of Chinese medicine names and prescription names are shown in Table 1. Then this study uses inside-outside-beginning (IOB) for labeling, where we label TCM names as B-TMC and I-TMC, symptoms as B-SYMP and I-SYMP, and others as O. For example, Ginseng Root (人蔘) in this text, "The effect of ginseng soothe the nerves and nourish the mind is mainly manifested in the promotion of learning and memory. Appropriate amount of ginseng can soothe the nerves and sleep, relieve stress. (人蔘的安神益智功效主要表現在促進學習記憶方面。適量人蔘可安神舒眠，緩解壓力)" is labeled as “人 B-TMC” and “蔘 I-TMC”, and other words are labeled with "O".

Table 1: *Example of Chinese medicine name and Prescription name.*

Chinese medicine name	Dahuricae Angelica Root (白芷), Field Mint (薄荷), Ginseng Root (人蔘), etc.
Prescription name	Gui Zhi Tang (桂枝湯), WEN PI TANG (溫脾湯), QING PI YIN (清脾飲), etc.

3. Proposed method

3.1. Word Embedding

Language model pre-training has been shown to be effective in improving many natural language processing tasks [19]. These include sentence-level tasks and token-level tasks, such as natural language inference and NER. In this study, we use BERT as the word embedding model, where BERT is a natural language pre-training model released by the Google AI team in recent years.

BERT is a kind of natural language pre-training model, and it has a more innovative training method than other pre-training language models. It uses the Transformer's encoder bidirectional connection, and uses two unsupervised prediction tasks when training the bidirectional language model, namely the Masked language model (MLM) and the Next sentence prediction (NSP). The difference between two-way and one-way is mainly due to the different training directions of the word language representation. The one-way language model will train the language representation of each word according to the words on the left or right side of each word. For example, suppose you want to get the language representation of the word "bank" in the "I accessed the bank account" sentence. The one-way language model trains the word "bank" based on "I accessed the" on the left side of "bank" instead of "account" on the right side of "bank". For a bidirectional language model, the word embedding of the word is trained by considering both "I accessed the" and "account". A good language representation of words is very important for NLP tasks, and bidirectional language models can read bidirectional information, so the language representation of words will be better than unidirectional ones.

But the BERT author explained that the two-way development of the general language model may be because it can indirectly "see itself", so that the prediction of the word only needs to directly fill in its known context information, In order to solve the problems faced by the bidirectional language model, BERT proposes the training technique of adding a mask to the task of predicting words, masking 15% of the words in the input sentence, and being able to predict those masked words. The task example is shown in Table 2. How to choose the proportion of occlusion is a problem. First, if 100% of the selected words are occluded, it will cause the model to learn contextual language representation only by occlusion. For unoccluded words, it is difficult to learn a good language representation. Second, because the masking itself does not appear in the actual prediction stage, in order to force the model to focus on all words, a certain percentage of selected words are not masked, but are replaced with other words or remain unchanged. Through this training mechanism, the model can learn a better contextual language representation.

Table 2: *Example of masked language model prediction task.*

Input:
The man [MASK1] to [MASK2] store
Label:
[MASK1] = went; [MASK2] = store

Another innovation added to the BERT training strategy is the relationship between two sentences not considered by other language models, which is also an important feature for many natural language tasks. Therefore, in order for the model to learn the relationship between sentences, two sentences A and B will be given, and the model will determine whether B is the next sentence of A in the real corpus. The task example is in Table 3, predicting two Inter-sentence associations learn deep bidirectional contextual representations. In addition, BERT also has experiments showing that adding a linear layer to the output of the model can perform well in a variety of natural language processing tasks through fine-tuning, suitable for natural language tasks such as Sentiment classification or Question answering (QA). The input of BERT is token embeddings, segment embeddings and position embeddings, as shown in Figure 1, and two special symbols [CLS], [SEP]. [CLS] can be used for subsequent natural language classification tasks, and [SEP] is used to distinguish two sentences.

There are many improved models based on BERT, including ALBERT [20], RoBERTa [21] and GPT2 [22]. This study will evaluate these different models for the best performance of the NER system.

Table 3: *Example of next sentence prediction task.*

Input:
The man went to the store [SEP] he bought a gallon of milk
Label:
IsNext
Input:
The man went to the store [SEP] penguins are flightless birds
Label:
NotNext

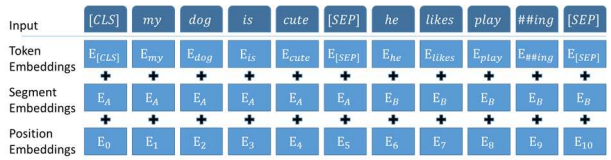

Figure 1: *Schematic diagram of BERT input representation.*

3.2. Bidirectional Long Short-Term Memory

The bidirectional long short-term memory (LSTM) is a special recurrent neural network (RNN). Different from the traditional RNN, LSTM uses three different gates to control the state of the Cell. These gates are Input Gate, Forget Gate and Output Gate. These three gates are represented by three green boxes in Figure 2, respectively.

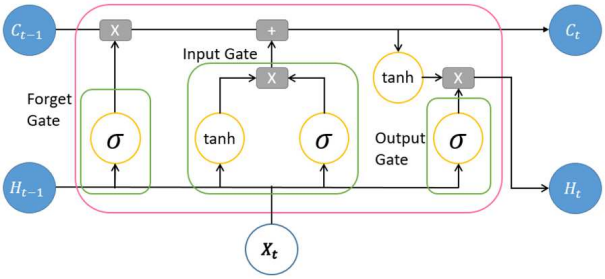

Figure 2: *Schematic diagram of LSTM model.*

The forgetting gate controls forgetting by using (1), where W_f and U_i represent the weight matrix to be multiplied by the output of the previous time point and the current input, h_{t-1} represents the output of the previous time point, and X_t represents the current input, b_f represents the bias vector, and the resulting f_t can decide which information should be forgotten. The input gate is divided into two small parts, one is called the candidate state vector $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ and the input gate vector $\boldsymbol{i_t}$, the operation method is (2) and (3), where W_c, W_i, U_c and U_i represent the weight matrix, and $\boldsymbol{b_c}$ and $\boldsymbol{b_i}$ represent the bias vector.

Using these two vectors $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$ and $\boldsymbol{i_t}$ to control how many Cell states are affected by the current input, the new Cell state $\boldsymbol{c_t}$ will be determined by f_t, $\boldsymbol{c_{t-1}}$, $\boldsymbol{i_t}$ and $\tilde{\boldsymbol{c}}_{\boldsymbol{t}}$, as shown in (4). The output gate is to control how many cell states will be output, as in (5), which is also determined by the current input $\boldsymbol{X_t}$ and the output h_{t-1} of the previous round. Finally, the output vector h_t of this round depends on the cell state $\boldsymbol{c_t}$ of this round and the vector $\boldsymbol{o_t}$ of the output gate, as shown in (6). Because of the mechanism of these gates, LSTM can remember long-term dependencies.

$$f_t = \sigma(W_f h_{t-1} + U_i X_t + \boldsymbol{b_f}) \quad (1)$$

$$\tilde{\boldsymbol{c}}_{\boldsymbol{t}} = tanh(W_c h_{t-1} + U_c X_t + \boldsymbol{b_c}) \quad (2)$$

$$\boldsymbol{i_t} = \sigma(W_i h_{t-1} + U_i X_t + \boldsymbol{b_i}) \quad (3)$$

$$\boldsymbol{c_t} = f_t \boldsymbol{c_{t-1}} + \boldsymbol{i_t} \tilde{\boldsymbol{c}}_{\boldsymbol{t}} \quad (4)$$

$$\boldsymbol{o_t} = \sigma(W_o \boldsymbol{h_{t-1}} + U_o \boldsymbol{X_t} + \boldsymbol{b_o}) \quad (5)$$

The output of most LSTMs will be one or more vectors, compared to the ground truth, get the error between the two, and then update the weights in the network by stochastic gradient descent or other optimization algorithms matrix. Due to the existence of several gates in the network, the possibility of gradient disappearance or explosion in the process of partial differentiation is greatly reduced, which is the advantage of LSTM over general RNN. The biLSTM is a bidirectional LSTM, and the architecture diagram is shown in Figure 3. BiLSTM is used to learn the dependencies of time series, and learn the key points that should be paid attention to in the sequence input by training the weights of the input gate, forgetting gate and output gate.

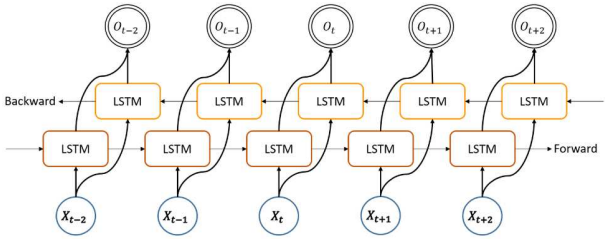

Figure 3: *Schematic diagram of biLSTM model.*

3.3. Condition Random Field

Each vector in the input sequence enters the biLSTM, and matches the hidden vector of the previous time point to judge the output of the current time point. This output will be used as a feature to enter the Condition Random Field (CRF) layer, and let CRF learn each weight of the feature function, as shown in Figure 4. The training method of CRF mainly has two steps. The first step is to generate feature functions from the training dataset, and initialize the weights corresponding to each feature function. The second step is to use maximum likelihood estimation, gradient descent and other methods to update the weights of each feature function until the weight changes converge. Using biLSTM and CRF as NER models, for CRF, the output sequence of biLSTM at each time point is the observation vector of CRF, and the labeled sequence of named entities can be compared with the predicted sequence predicted by CRF to calculate the gradient of the error function. Through the back-propagation algorithm, this error gradient is fed back to biLSTM and CRF, and the weights can be updated to minimize the error.

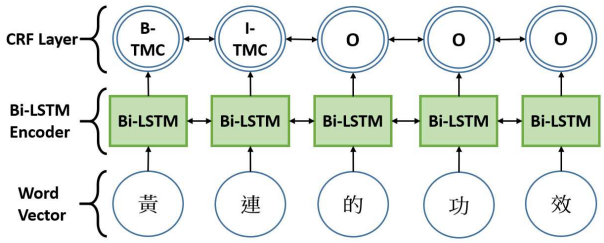

Figure 4: *Schematic diagram of NER model.*

4. Experimental Results and Discussion

This study proposes a BERT-biLSTM-CRF model for the NER task, which uses BERT to represent the information relationship of a single sentence, and then judges the labeled positions of named entities through the bidirectional LSTM and CRF models. This study compares the performance of BERT, ALBERT, RoBERTA and GPT2 combined with LSTM and CRF models to determine the architecture of the final NER system. The experimental results of the test set show that the Precision of BERT is 0.86, the Recall is 0.91 and the F1-score is 0.89; the Precision of ALBERT is 0.93, the Recall is 0.94 and the F1-score is 0.93; the Precision of RoBERTA is **0.96**, the Recall is **0.96** and F1 -score is **0.96**; Precision for GPT2 is 0.93, Recall is 0.92 and F1-score is 0.92.

Tables 4 and 5 show the experimental results of different models on TMC, SYMP tags, micro avg, macro avg and weighted avg, respectively. Experimental results show that the RoBERTa model outperforms BERT, ALBERT and GPT2 models. We believe that in the TCM corpus we collected, the RoBERTa model is more capable of extracting semantically relevant information, which leads to the best overall performance.

Table 4: *Evaluation of BERT and ALBERT.*

	BERT			ALBERT		
	P	**R**	**F1**	**P**	**R**	**F1**
SYMP	0.66	0.83	0.74	0.89	0.88	0.88
TMC	0.5	0.56	0.53	0.68	0.75	0.71
micro avg	0.86	0.91	0.89	0.93	0.94	0.93
macro avg	0.79	0.85	0.82	0.89	0.9	0.9
weighted avg	0.88	0.91	0.89	0.93	0.94	0.94

P: Precision; **R**: Recall; **F1**: F1-score

Table 5: *Example of next sentence prediction task.*

	RoBERTa			GPT2		
	P	**R**	**F1**	**P**	**R**	**F1**
SYMP	**0.92**	**0.91**	**0.92**	0.82	0.79	0.81
TMC	**0.8**	**0.79**	**0.79**	0.66	0.62	0.64
micro avg	**0.96**	**0.96**	**0.96**	0.93	0.92	0.92
macro avg	**0.93**	**0.92**	**0.93**	0.87	0.85	0.86
weighted avg	**0.96**	**0.96**	**0.96**	0.93	0.92	0.92

P: Precision; **R**: Recall; **F1**: F1-score

5. Conclusion and future work

In this study, NER was constructed and applied to the recognition of Chinese medicine names and disease names. The results can be further used in a human-computer dialogue system to provide people with correct Chinese medicine medication reminders. In addition, this study used web crawlers to organize network resources into a TCM named entity corpus, which included a total of 1097 articles, 1412 disease names, and 38714 TCM names. Then we annotated each article with the Chinese medicine name and BIO annotation method. Finally, the experimental results show that the RoBERTa combined biLSTM and CRF NER system achieves the best system performance, where Precision is 0.96, Recall is 0.96, and F1-score is 0.96.

In future work, we hope to obtain more Chinese medicine dialogue datasets so that we can train NER systems that are more suitable for a dialogue system. In addition, we also hope to add a self-attention mechanism to the NER system to improve system performance. Finally, we hope to expand the labels of NER, so that NER can identify more Chinese medicine-related named entities.

6. Acknowledgements

This work was supported by the Slovak Research and Development Agency (APVV) under the contract no. SK-TW-21-0002 (MOST-SRDA contract no. 111-2927-I-031-501), by Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and the Slovak Academy of Sciences under the research projects VEGA 1/0753/20 & VEGA 2/0165/21 and by National Science and Technology Council of Taiwan under the research project: Technologies to Support Response Generation for Multilingual Intelligent Agent.

7. References

[1] Wiki, "Traditional Chinese medicine," Retrieved October 11, 2022, from https://en.wikipedia.org/wiki/Traditional_Chinese_medicine.

[2] C.-L. Yan, L.-H. Huang, and M.-G. Yeh, "Pharmacists intervene to improve the safety of traditional Chinese medicine for the public (藥師介入提升民眾中醫藥就醫用藥安全)," *Journal of Pharmacy* (藥學雜誌), vol. 29, no. 3, 2013. Retrieved October 11, 2022, from https://www.taiwan-pharma.org.tw/magazine/116/030.pdf.

[3] M.-G. Yeh, "Stop, watch, listen, choose, and use specialty of traditional Chinese medicine medicine (中醫藥就醫用藥之停、看、聽、選、用專業)," *MOHW Nantou Hospital* (南投醫院). Retrieved October 12, 2020 from https://www.nant.mohw.gov.tw/public/ufile/c909c79547bd1528856c92f9a08e9361.pdf.

[4] R.-J. Yang, "Medication safety survey and knowledge research on medical care of "elderly" and "women" (「老人」及「婦女」醫學保健之用藥安全調查與知能研究)," *Annual Report of Traditional Chinese Medicine, Department of Ministry of Health and Welfare* (行政院衛生署中醫藥年報), vol. 1, no. 6, pp. 1-120, 2012. Retrieved October 12, 2020, from https://dep.mohw.gov.tw/DOCMAP/dl-11452-125fb77a-468d-45b4-b569-193af7e16ac4.html.

[5] H.-J. Wang, S.-L. Lin, P. Chang, Y.-W. Wang, and W.-C. Chen, "Analysis of 520 cases of misuse of traditional Chinese medicine in elderly patients (老年患者中藥用藥錯誤報告 520 例分析)," Adverse Drug Reactions Journal (藥物不良反應雜誌), vol. 17, no. 5, pp. 353, 2015. Retrieved October 12, 2020, from http://www.cadrj.com/CN/abstract/abstract3689.shtml.

[6] M.-H. Su, C.-H. Wu, K.-Y. Huang, and C.-K. Chen, "Attention-based dialog state tracking for conversational interview coaching," *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 6144-6148, 2018. DOI: https://doi.org/10.1109/ICASSP.2018.8461494.

[7] M.-H. Su, C.-H. Wu, and L.-Y. Chen, "Attention-Based Response Generation Using Parallel Double Q-Learning for Dialog Policy Decision in a Conversational System," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 131-143, 2019. DOI: https://doi.org/10.1109/TASLP.2019.2949687.

[8] A. Chhillar, S. Thakur, and A. Rana, "Survey of Plant Disease Detection Using Image Classification Techniques," *Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization* (*ICRITO*), pp. 1339-1344, 2020. DOI: https://doi.org/10.1109/ICRITO48877.2020.9197933.

[9] A. D. McCarthy, L. Puzon, and J. Pino, "SkinAugment: auto-encoding speaker conversions for automatic speech translation," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 7924-7928, 2020. DOI: https://doi.org/10.1109/lCASSP40776.2020.9053406.

[10] N. S. Pagad, and N. Pradeep, "Clinical named entity recognition methods: an overview," *Proceedings of International Conference on Innovative Computing and Communications*, pp. 151-165, 2022. DOI: https://doi.org/10.1007/978-981-16-2597-8_13.

[11] J. P. Chiu, and E. Nichols, E. "Named entity recognition with bidirectional LSTM-CNNs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357-370, 2016. DOI: https://doi.org/10.1162/tacl_a_00104.

[12] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260-270, 2016. DOI: https://doi.org/10.48550/arXiv.1603.01360.

[13] E. Strubell, P. Verga, D. Belanger, and A. McCallum, "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions," *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pp. 2670-2680, 2017. DOI: http://dx.doi.org/10.18653/v1/D17-1283.

[14] E. F. Tjong Kim Sang, and F. De Meulder, "Introduction to the CoNLL-2003 shared task: language-independent named entity recognition," *Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142-147, 2003.

[15] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang, "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes," *Proceedings of the Joint Conference on EMNLP and CoNLL-Shared Task*, pp. 1-40, 2012.

[16] Ö. Sevgili, A, Shelmanov, M. Arkhipov, A. Panchenko, and C. Biemann, "Neural entity linking: A survey of models based on deep learning," *Semantic Web*, vol. 13, no. 3, pp. 527-570, 2022. DOI: http://dx.doi.org/10.3233/SW-222986.

[17] KingNet website, Retrieved October 11, 2022, from https://www.kingnet.com.tw/tcm/.

[18] CloudTCM website, Retrieved October 11, 2022, from https://cloudtcm.com/.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. DOI: https://doi.org/10.48550/arXiv.1810.04805.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite Bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019. DOI: https://doi.org/10.48550/arXiv.1909.11942.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Roberta: A robustly optimized Bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. DOI: https://doi.org/10.48550/arXiv.1907.11692.

[22] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, and T. Nilsson, "GPT2: Empirical slant delay model for radio space geodetic techniques," *Geophysical research letters*, vol. 40, no. 6, pp. 1069-1073, 2013. DOI: https://doi.org/10.1002/grl.50288.