BuTTER: **B**idirec**T**ional LS**T**M for Food Named-**E**ntity **R**ecognition

1st Gjorgjina Cenikj
[1]*Computer Systems Department*
Jožef Stefan Institute
Ljubljana, Slovenia
[2]*Jožef Stefan International Postgraduate School*
Ljubljana, Slovenia
gjorgjina.cenikj@ijs.si

2nd Gorjan Popovski
[1]*Computer Systems Department*
Jožef Stefan Institute
Ljubljana, Slovenia
[2]*Jožef Stefan International Postgraduate School*
Ljubljana, Slovenia
gorjan.popovski@ijs.si

3rd Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

4th Barbara Koroušić Seljak
[1]*Computer Systems Department*
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

5th Tome Eftimov
Computer Systems Department
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

***Abstract*—In the modern era of big data, one of the biggest challenges is to find an efficient way of extracting information from unstructured data and structuring it in a form that can be interpreted and utilized by both humans and computers. In this paper, we focus on the domain of food and nutrition by introducing a Machine Learning (ML) based Named Entity Recognition (NER) method, which is a crucial step in extracting information from unstructured textual data. To the best of our knowledge, this is the first corpus-based food NER method that has been enabled by the recently published FoodBase corpus. The method is based on Bidirectional Long Short-Term Memory (BiLSTM) in conjunction with Conditional Random Fields (CRF) and Representation Learning (RL). Our experiments show that, despite the relatively small amount of annotated data, BuTTER is able to successfully identify food entities from raw text, with the best of the proposed models achieving an average macro F1 score of 0.946.**

***Index Terms*—information extraction, named-entity recognition, food, Bidirectional Long Short-Term Memory, Conditional Random Fields**

I. Introduction

Information extraction (IE) from biomedical scientific literature is an extremely important task in order to follow newly published knowledge in the form of unstructured text [1]–[3]. This extracted information can be used to improve decisions taken by clinical Decision Support Systems that are supported by predictive healthcare tools [4]. To this end, Question-Answering (QA) systems related to healthcare are extremely welcome [5], where underlying state-of-the-art models are based on deep neural networks [6] trained on self-created datasets of healthcare-related questions or knowledge graphs (KGs) describing the biomedical and health information [7]. KGs represent a collection of interlinked descriptions of entities – objects, events or concepts, by using semantic metadata, and they provide frameworks for data integration, unification, analytics and sharing [8], [9].

In most cases, KG creation is performed by literature mining where Natural Language Processing (NLP) methods are used to extract domain entities and the relations between them. To do this, IE methods tackling the Named Entity Recognition (NER) task can be applied to automatically detect and identify text phrases that represent domain entities [10]. To support the biomedical domain in this direction, a large amount of work has been already done [10]–[16], where NER methods were developed as a result of the existence of diverse biomedical vocabularies and standards together with the collection of a large amount of annotated biomedical data (e.g. in the domain of drugs, diseases and other treatments) from numerous biomedical NLP workshops [17]–[24].

However, unlike the large amount of work that has been done in the biomedical domain, the food and nutrition domain is still low-resourced. Nonetheless, moving to the era of personalized medicine, food is one of the environmental factors which affects the human health [25]. To understand the impact of food on our health, detecting food entities from scientific literature is crucial for several applications, such as food-drug interactions and food-disease interactions. Considering these demands, several NER methods have already been proposed in the food domain, some based on computational linguistics rules [26], [27] and others utilizing semantic information that is a part of various semantic resources [28], [29]. Each NER method is developed to fit some specific application. To this

This research was supported by the Slovenian Research Agency (research core grant number P2-0098 and project grant number PR-10465); and the European Union's Horizon 2020 research and innovation programme (FNS-Cloud, Food Nutrition Security) (grant agreement 863059);

end, there are still no ML-based NER methods, which are proven to be more robust than the rule-based methods, since an annotated corpus with food entities did not exist until recently.

At the end of 2019, to the best of our knowledge, the first annotated food corpus, known as FoodBase, was published [30]. It consists of food entities annotated with their corresponding Hansard food semantic tags [31], extracted from 1,000 recipes.

Having an annotated corpus with food entities facilitates the creation of Machine Learning (ML) based food NER methods. These extracted entities can be further combined with other health-related entities in order to create a more complex heterogeneous KG. In this paper, we present the first ML-based food NER method that utilizes the annotated food concepts available in the FoodBase corpus.

In the remainder of the paper, we first present a general overview of the types of NER methods, followed by the proposed methodology for food NER. Then the results are discussed, finalizing with the conclusions of the paper.

II. Related Work

In this section, we provide an overview of various NER methodologies that can be used for IE from different domains, followed by a summary of word-based representations that can be used to represent textual data before applying ML methods.

A. NER Methods

A comprehensive survey of different NER methods is presented in [32]. The first developed NER methods are known as *dictionary-based* [33], which can extract only the entities that are mentioned in a selected domain dictionary. They have further been improved by proposing *rule-based* [26], [27], [34] methods that use dictionaries in a combination with rules that describe the characteristics of the domain entities. Significant weaknesses of such approaches are the limited number of entities that can be extracted (i.e. only those present in the dictionary), as well as the time needed to manually create the domain specific rules. However, they can still provide promising results when annotated corpora required for training ML-based method do not exist. Hence, such NER methods are especially used for low-resourced domains.

More robust results can be achieved by training *ML-based* methods [35], [36], which learn a supervised ML model by using an annotated corpus. This is also known as the sequence tagging task, where the most commonly used methods are Hidden Markov Models (HMMs) [37], Maximum entropy Markov models (MEMMs) [38], and Conditional Random Fields (CRF) [39]. Recently, state-of-the-art results in sequence tagging have been achieved by utilizing deep neural networks models, such as: long short-term memory (LSTM) networks, bidirectional LSTM networks (BI-LSTM), LSTM networks with a CRF layer (LSTM-CRF), and bidirectional LSTM networks with a CRF layer (BILSTM-CRF) [40]–[42]. Their only weakness is that they require a large amount of annotated data. However, when such data is unavailable, they can be combined with *active learning*, where semi-supervised learning is used to train a model that does not require a large annotated corpus, but instead interacts with the user to query for new annotations that are further used for iteratively improving the model [43].

B. Representation Learning

Three word-based representation learning methods are explained below, which will further be used in our proposed methodology.

1) GloVe: GloVe is an unsupervised algorithm for distributed word representation which is based on the assumption that the ratios of the co-occurrence probabilities of words can be used to obtain representative word mappings in the form of a vector space where the distance between words is based on their semantic similarity. The goal is to learn vector representations of words such that their dot product is equal to the logarithm of the probability of the words' co-occurrence, i.e. the idea is to associate the ratios of the co-occurrence probabilities with the difference of the vector representations in the vector space.

2) Word2Vec: Word2Vec is a two-layer neural network that produces vectorized representations of words from a given input corpus in such a manner that words with similar contexts are grouped close to one another in the vector space. It generates a vocabulary where each word is mapped to a corresponding vector representation, which can be used to find relationships between words or represent words in a downstream ML task. Depending on the training objective and the model architecture, two general variants are distinguished: continuous bag-of-words (CBOW) and continuous skip-gram. The CBOW model aims to predict target words from the context they appear in, while the skip-gram model has the opposite objective: predicting the context of a given target word.

3) FastText: FastText is based on the premise that the morphological structure of words carries important information about their meaning. It extends the idea of the continuous skip-gram model by additionally exploiting subword information to construct word embeddings. The representation of each word is obtained as a sum of the vectors of the character n-grams it is composed of. By virtue of this, FastText can generate embeddings for words that are not present in its vocabulary, provided that at least one of the character n-grams was present in its training data.

III. Methodology

Our proposed methodology is based on the evaluation of the three aforementioned pre-trained representation models in combination with a Neural Network architecture based on BiLSTM and CRF. The first step involves representing the textual data by using a representation method and then using it to train a food NER method utilizing BiLSTMs and CRFs. When representing the textual data we have also explored two preprocessing steps: lemmatization and handling of out-of-vocabulary (OOV) words.

A. Pre-processing

Due to the use of pre-trained word representation models and the absence of some of the words in our dataset in the vocabularies of Word2Vec, GloVe, and FastText, additional preprocessing was applied in order to get representations for the OOV words. As the first pre-processing step, each token was converted to lowercase. Due to the fact that most of the OOV words were either numbers or compound words, the num2words library was used to transform arabic numerals into their appropriate textual equivalent. Then, the OOV words were split using the punctuation characters they contain, and a representation for the whole word was generated using the average of the embeddings of the obtained substrings. In the cases when the vocabulary did not contain any of the substrings, zero vectors were used instead.

In the case when custom word embeddings are trained instead of using pre-trained representation models, we also examine how lemmatization impacts the model performance.

B. Classification Models

We treat the task of identifying food entities in raw text as a classification problem where the classes are the tags from the Inside-outside-beginning (IOB) tagging scheme, i.e. the goal is to determine whether each token in the text is outside (O), inside (I) or at the beginning (B) of a food entity. Since the technical implementation requires all sentences to be of equal length, we also include an additional padding class (P). This is due to the nature of LSTM models, where the input size is always fixed.

As the joint use of BiLSTMs and CRFs has been shown to outperform other recurrent neural networks in sequence tagging tasks [44], we use this architecture in order to exploit the benefits of BiLSTMs for capturing the sequential dependencies of the input tokens in both directions, left-to-right and right-to-left, and the CRF layer for capturing the relationships amongst the labels, allowing an optimal joint prediction of all the labels in the sentence.

Additionally, we examine the impact of supplementing the word embeddings generated by each representation model with character embeddings. The motivation behind this is to generate a richer contextual embedding for each word, which takes into account character-level features such as prefix and postfix, and can aid in the representation of rarely occurring words whose word embeddings might not be trained well.

Based on whether character embeddings are included in the representation of each word, we distinguish two neural network models: BiLSTM-CRF and Char-BiLSTM-CRF, which differ only in their input layers.

The architecture of the BiLSTM-CRF model is presented in Fig. 1. It has a single input layer which gets the ids of the words in the vocabulary, followed by an embedding layer which represents each word using the embeddings generated by one of the representation models or by training to learn custom weights to represent each word. We refer to the model which learns custom weights for the word embedding layer as a lexical model. The obtained word representations are delegated to a BiLSTM, a fully-connected, time-distributed layer and a CRF layer, which outputs the probability of the word belonging to each of the classes.

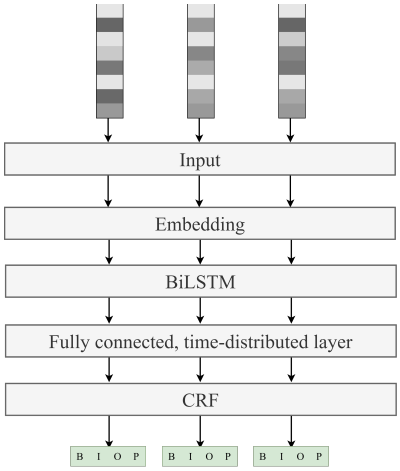


Fig. 1: Architecture of the BiLSTM-CRF model

The Char-BiLSTM-CRF contains an additional stack of input and embedding layers for creating the character embeddings. This stack of layers receives as input the indices of the characters of each word in the character vocabulary, which contains all of the unique characters in the dataset. The weights of the time-distributed embedding layer for generating the character embeddings are trained together with the rest of the layers in the model, on the NER task. The character embeddings are then processed by a time-distributed LSTM layer, and concatenated with the word embeddings. The concatenated embeddings are passed to the same sequence of downstream layers as in the BiLSTM-CRF model, i.e. a BiLSTM, a fully-connected, time-distributed layer and a CRF layer. The complete architecture of the Char-BiLSTM-CRF model is depicted in Fig. 2.

IV. Experimental Design

The pre-trained language representation models GloVe, Word2Vec, and FastText were borrowed from the gensim-data[1] repository. The FastText model was trained on the Wikipedia 2017, UMBC web base corpus and the statmt news dataset, the GloVe model was trained on the Wikipedia 2014 and Gigaword 5 corpora, while the Word2Vec model was trained on the Google News dataset. For the sake of fair model comparison, all of the vectors produced by these models were of dimension 300. The datasets on which these methods have been trained are of substantial size, making bias unlikely to have a notable effect and the texts in the datasets concern various topics.

The custom-trained word embeddings used in the lexical model are of dimension 300, while the character-based representations of words were set to be of dimension 20. The choice of 300 is to be consistent with the pre-trained word

[1] https://github.com/RaRe-Technologies/gensim-data

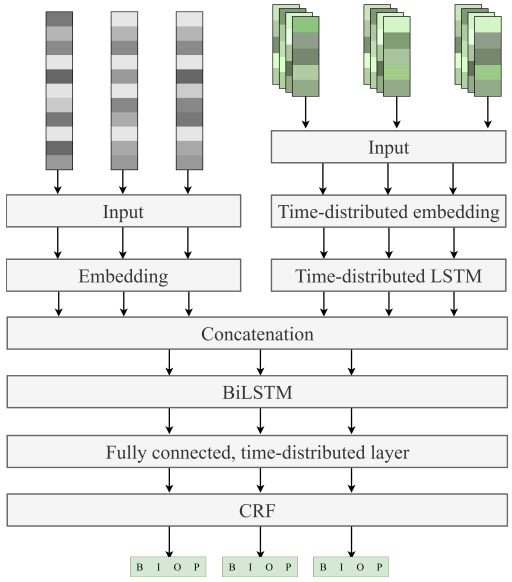


Fig. 2: Architecture of the Char-BiLSTM-CRF model

representation models, while the choice of 20 for the character-based representations is such because of the nature of the set of characters and its low cardinality.

The hyperparameters that define the architecture of the model were chosen according to a paper on utilizing BiLSTM-CRF [45]. Other hyperparameters which were not defined in the architecture and regard technical details were chosen to be the same for each model. For example, each model used the same batch size, weight initialization, etc. Finally, all stochasticity was eliminated by fixing the random seeds as well as the hash seeds. With this, each run of the same model produced identical results.

Before being passed as inputs to the model, the sentences are converted to arrays of integers in the range $[0, 3125]$, where 3125 is the length of the vocabulary, i.e. the number of unique words in the dataset. Since the word embedding layer requires arrays of equal length to be given as inputs, all sentence representations are padded up in order to reach the length of 50, since the longest sentence in our dataset was of length 45.

Similarly, for the generation of character-based embeddings, all words are represented as arrays of character indices in the character vocabulary, and are padded up to contain 18 elements, which is in fact the length of the longest word in the dataset.

The hidden dimensions of the fully connected time-distributed layer and the LSTMs in the BiLSTM layer are set to 50. The models are optimized using the rmsprop optimizer with the smoothing term set to 10^{-7}, the gradient moving average decay factor set to 0.9 and the learning rate set to 0.001.

Each model was trained until the improvement in validation loss of 5 consecutive epochs did not surpass $5 \cdot 10^{-3}$. For each fold, 10% of the training set was taken out and used as a validation set. This is a common technique used to terminate training in order to reduce overfitting and is referred to as early stopping.

Finally, the total number of trained models is 16. This is the result of combining 3 word representation models with 2 methods for pre-processing totaling 6 models, with two additional models based on lexical features, one lemmatized and one not. This means that there are 8 representation models, each trained using the aforementioned BiLSTM-CRF and Char-BiLSTM-CRF architectures, to produce a total of 16 models.

The full source code can be found on GitHub. [2]

V. Evaluation

The evaluation of the proposed methodology has been done using the FoodBase corpus and stratified 5-fold cross validation. The folds were generated using stratified sampling since the FoodBase corpus consists of 5 classes of recipes. The reported evaluation metric is the commonly used F1 Score, which in this case is computed by using macro averaging, meaning that the F1 Score is computed for each class separately (I, O, B) and then averaged. The rationale behind this is that the dataset is heavily unbalanced (The O class appears much more compared to B and I due to the nature of the task, i.e. NER). Then, for each model we report the average of the macro averaged F1 Score for each fold. The final F1 Score lies in the range $[0, 1]$, where a theoretically ideal NER method would have an F1 Score of 1.

Next, a brief explanation of the corpus is presented, followed by the experimental results and discussion.

A. FoodBase corpus

In 2019, the first annotated food corpus known as FoodBase became available [30]. It consists of 1,000 recipes, from which food entities are extracted and annotated with the Hansard food semantic tags [31]. Its existence opens new directions in food NER, by allowing development of ML-based NER.

B. Results and discussion

After performing 5-fold Cross Validation we obtain five different macro F1 Scores for each model, which are then averaged. These averaged F1 Scores are presented in Table I, while their convergence time, i.e. average number of epoch needed to satisfy the early stopping criterion, is presented in Table II.

From the obtained results, it is obvious that the lexical model outperforms the other models both in terms of the achieved F1 macro score and the number of epochs needed for convergence, regardless of whether additional character embeddings are included. The model with the highest macro F1 score is the lemmatized lexical model using the BiLSTM-CRF architecture, achieving value of 0.94640. Regarding the Char-BiLSTM-CRF architecture, the un-lemmatized lexical model is once again with the highest macro F1 score, obtaining a value of 0.94603.

[2] https://github.com/gjorgjinac/butter

Lemmatization has a positive impact on the performance of the BiLSTM-CRF model, but a slightly negative impact to the Char-BiLSTM-CRF model. This may be attributed to the fact that lemmatization would decrease the amount of postfix information that is given to the model through the character embeddings.

Out of the models that use pretrained word representations, GloVe achieves the highest F1 macro scores (0.91573 and 0.93123 for the BiLSTM-CRF and Char-BiLSTM-CRF models respectively), while Word2Vec converges in the lowest amount of epochs. While the difference in the F1 macro scores of the Word2Vec and FastText models is small, the convergence of the FastText model requires a considerably larger number of epochs.

The proposed handling of OOV words also results in a small, but consistent increase of the F1 macro scores of both the BiLSTM-CRF and the Char-BiLSTM-CRF model. By applying the additional preprocessing step, the percentage of words being represented as zero vectors was reduced from 11.35% to 1.67% for the Word2Vec model, from 4.02% to 0.75% for the FastText model, and from 6.01% to 1.14% for the GloVe model. It is therefore unsurprising that this preprocessing has the greatest impact on the Word2Vec model, where the difference of the F1 macro scores of the model where the OOV words are handled and the model where they are simply represented as zero vectors, is roughly twice as high as that of the GloVe and FastText models.

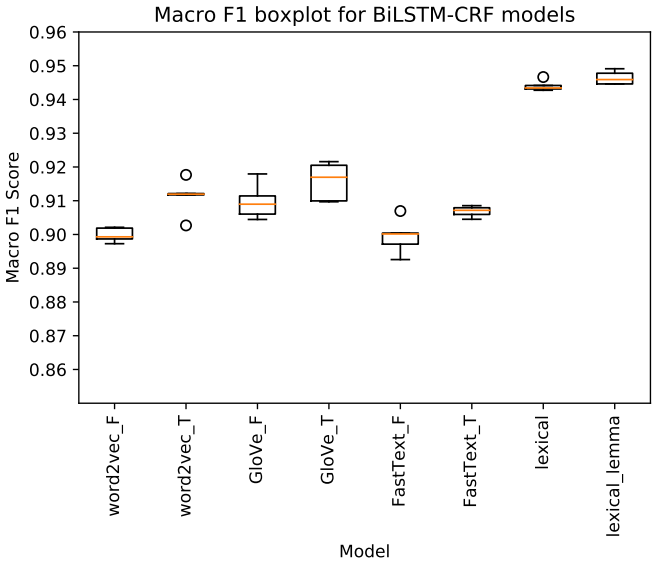
The inclusion of character embeddings in the Char-BiLSTM-CRF model brings in a consistent improvement over the BiLSTM-CRF model, observed through the increase in the F1 macro scores and the decrease in the number of epochs needed for convergence.

Finally, it is important to note that all the evaluated models achieve a macro F1 score of at least 0.89.

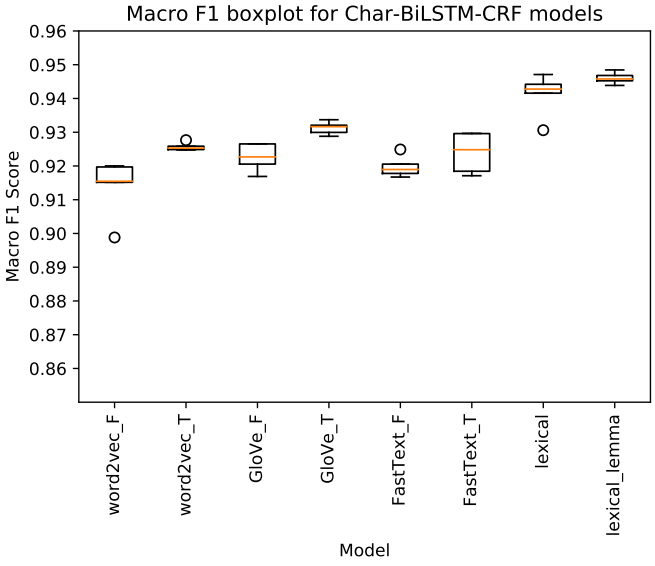
In Figure 3, boxplots of the macro F1 score for all folds are shown. From it we can note that the two most robust models with the BiLSTM-CRF architecture are word2vec with OOV handling, and the un-lemmatized lexical model. Similarly, for the Char-BiLSTM-CRF architecture, the word2vec with OOV handling, as well as the lemmatized lexical model, are the most robust.

Fig. 3: Macro F1 boxplots. The suffix *_F* denotes that the OOV words are not handled, while *_T* denotes that the missing values are handled.

(a) Boxplot for the models using the BiLSTM-CRF architecture



(b) Boxplot for the models using the Char-BiLSTM-CRF architecture



VI. Conclusions

In this paper, we present BuTTER, a Machine Learning (ML) based Named-Entity Recognition (NER) method trained by using recently published FoodBase corpus, that leverages Bidirectional Long Short-Term Memory (BiLSTM) and Conditional Random Fields (CRF) for the identification of food entities from raw text. We explore and evaluate the use of the pretrained word representation models GloVe, Word2Vec, FastText, as well as training our own embeddings. We also consider the performance impact of complementing them with different preprocessing techniques and character embeddings. The experimental results indicate that the additional use of character embeddings is beneficial for the model's performance, and that the lexical model which trains custom embeddings on our own dataset gives the best results, with an F1 Score of 0.94640.

The fact that BuTTER is, to the best of our knowledge, the first corpus-based NER method and the obtained results are very promising, once again emphasizes the importance of generating new resources which will enable the development of similar methods that are incredibly essential, and yet, are lacking in the food domain.

Our future efforts will be dedicated towards the application of BuTTER for the generation of KG and unification of concepts from the food and biomedical domains, which will

TABLE I: Macro F1 scores for each model, using 300D word embeddings. Each macro F1 score is obtained by using stratified K-fold Cross-Validation (k=5). (underlined values are best per subtable, while the bold value is the best from the whole table.)

(a) Macro F1 scores referring to the BiLSTM-CRF model.

	OOV words handled	
model	$\perp$	$\top$
GloVe	0.90976	0.91573
Word2Vec	0.89985	0.911225
FastText	0.89944	0.906811
	lemmatized	
	$\perp$	$\top$
lexical	0.94401	**0.94640**

(b) Macro F1 scores referring to the Char-BiLSTM-CRF model.

	OOV words handled	
model	$\perp$	$\top$
GloVe	0.92264	0.93123
Word2Vec	0.91383	0.92568
FastText	0.91978	0.92392
	lemmatized	
	$\perp$	$\top$
lexical	0.94603	0.94125

TABLE II: Average number of epochs required to reach the early stopping criteria for each model. The reported numbers are averages over the number of epochs across the 5 folds. (The underlined values represent the best results per subtable, while the bold value is the best from the whole table.)

(a) Average number of epochs required for convergence of the BiLSTM-CRF model.

	OOV words handled	
model	$\perp$	$\top$
GloVe	155.2	158.4
Word2Vec	146.0	147.4
FastText	233.0	199.8
	lemmatized	
	$\perp$	$\top$
lexical	130.4	137.8

(b) Average number of epochs required for convergence of the Char-BiLSTM-CRF model.

	OOV words handled	
model	$\perp$	$\top$
GloVe	125.2	149.0
Word2Vec	131.2	140.8
FastText	185.4	186.0
	lemmatized	
	$\perp$	$\top$
lexical	**118.8**	134.4

aid the analysis of the effects of food on human health. Additionally, more state-of-the-art and sophisticated word and character Representation Learning (RL) methods can be utilized to further improve the results.

ACKNOWLEDGMENT

This work is supported by AdFutura. Another financial support that this project received was from the Slovenian Research Agency (research core fundings No. P2-0098, PR-10465), as well as the European Union's Horizon 2020 research and innovation programme (FNS-Cloud, Food Nutrition Security) (grant agreement 863059). The authors acknowledge the financial support from all three funding parties.

REFERENCES

[1] C. Sivaparthipan, N. Karthikeyan, and S. Karthik, "Designing statistical assessment healthcare information system for diabetics analysis using big data," *Multimedia Tools and Applications*, vol. 79, no. 13, pp. 8431–8444, 2020.

[2] K. Adnan, R. Akbar, S. W. Khor, and A. B. A. Ali, "Role and challenges of unstructured big data in healthcare," in *Data Management, Analytics and Innovation*. Springer, 2020, pp. 301–323.

[3] E. Ogbuju and G. Obunadike, "Information extraction from electronic medical records using natural language processing techniques," *Journal of Applied Sciences and Environmental Management*, vol. 24, no. 6, pp. 1027–1033, 2020.

[4] S. Nijjer, K. Saurabh, and S. Raj, "Predictive big data analytics in healthcare," in *Big Data Analytics and Intelligence: A Perspective for Health Care*. Emerald Publishing Limited, 2020.

[5] F. Luo, X. Wang, Q. Wu, J. Liang, X. Qiu, and Z. Bao, "Hqadeephelper: A deep learning system for healthcare question answering," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 194–197.

[6] A. Nentidis, A. Krithara, K. Bougiatiotis, M. Krallinger, C. Rodriguez-Penagos, M. Villegas, and G. Paliouras, "Overview of bioasq 2020: The eighth bioasq challenge on large-scale biomedical semantic indexing and question answering," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2020, pp. 194–214.

[7] L. LiuQiao, L. DuanHong *et al.*, "Knowledge graph construction techniques," *Journal of computer research and development*, vol. 53, no. 3, p. 582, 2016.

[8] P. Ernst, A. Siu, and G. Weikum, "Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences," *BMC bioinformatics*, vol. 16, no. 1, p. 157, 2015.

[9] S. Sang, Z. Yang, X. Liu, L. Wang, H. Lin, J. Wang, and M. Dumontier, "Gredel: A knowledge graph embedding based method for drug discovery from biomedical literatures," *IEEE Access*, vol. 7, pp. 8404–8415, 2018.

[10] T. H. Dang, H.-Q. Le, T. M. Nguyen, and S. T. Vu, "D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information," *Bioinformatics*, vol. 34, no. 20, pp. 3539–3546, 2018.

[11] J. M. Giorgi and G. D. Bader, "Transfer learning for biomedical named entity recognition with neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4087–4094, 2018.

[12] W. Yoon, C. H. So, J. Lee, and J. Kang, "Collabonet: collaboration of deep neural networks for biomedical named entity recognition," *BMC bioinformatics*, vol. 20, no. 10, p. 249, 2019.

[13] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type biomedical named entity recognition with deep multi-task learning," *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2019.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] H. Zhou, S. Ning, Z. Liu, C. Lang, Z. Liu, and B. Lei, "Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes," *BMC bioinformatics*, vol. 21, no. 1, p. 35, 2020.

[16] M. Cho, J. Ha, C. Park, and S. Park, "Combinatorial feature embedding based on cnn and lstm for biomedical named entity recognition," *Journal of Biomedical Informatics*, vol. 103, p. 103381, 2020.

[17] C. N. Arighi, B. Carterette, K. B. Cohen, M. Krallinger, W. J. Wilbur, P. Fey, R. Dodson, L. Cooper, C. E. Van Slyke, W. Dahdul *et al.*, "An overview of the biocreative 2012 workshop track iii: interactive text mining task," *Database*, vol. 2013, 2013.

[18] Y. Mao, K. Van Auken, D. Li, C. N. Arighi, P. McQuilton, G. T. Hayman, S. Tweedie, M. L. Schaeffer, S. J. Laulederkind, S.-J. Wang *et al.*, "Overview of the gene ontology task at biocreative iv," *Database*, vol. 2014, 2014.

[19] C.-H. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu, "Overview of the biocreative v chemical disease relation (cdr) task," in *Proceedings of the fifth BioCreative challenge evaluation workshop*, vol. 14, 2015.

[20] S. Pyysalo, T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou, "Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013," *BMC bioinformatics*, vol. 16, no. S10, p. S2, 2015.

[21] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1," *Journal of biomedical informatics*, vol. 58, pp. S11–S19, 2015.

[22] L. Deléger, R. Bossy, E. Chaix, M. Ba, A. Ferré, P. Bessieres, and C. Nédellec, "Overview of the bacteria biotope task at bionlp shared task 2016," in *Proceedings of the 4th BioNLP shared task workshop*, 2016, pp. 12–22.

[23] K. B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, "Bionlp 2017," in *BioNLP 2017*, 2017.

[24] Y. Wang, K. Zhou, M. Gachloo, and J. Xia, "An overview of the active gene annotation corpus and the bionlp ost 2019 agac track tasks," in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 62–71.

[25] F. Johan and G. Owen. (2020) Exponential roadmap. [Online]. Available: https://exponentialroadmap.org/wp-content/uploads/2019/09/Exponential-Roadmap-1.5-September-19-2019.pdf

[26] T. Eftimov, B. Koroušić Seljak, and P. Korošec, "A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations," *PloS One*, vol. 12, no. 6, p. e0179488, 2017.

[27] G. Popovski, S. Kochev, B. K. Seljak, and T. Eftimov, "Foodie: a rule-based named-entity recognition method for food information extraction," in *In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019, pp. 915–922.

[28] C. Jonquet, N. Shah, C. Youn, C. Callendar, M.-A. Storey, and M. Musen, "Ncbo annotator: semantic annotation of biomedical data," in *International Semantic Web Conference, Poster and Demo session*, vol. 110, 2009.

[29] G. Popovski, B. K. Seljak, and T. Eftimov, "A survey of named-entity recognition methods for food information extraction," *IEEE Access*, vol. 8, pp. 31 586–31 594, 2020.

[30] ——, "Foodbase corpus: a new resource of annotated food entities," *Database*, vol. 2019, 2019.

[31] M. Alexander and J. Anderson, "The hansard corpus, 1803-2003," 2012.

[32] V. Yadav and S. Bethard, "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[33] X. Zhou, X. Zhang, and X. Hu, "Maxmatcher: Biological concept extraction using approximate dictionary lookup," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2006, pp. 1145–1149.

[34] D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck, "Prominer: rule-based protein and gene entity recognition," *BMC bioinformatics*, vol. 6, no. 1, p. S14, 2005.

[35] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou, "Using text mining techniques to extract phenotypic information from the phenochf corpus," *BMC medical informatics and decision making*, vol. 15, no. 2, p. 1, 2015.

[36] R. Leaman, C.-H. Wei, C. Zou, and Z. Lu, "Mining patents with tmchem, gnormplus and an ensemble of open systems," in *Proce. The fifth BioCreative challenge evaluation workshop*, 2015, pp. 140–146.

[37] C. Harshitha and N. Sunitha, "Topic identification for semantic grouping based on hidden markov model," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, 2020, pp. 932–937.

[38] F. Alam and M. A. Islam, "A proposed model for bengali named entity recognition using maximum entropy markov model incorporated with rich linguistic feature set," in *Proceedings of the International Conference on Computing Advancements*, 2020, pp. 1–6.

[39] X. Sun, S. Sun, M. Yin, and H. Yang, "Hybrid neural conditional random fields for multi-view sequence labeling," *Knowledge-Based Systems*, vol. 189, p. 105151, 2020.

[40] K. Xu, Z. Zhou, T. Hao, and W. Liu, "A bidirectional lstm and conditional random fields approach to medical named entity recognition," in *International Conference on Advanced Intelligent Systems and Informatics*. Springer, 2017, pp. 355–365.

[41] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on rnn and crf models for de-identification of medical free text," *Journal of Big Data*, vol. 7, no. 1, pp. 1–22, 2020.

[42] C. Ronran and S. Lee, "Effect of character and word features in bidirectional lstm-crf for ner," in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2020, pp. 613–616.

[43] G. Yu, Y. Yang, X. Wang, H. Zhen, G. He, Z. Li, Y. Zhao, Q. Shu, and L. Shu, "Adversarial active learning for the identification of medical concepts and annotation inconsistency," *Journal of Biomedical Informatics*, vol. 108, p. 103481, 2020.

[44] Y. Jiang, Z. Liu, and S. Ponnusamy, "Univalent harmonic mappings and lift to the minimal surfaces," 2015.

[45] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.