VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

VLSP 2021 - NER Challenge: Named Entity Recognition for Vietnamese

Ha My Linh[1,*], Do Duy Dao[1], Nguyen Thi Minh Huyen[1], Ngo The Quyen[1], Doan Xuan Dung[2]

[1]*VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*
[2]*Viettel Cyberspace Center, Viettel Group, Hanoi, Vietnam*

Received 14 May 2021
Revised 27 August 2021; Accepted 1 November 2021

Abstract: Named entities (NE) are phrases that contain the names of persons, organizations, locations, times, quantities, email, phone number, etc., in a document. Named Entity Recognition (NER) is a fundamental task that is useful in many applications, especially in information extraction and question answering. Shared tasks on NER provides several reference datasets in many languages. In the 2016 and 2018 editions of the VLSP workshop series, reference NER datasets have been published with only three main entity categories: person, organization and location. At the VLSP 2021 workshop, another challenge on NER is organized for dealing with an extended set of 14 main entity types and 26 sub-entity types. This paper describes the published datasets and the evaluated systems in the framework of the VLSP 2021 evaluation campaign.

Keywords: Named Entity Recognition for Vietnamese.

1. Introduction

Named entities (NE) are phrases that contain the names of persons, organizations, locations, times, quantities, email, phone number, etc., in a document. Below is an example with three common entity types.

- Anh [PER Thanh] là cán bộ
- [ORGANIZATION Uỷ ban nhân dân
- [LOCATION thành phố Hà Nội]

This sentence has three named entities: "Thanh" is a person, "Uỷ ban nhân dân thành phố Hà Nội" is an organization and "thành phố Hà Nội" is a location.

Named Entity Recognition (NER) is a fundamental task that is useful in many applications, especially in information extraction and question answering. This task has attracted much attention since the 1990s.

In 1995, the 6th Message Understanding Conference (MUC) started a shared task for evaluating English NER systems [1]. Later, NER

* Corresponding author.
E-mail address: halinh.hus@gmail.com

https://doi.org/10.25073/2588-1086/vnucsce.362

systems for Dutch and Turkish were also evaluated in CoNLL 2002 [2] and CoNLL 2003 [3] shared tasks. In these evaluation tasks, four named entities were considered, consisting of names of persons, organizations, locations, and names of some miscellaneous entities. While the task is not new, researchers continue to work on it for dealing with extensible sets of entity types and in different fields.For example, in the tourism field, interesting entities should be hotel or resort names, locations, resort activities, etc. Or in the medical field, important entities may be disease names, drug names, clinical symptoms, etc. NER datasets covering several domains have been developed for many languages such as English (CoNLL2003 [3] (news), WNUT2017 [4] (social media), i2b2 (medical), etc.), German (GermEval2014 [5], [6], etc.), Dutch (CoNLL2002 [2], etc.). . . In addition, several multilingual NER datasets have also been built such as WikiNEuRal [7], WikiNER [8], MEANTIME corpus [9], etc.

Recently, the SHINRA project has attracted many teams to participate in the named entity classification task: classify Wikipedia pages in 30 languages into about 220 fine-grained named entity categories, with a huge training dataset (more than 100K pages).

One of the missions of the Association for Vietnamese Language and Speech Processing (VLSP) is to provide the VLSP community with public reference datasets for natural language processing (NLP) tasks. Regarding the NER task, for Vietnamese language, VLSP 2016 and VLSP 2018 workshops [10] have provided two NER datasets with three main entities (person, organization, and location) with nested levels. In 2021, a reference COVID-19 NER dataset [11] is published with 10 categories of non-nested entities (patient id, person name, age, gender, occupation, location, organization, symptom and disease, transportation, and date). These datasets constitute essential resources for developing and evaluating applications involving the NER task, but the need for NER datasets of richer set of entities is still important.

In this 2021 edition of the VLSP workshop series, another NER challenge has been organized. Participants are provided with a new NER dataset containing an extended set of entities (14 main categories and 26 subcategories). Entity types are inspired from the named entity types supported by the Azure Cognitive Service for Language from Microsoft. These entities are defined in detail and accompanied by specific examples. The training and test datasets are made publicly available for research in NER after the challenge.

This paper has two contributions. First, we introduce a benchmarking dataset for the Vietnamese NER task, with a rich set of common entity types. Second, the shared task allows to evaluate the performance of different models which are submitted by participants in the VLSP community, sharing the knowledge in the field.

The remaining sections of this paper are as follows. Section 2 introduces the task description and the definition of named entity types, as well as the datasets built for training and testing. Section 3 presents the submitted systems and discusses their achieved results. In the final section, we conclude the paper with some discussions on the work perspectives.

2. Shared task description

2.1. Task description

Similarly to previous competitions, several steps are undertaken to organize the VLSP 2021 NER shared task:

- Definition of entity types and annotation guidelines.
- Collection and annotation of the training and test datasets.
- Distribution of datasets to participants following the campaign schedule.
- Evaluation the test results submitted by the participant teams.

As mentioned above, the VLSP 2016 and 2018 NER shared tasks dealt with the recognition of three common entity types in documents, taking into account nested entities. The scope of VLSP 2021's campaign is to assess

the ability to recognize entities in several categories (14 main types, 26 subtypes and 1 generic type), as an extension of the datasets published in VLSP 2018 [10], with the definition of more entity types to be able to fully capture the meaningful entity information in the document.

For data collection and annotation, we make use of the whole dataset from VLSP 2018 and extend its annotation with the newly defined entity types. In addition, we have collected additional data from media sites to be able to supplement some of the less common entity types. The data provided to the teams is in MUC format, but we also provide tools for easily converting data to other formats such as CoNLL [2]. In the following subsections, we will present the definition of entity types, the data format, as well as the data collection and annotation.

2.2. Named Entity Types

In the 2016 and 2018 VLSP workshops, the NER shared task consists in evaluating the performance of NE recognizers for three entity types: names of persons (PER), organizations (ORG), and locations (LOC). This year, an extended set of named entity types are defined based on the NER labels supported by the Microsoft Azure Cognitive Service for language3. The new set is composed of 14 main entity types and 26 sub-entity types. Sub-entity types are given to better describe the main types. For example, ORG contains sub-labels: ORG-Medical (Medical companies and groups), ORG-Stock (Stock exchange groups), or ORG-Sports (Sports-related organizations). Because there are more labels, this year's shared task will be more challenging both for the organizers building the NER dataset and for the participating teams developing NER models.

The main entity types are shortly described in Table 1. In addition to these 14 categories, 26 subcategories have been defined for DateTime, Event, Location, Organization, Product, and Quantity. Detail annotation guidelines for the VLSP 2021 NER challenge can be found on the shared task website.

2.3. Data Format

For this year's competition, we provided the teams with data in MUC format [1] which contains only NE information annotated using the markup language.

For example, named entities in the Vietnamese sentence "Anh Thanh là cán bộ Ủy ban nhân dân Thành phố Hà Nội." are annotated in MUC format as follows:

<ENAMEX TYPE="PERSON"> Anh Thanh </ENAMEX> là cán bộ <ENAMEX TYPE="ORGANIZATION"> Uỷ ban nhân dân <ENAMEX TYPE="LOCATION"> thành phố Hà Nội </ENAMEX> </ENAMEX>.

The tag pair <ENAMEX> </ENAMEX> is used to label each named entity appearing in the text, while the entity type is given by the "TYPE" attribute.

In addition, a tool is built to convert datasets from MUC format to column data format as shown in Table 2.

2.4. Data Collection and Annotation

2.4.1. Data Collection

The VLSP 2021 NER corpus is composed of two packages. The first package is the whole dataset published by the VLSP 2018 NER challenge. The second package is a set of articles which are newly collected from news websites such as vnexpress.vn, baomoi.com, zingnew.vn, etc. This new package contains articles that cover missing entity types and domains in the VLSP 2018 dataset. After the annotation process (presented below), we finally obtained an annotated dataset that includes 1282 articles from VLSP 2018 NER challenge, and 824 new articles belonging to several domains such as life, science and technology, education, sport, law, entertainment, etc,…

2.4.2. Annotation Procedure

For data annotation, we use WebAnno, a general-purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations.

Table 1: Named entity categories:

No	Category	Description	Example
1	Person	Names of people	Ông Mạnh là giảng viên trường Đại học Khoa học Tự nhiên.
2	PersonType	Job types or roles held by a person	Ông Mạnh là giảng viên trường Đại học Khoa học Tự nhiên.
3	Location	Natural and human made landmarks, structures, Geographical features,and geopolitical entities	Hôm nay, tôi lên Hà Nội làm hồ sơ nhập học.
4	Organization	Companies, political groups, musical bands, sport clubs, government bodies, and public organizations.	Ông Mạnh là giảng viên trường Đại học Khoa học Tự nhiên.
5	Event	Historical, social, and naturally occurring events.	Chiến dịch Điện Biên Phủ diễn ra tại lòng chảo Mường Thanh, Điện Biên, Lai Châu.
6	Product	Physical objects of various categories.	Tôi mới được ba mua cho một chiếc điện thoại Iphone XS Max.
7	Skill	A capability, skill, or expertise	Anh ta rất thông thạo tiếng Pháp.
8	Address	Full mailing addresses.	Nếu có thắc mắc, hãy liên lạc với tôi qua địa chỉ: Hà Lan, Số nhà 34, Ngõ 75, Thanh Xuân, Hà Nội. Hoặc gửi thư qua địa chỉ email: halan@gmail.com, số điện thoại: 03476229456.
9	Phone number	Phone numbers.	Nếu có thắc mắc, hãy liên lạc với tôi qua địa chỉ: Hà Lan, Số nhà 34, Ngõ 75, Thanh Xuân, Hà Nội. Hoặc gửi thư qua địa chỉ email: alan@gmail.com,số điện thoại: 03476229456.
10	Email	Email addresses.	Nếu có thắc mắc, hãy liên lạc với tôi qua địa chỉ: Hà Lan, Số nhà 34, Ngõ 75, Thanh Xuân, Hà Nội. Hoặc gửi thư qua địa chỉ email: halan@gmail.com, số điện thoại: 03476229456.
11	URL	URLs to websites.	Mọi dữ liệu được lấy từ trang web: https://vnexpress.net/.
12	IP	Network IP addresses	Bước 1: Truy cập vào máy chủ theo địa chỉ IP: 192.168.10.3. Sau đó đăng nhập và làm bài tập.
13	DateTime	Dates and times of day	Hôm nay, tôi lên Hà Nội làm hồ sơ nhập học.
14	Quantity	Numerical measurements and	Lớp tôi có 23 bạn nam và 25 bạn nữ

Table 2: A Vietnamese sentence in column data format:

Word	POS	Chunk	NE	Nested NE
Anh	N	B-NP	O	O
Thanh	NPP	B-NP	B-PER	O
là	V	B-VP	O	O
cán_bộ	N	B-NP	O	O
Uỷ_ban	N	B-NP	B-ORG	O
nhân_dân	N	I-NP	I-ORG	O
Thành_phố	N	I-NP	I-ORG	B-LOC
Hà_Nội	NPP	I-NP	I-ORG	I-LOC
.	.	O	O	O

whole dataset published by the VLSP 2018 NER challenge. The second package is a set of articles which are newly collected from news websites such as vnexpress.vn, baomoi.com, zingnew.vn, etc. This new package contains articles that cover missing entity types and domains in the VLSP 2018 dataset.

After the annotation process (presented below), we finally obtained an annotated dataset that includes 1282 articles from VLSP 2018 NER challenge, and 824 new articles belonging to several domains such as life, science and technology, education, sport, law, entertainment, etc.

2.4.2. Annotation Procedure

For data annotation, we use WebAnno, a general-purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations. Custom annotation layers can be defined, allowing WebAnno to be also used for non-linguistic annotation tasks. Furthermore, WebAnno is a multi-user tool supporting different roles such as annotator, curator, and project manager. The progress and quality

2.4.2. Annotation Procedure

For data annotation, we use WebAnno, a general-purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations. Custom annotation layers can be defined, allowing WebAnno to be also used for non-linguistic annotation tasks. Furthermore, WebAnno is a multi-user tool supporting different roles such as annotator, curator, and project manager. The progress and quality of annotation projects can be monitored and measured in terms of inter-annotator agreement.

The NER data annotation is done by 6 annotators and 3 experts. We trained annotators with definitions and examples of named entity types before assigning data packages to each of them. Due to limited time, we didn't perform cross-labelling. During the annotation process, if there was any problem, annotators would discuss with experts. The annotations produced by annotators are finally reviewed by experts. We export annotated data in TSV format (an output format of Webanno) and then convert the data to the MUC format.

2.5. Dataset Splits

The annotated dataset is split into a training set and a test set. Since the VLSP 2018 NER dataset was already published, its updated version with new entity types is included in the training set. To provide a reliable model evaluation, for the package of 848 articles, we design a split ensuring a relatively balanced coverage of entity types, based on article domains. Specifically, our dataset includes:

- A training dataset composed of 1830 articles, in which 1282 articles are developed from the VLSP 2018 NER dataset, and 538 are new articles. This dataset contains in total 81,173 named entities.
- A test dataset composed of 310 new articles, with a total number of 19,538 named entities.

Table 3 shows a statistic of entities labels appearing in the training and test datasets. It can be seen that some common NEs such as DATETIME, PERSON, PERSONTYPE, QUANTITY,... have a high frequency in the dataset. Other NEs like IP, URL, EMAIL,... have relatively few occurrences. This will also affect

to the training result of NER models. In addition, due to short time for the corpus annotation, we couldn't ensure a similar distribution between first-level and nested-level of NEs in training and test datasets. This shortcoming will be addressed in future work to obtain a high quality benchmark dataset.

Table 3: Data statistic:

NE types	Training data		Test data		NE types	Training data		Test data	
	First-level	Nested-level	First-level	Nested-level		First-level	Nested-level	First-level	Nested Level
ADDRESS	85	6	17	6	ORGANIZATION-MED	171		82	61
DATETIME	3223	112	816	7	ORGANIZATION-	1492	505	372	45
					SPORTS				
DATETIME-DATE	2224	14	824	6	ORGANIZATION-STOCK	16	0	35	2
DATETIME-DATERANGE	310	97	46	102	PERSON	15076	-20	2719	10
DATETIME-DURATION	1065	274	193	306	PERSONTYPE	4680	300	704	129
DATETIME-SET	34	10	2	2	PHONENUMBER	282	0	10	0
DATETIME-TIME	344	52	35	31	PRODUCT	2439	58	417	64
DATETIME-TIMERANGE	187	168	13	124	PRODUCT-AWARD	1266	6	155	8
EMAIL	75	1	2	0	PRODUCT-COM	35	0	55	31
EVENT	366	170	124	66	PRODUCT-LEGAL	283	22	142	33
EVENT-CUL	293	47	14	3	QUANTITY	3728	114	104	59
EVENT-GAMESHOW	379	42	54	3	QUANTITY-AGE	338	183	16	246
EVENT-NATURAL	165	0	7	2	QUANTITY-CUR	990	206	157	347
EVENT-SPORT	324	248	89	65	QUANTITY-DIM	408	131	146	153
IP	117	0	15	0	QUANTITY-NUM	6075	3	3479	2
LOCATION	7451	23	531	29	QUANTITY-ORD	937	0	486	36
LOCATION-GEO	453	5	138	2	QUANTITY-PER	979	34	359	11
LOCATION-GPE	9799	71	3096	22	QUANTITY-TEM	83	7	1	10
LOCATION-STRUC	641	78	140	9	SKILL	40	2	1	1
MISCELLANEOUS	662	4	0	0	URL	317	3	9	0
ORGANIZATION	8329	1921	1452	448	Overall	76161	5012	17057	2481

3. Submissions and Evaluations

In this section, we first present an overview of methods for NER. We continue with an introduction of submitted models and the evaluation metrics for NER performance. And

finally, we discuss the evaluation results of these models.

3.1. Methods for NER

The NER problem can be formulated as follows. Given a predefined set of entity types, and a sequence of tokens (words or other lexical units) s = w1, w2, · · · , wN , a NER system needs to output triplets Is, Ie, t , where Is and Ie are respectively the start and end positions of an entity of type t mentioned in s. Annotated NER data can be represented in a number of formats, such as XML with tags marking the entities, or the BIO format. Among these formats, BIO is the most commonly used. In the BIO format, for every token in the document, a label B-eType, or I-eType or O is assigned to that token if it is, respectively, the start or end token of an eType entity, or outside of any entity. The NER problem is consequently a sequence labeling problem.

The approaches for NER can be classified into 4 types [12] as below.

- Knowledge-based systems: These systems do not require labeled data, and operate on lexical resources combined with domain-specific expert knowledge. This approach often gives high precision, but quite low recall due to limited lexical resources.
- Unsupervised and bootstrapped systems: These systems use little or even no entity annotated data to build models, but information that can be extracted from raw text such as TF-IDF, or linguistic information such as orthography, part-of-speech (POS), . . .combined with patterns to create entity extractors and classifiers.
- Feature-engineered supervised systems: Supervised machine learning methods such as SVM, or especially methods for sequence labeling such as HMM, CRF are widely used from 2000 to 2016. In these systems, tokens are vectorized based on a predefined set of features, usually based on orthographic, lexicons, prefixes, suffixes, n-grams, ...
- Feature-inferring neural network systems: With the rise of neural networks, in recent systems, manually constructed feature vectors are replaced by word embeddings. The pre-trained word embeddings contain word representations in n-dimensional space, which are trained on a large raw corpus using an unsupervised method like CBOW or Skip-gram. Neural networks for the NER problem can be built according to several architectures such as word level, character level, character + word level, character + word + affix level. Research as well as experiment results on NER datasets have shown that pre-trained word embeddings combined with neural network models such as RNN, or Bi-LSTM have made a huge leap in the performance of NER systems. In recent years, neural networks along with pre-trained word embeddings have been increasingly used in text processing tasks, including NER.

For Vietnamese, several works on NER have been undertaken (e.g. [13], [14], [15]), but only recently have some important benchmark datasets been published, especially in the framework of the VLSP 2016 and 2018 NER challenge. The systems developed for these two challenges also show a clear shift in terms of approach as described above. In VLSP 2016, teams mostly used the CRF model with manual features. In VLSP 2018, most of the teams used word embeddings like Fastext or Glove with a classification model using CRF or a combination of CRF and LSTM. It is worth to equally mention a new work in 2021 [11], in which the authors introduced a manually-annotated COVID-19 domain-specific dataset for Vietnamese, built and evaluated a NER system by fine-tuning the pre-trained language models PhoBERT [16] for Vietnamese.

3.2. Submissions

During the evaluation campaign, we recorded 23 pre-registered participants, of which 6 teams signed the user agreement to access the VLSP 2021 NER datasets. However, at the end of the challenge, only 4 teams (named NER1, NER2, NER3, NER4) submitted technical reports describing their methods and results in detail.

Based on the technical reports, we have some summaries of the methods that the participating teams used to solve this problem as follows:

- NER1: The authors consider words as spans, and firstly tackle this problem as a span labeling task. Then, they extract nested entities and classify them into different levels. For each level, they use a pretrained encoder to get word representations. After learning the representations of tokens, they conduct fine-tuning on the pre-trained XLM-RoBERTa model. Lastly, they apply a model ensemble from 4 level trained models to achieve the final result.
- NER2: The authors adopt a technique from dependency parsing (Biaffine dependency parsing model) to tackle the problem of nested entities. They also apply the Coteaching+ technique to enhance the overall performance and propose an ensemble algorithm to combine predictions.
- NER3: The authors propose a two stage model for nested NER. They utilize an entity proposal module to filter the easy non-entity spans for efficient training. In addition, they combine all variants of the model to improve overall accuracy of their system which contains three modules: text representation (use PhoBERT, processes segmented sentence to contextual representation), entity proposal (generate entity candidates), entity classification module (classifies entity candidates).
- NER4: The author integrated the deep contextualized embedding, which includes word embedding, ELMo, and BERT representation, into a bipartite flat-graph network for Vietnamese nested named entity recognition dataset.

3.3. Evaluation metrics

In the NER challenge, we use the F1 scores to compute the performance of NER systems (for each entity type):

$$\mathrm{F1} = \frac{2*P*R}{(P+R)} \quad (1)$$

where P (Precision), and R (Recall) are determined as follows: E_{true}

$$\mathrm{P} = \frac{NE_{\mathrm{true}}}{NE_{\mathrm{sys}}} \quad (2)$$

$$\mathrm{P} = \frac{NE_{\mathrm{true}}}{NE_{\mathrm{ref}}} \quad (3)$$

where:

- NEref : The number of NEs in gold data
- NEsys: The number of NEs in recognizing system
- NEtrue: The number of correctly recognized NEs

The F1 score will be calculated for each NE type. However, to evaluate the systems in more detail, we also calculated F1 scores according to the NE level: Top-level, Nested-level and overall (NEs belonging to both top-level and nested-level).

3.4. Results

We evaluate the results of the teams based on F1 score, as presented in 3.3. Team NER1 submitted one model, while teams NER2, NER3, and NER4 submitted three results.

Based on the results submitted by the teams, we evaluate the results on NE types, the details of which are described in Table 4. In this table, it can be seen that most of models get a higher performance for specific entity types, such as PERSON, LOCATION-GPE, or ORG. In addition, some other entities such as EMAIL, PHONENUMBER, IP, DATETIME, or QUANTITY subtypes also have high F1 score for systems having relied on the lexical characteristics of these entities to build regular expressions to recognize them.

We also provide statistics on the scores of the teams based on the top level and nested level as shown in Table 5. It can be seen that the F1 score of the top-level is much higher than the F1 score of the nested level, by a margin of 3-5%. This is understandable because the number of entities of the top level is higher than that of the nested level in both training and testing dataset.

The best overall results are achieved by team NER3, with P = 64.87%, R = 60.81% and F1 = 62.71%. Table 6 presents the final ranking

results. It can be seen that the achieved results are still limited, which suggests that improvement is needed not only in the scale and the quality of the training and test data splits, but also in the approaches for the problem.

Table 4: NER 2021 results by NE types:

		NER1		**NER2**			**NER3**			**NER4**	
NE types	**NEs**	**Model 1**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
ADDRESS	23	8,51	15,38	0,00	14,81	0,00	0,00	0,00	**18,18**	12,50	5,41
DATETIME	823	49,46	**54,69**	9,81	52,30	21,14	20,95	22,53	44,83	39,87	38,73
DATETIME-DATE	830	46,66	55,07	21,98	55,83	57,39	57,75	**57,91**	51,00	50,17	49,98
DATETIME-DATERANGE	148	22,22	3,99	4,95	3,99	**30,39**	28,57	29,96	10,62	10,19	11,91
DATETIME-DURATION	499	70,25	73,15	9,23	**78,50**	76,19	77,12	77,49	64,73	59,82	66,35
DATETIME-SET	4	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
DATETIME-TIME	66	49,81	43,22	11,83	44,69	64,12	**66,67**	62,73	30,93	28,57	38,46
DATETIME-TIMERANGE	137	53,22	0,00	6,89	0,00	54,39	**58,58**	58,30	30,77	39,51	44,67
EMAIL	2	**100,00**	**100,00**	0,00	**100,00**	**100,00**	66,67	66,67	0,00	66,67	0,00
EVENT	190	17,94	9,50	10,28	9,30	**24,37**	19,77	20,15	6,87	10,63	9,79
EVENT-CUL	17	57,97	9,76	19,67	10,26	47,76	53,73	**58,67**	25,40	29,63	23,53
EVENT-GAMESHOW	57	41,32	9,02	12,74	8,82	41,12	51,02	**52,26**	7,92	17,39	12,96
EVENT-NATURAL	9	30,77	12,50	13,33	18,18	18,18	**36,36**	**36,36**	18,18	18,18	0,00
EVENT-SPORT	154	51,44	53,51	35,84	51,84	54,19	55,60	**57,35**	35,32	49,38	44,68
IP	15	**96,55**	71,79	0,00	88,24	0,00	0,00	0,00	22,22	56,00	75,86
LOCATION	560	21,78	26,34	4,55	**27,19**	20,31	20,30	20,13	16,51	16,41	17,98
LOCATION-GEO	140	37,15	25,38	29,02	21,05	31,63	35,42	**41,04**	38,68	27,14	25,48
LOCATION-GPE	3118	65,69	71,06	46,31	71,34	74,71	75,79	**75,96**	63,89	61,17	62,19
LOCATION-STRUC	149	55,56	42,37	36,19	43,27	53,23	59,20	**63,31**	29,46	32,43	36,36
ORGANIZATION	1900	65,72	60,91	41,03	64,24	66,18	66,87	**68,21**	48,70	51,07	52,27
ORGANIZATION-MED	143	59,86	56,47	52,40	44,19	62,87	64,47	**65,16**	43,48	49,62	37,72
ORGANIZATION-SPORTS	417	60,80	74,16	49,17	**74,75**	72,56	74,62	73,25	44,88	57,01	60,80
ORGANIZATION-STOCK	37	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PERSON	2729	**92,16**	85,67	60,95	89,46	90,21	91,27	91,37	74,24	81,85	84,64
PERSONTYPE	833	21,46	52,28	33,98	53,30	53,04	**56,19**	55,11	44,41	48,30	49,19
PHONENUMBER	10	81,82	70,00	70,59	70,00	73,68	**90,00**	**90,00**	19,05	40,00	33,33
PRODUCT	481	34,59	27,70	12,65	29,94	36,97	37,43	**38,65**	18,55	21,18	22,34
PRODUCT-AWARD	86	16,00	0,00	22,47	4,57	16,49	**20,94**	17,11	0,00	0,00	0,00
PRODUCT-COM	163	32,60	38,32	22,65	41,40	39,40	38,52	**42,84**	15,52	16,37	19,13
PRODUCT-LEGAL	175	10,15	33,62	35,79	34,67	37,57	38,30	**40,40**	15,92	7,27	6,15
QUANTITY	163	6,62	6,26	1,21	7,37	8,31	8,74	**8,82**	8,57	5,08	7,94
QUANTITY-AGE	262	73,91	85,88	42,35	**86,99**	81,08	83,70	83,26	72,93	67,29	65,06
QUANTITY-CUR	504	75,11	87,20	50,07	**87,51**	85,73	86,83	86,64	80,51	71,22	73,27
QUANTITY-DIM	299	43,54	64,56	36,58	**66,08**	60,54	60,59	59,70	29,83	23,80	28,19
QUANTITY-NUM	3481	43,95	34,58	35,20	30,88	46,35	46,41	**47,18**	37,28	34,92	39,31

QUANTITY-ORD	522	32,61	25,50	33,93	23,83	31,16	32,54	**35,97**	18,23	15,79	20,02
QUANTITY-PER	370	89,60	87,93	69,18	88,95	94,37	94,90	**95,62**	89,55	87,70	86,86
QUANTITY-TEM	11	**90,91**	73,33	30,77	81,48	58,33	60,87	75,00	66,67	41,67	29,63
SKILL	2	0,00	0,00	0,00	0,00	**44,44**	0,00	0,00	0,00	0,00	0,00
URL	9	77,42	80,00	51,28	**90,32**	75,68	64,86	57,14	48,48	28,07	48,48

Table 5: Overall evaluation:

		Top-level evaluation			Nested evaluation			Overall		
Team	**Model**	**P**	**R**	F_1	**P**	**R**	F_1	**P**	**R**	F_1
NER1	1	57.77	57.52	**57.65**	54.23	52.61	**53.41**	55.89	54.88	**55.38**
NER2	1	66.43	56.03	60.79	64.48	50.21	56.46	65.42	52.90	58.50
	2	64.69	57.72	61.01	61.31	51.74	56.12	62.92	54.51	58.41
	3	65.99	59.17	**62.39**	62.18	53.22	**57.35**	63.98	55.97	**59.71**
NER3	1	62.64	63.75	63.19	59.99	58.34	59.15	61.24	60.84	61.04
	2	64.43	64.21	64.32	62.37	58.59	60.42	63.36	61.19	62.26
	3	64.52	65.22	**64.87**	62.09	59.58	**60.81**	63.24	62.19	**62.71**
NER4	1	59.03	45.59	51.44	58.57	40.79	48.09	58.79	43.01	49.68
	2	56.00	47.65	51.49	55.14	42.32	47.89	55.56	44.79	49.59
	3	57.61	50.07	**53.58**	56.96	44.65	**50.06**	57.28	47.16	**51.73**

Table 6: Ranking results:

No.	Top-level	Nested	Overall	Rank
NER3	64.87	60.81	62.71	1
NER2	62.39	57.35	59.71	2
NER1	57.65	53.41	55.38	3
NER4	53.58	50.06	51.73	4

4. Conclusion

In this article, we have presented the datasets and evaluation results of the VLSP 2021 NER shared task, which was held along with other competitions during the 8th VLSP workshop. This year's competition has a significant change from VLSP 2016 and 2018 NER shared tasks in term of named entity types. The set of entity types are defined based on those supported by Azure Cognitive Service for Language (Microsoft), which include 14 main types and 26 sub-types. Because there are an important number of entity types, the performance of the participant systems are still limited, and there is much room for improvement.

The built datasets are made available to the VLSP community for research purpose. We plan a revision and extension of those datasets for improving the annotation quality and reducing data imbalance, allowing better performances for NER systems. Another NER challenge for the next edition of VLSP workshop is also scheduled.

Acknowledgement

This work is partially supported by VINIF and the NLP group at VNU University of Science. Ha My Linh was funded by Vingroup Joint Stock Company and supported by the Domestic Master/ PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), Vingroup Big Data Institute (VINBIGDATA), code VINIF.2021.TS.028.

References

[1] B. M. Sundheim, Overview of Results of the MUC-6 Evaluation, in: Proceedings of the 6th Conference on Message Understanding, MUC6 '95, Association for Computational Linguistics, USA, 1995, pp. 13–31. doi:10.3115/1072399.1072402.

[2] E. F. Tjong Kim Sang, Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, in: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002), 2002.

[3] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147.

[4] L. Derczynski, E. Nichols, M. van Erp, N. Limsopatham, Results of the WNUT2017 shared task on novel and emerging entity recognition, in: Proceedings of the 3rd Workshop on Noisy User-generated Text, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 140–147. doi:10.18653/v1/W17-4418.

[5] D. Benikova, C. Biemann, M. Kisselew, S. Padó, Germeval 2014 named entity recognition shared task: Companion paper, 2014. URL https://nbn-resolving.org/urn: nbn:de:gbv:hil2-opus-3006

[6] E. Leitner, G. Rehm, J. Moreno-Schneider, Fine-grained Named Entity Recognition in Legal Documents, in: M. Acosta, P. Cudré-Mauroux, Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019), no. 11702 in Lecture Notes in Computer Science, Springer, Karlsruhe, Germany, 2019, pp. 272–287.

[7] S. Tedeschi, V. Maiorca, N. Campolungo, Cecconi, R. Navigli, WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER, in: Findings of theAssociation for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2521–2533.

[8] J. Nothman, N. Ringland, W. Radford, Murphy, J. R. Curran, Learning multilingual named entity recognition from wikipedia, Artificial Intelligence, Vol. 194, 2013, pp. 151–175,https://doi.org/10.1016/j.artint.2012.03.006.

[9] A.-L. Minard, M. Speranza, R. Urizar, Altuna, M. van Erp, A. Schoen, C. van Son, MEANTIME, the NewsReader multilingual event and time corpus, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portoroz,˘ Slovenia, 2016, pp. 4417–4422.

[10] N. T. M. Huyen, N. T. Quyen, X. V. Luong, T. M. Vu, H. N. T. Thu, VLSP Shared task: Named Entity Recognition, Jounral of Compiter Sicence and Cybernectics.

[11] T. H. Truong, M. H. Dao, D. Q. Nguyen, COVID-19 Named Entity Recognition for Vietnamese, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, p. 2146–2153.

[12] V. Yadav, S. Bethard, A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158.

[13] Nguyen, Truc-Vien T. and Cao, Tru, VN-KIM IE: Automatic Extraction of Vietnamese Named-Entities on the Web, New Generation Computing, Vol. 25, 2007, pp. 277–292, Doi:10.1007/s00354-007-0018-4.

[14] H. T. Nguyen, T. H. Cao, Named Entity Disambiguation: A Hybrid Approach, International Journal of Computational Intelligence Systems 5 (2012) 1052–1067, https://doi.org/10.1080/18756891.2012.747661.

[15] Q. H. Pham, M.-L. Nguyen, B. T. Nguyen, N. V. Cuong, Semi-supervised Learning for Vietnamese Named Entity Recognition using Online Conditional Random Fields, in: Proceedings of the Fifth Named Entity Workshop, Association for Computational Linguistics, Beijing, China, 2015, pp. 50–55. doi:10.18653/v1/W15-3907.

[16] D. Q. Nguyen, A. T. Nguyen, PhoBERT: Pre-trained language models for Vietnamese, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1037–1042.