Overview of VLSP RelEx shared task: A Data Challenge for Semantic Relation Extraction from Vietnamese News

Mai-Vu Tran[1], **Hoang-Quynh Le**[1], **Duy-Cat Can**[1]
Huyen Nguyen[2], **Linh Nguyen Tran Ngoc**[3] **and Tam Doan Thanh**[4]
[1]VNU University of Engineering and Technology, Hanoi, Vietnam.
{vutm, lhquynh, catcd}@vnu.edu.vn
[2]Hanoi University of Science, Vietnam National University, Vietnam.
huyenntm@hus.edu.vn
[3]Viettel Big Data Analytics Center, Viettel Telecommunication Company, Viettel Group.
linhntn3@viettel.com.vn
[4]doanthanhtam283@gmail.com

Abstract

This paper reports the overview of RelEx shared task for semantic relation extraction from Vietnamese News, which is hosted at the seventh annual workshop on Vietnamese Language and Speech Processing (VLSP 2020). This task focuses on classifying entity pairs in Vietnamese News text into four different, non-overlapping categories of semantic relations defined in advance. In order to generate a fair benchmark, we build a human-annotated dataset of 1,056 documents and 5,900 instances of semantic relations, collected from Vietnamese News in several domains. All models will be evaluated in terms of macro- and micro-averaged F1 scores, two typical evaluation metrics for semantic relation extraction problem.

1 Introduction

The rapid growth of volume and variety of news brings an unprecedented opportunity to explore electronic text but an enormous challenge when facing a massive amount of unstructured and semi-structured data. Recent research progress in text mining needs to be supported by Information Extraction (IE) and Natural Language Processing (NLP) techniques. One of the most fundamental sub-tasks of IE is Relation Extraction (RE). It is the task of identifying and determining the semantic relations between pairs of named entity mentions (or nominals) in the text (Aggarwal, 2015). Receiving the (set of) document(s) as an input, the relation extraction system aims to extract all predefined relationships mentioned in this document by identifying the corresponding entities and determining the type of relationship between each pair of entities (see examples in Figure 1).

Evidence: 23353704	Tại buổi họp báo, ông **Nguyễn Quang Huyền**, Phó Cục trưởng **[Cục Quản lý và Giám sát Bảo hiểm]** [**Bộ Tài chính**] cho biết,	
Relation type: AFFILIATION	*Entity 1:* PER Nguyễn Quang Huyền	*Entity 2:* ORG Cục Quản lý và Giám sát Bảo hiểm
Relation type: AFFILIATION	*Entity 1:* ORG Nguyễn Quang Huyền	*Entity 2:* ORG Bộ Tài chính
Relation type: PART-WHOLE	*Entity 1:* ORG Cục Quản lý và Giám sát Bảo hiểm	*Entity 2:* ORG Bộ tài chính

Figure 1: Relation examples.

RE is of significant importance to many fields and applications, ranging from ontology building (Thukral et al., 2018), improving the access to scientific literature (Gábor et al., 2018), question answering (Lukovnikov et al., 2017; Das et al., 2017) to major life events extraction (Li et al., 2014; Cavalin et al., 2016) and many other applications. However, manually curating relations is plagued by its high cost and the rapid growth of the electronic text.

For English, several challenge evaluations have been organized such as Semantic Evaluation (SemEval) (Gábor et al., 2018; Hendrickx et al., 2010), BioNLP shared task (Deléger et al., 2016), and Automatic Content Extraction (ACE) (Walker et al., 2006). These challenges evaluations attracted many scientists worldwide to attend and publish their latest research on semantic relation extraction. Many approaches are proposed for RE in English texts, ranging from knowledge-based methods to machine learning-based methods (Bach and Badaskar, 2007; Dongmei et al., 2020). Studies on this problem for Vietnamese text are still in the early stages with a few initial achievements. In recent years, there has been a growing interest to develop computational ap-

proaches for extracting semantic relations in Vietnamese text automatically with proposals of several methods. Despite these attempts, the lack of a comprehensive benchmarking dataset has limited the comparison of different techniques. RelEx challenge task in VLSP was set up to provide an opportunity for researchers to propose, assess and advance their researches.

The remainder of the paper is organized as follows. Section 2 gives the description about RelEx shared task. The next section describes the data collection and annotation methodologies. Subsequently, section 4 describes the competition, approaches and respective results. Finally, Section 5 concludes the paper.

2 RelEx 2020 Challenge

As the first shared task of relation extraction for Vietnamese text, we go from typical relations between three fundamental entities in News domain: $Location$, $Organization$ and $Person$. All semantic relations between nominals other than the aforementioned entities were excluded. Based on these three types of annotated entities, we selected four relation types with coverage sufficiently broad to be of general and practical interest. Our selection is referenced and modified based on the relation types and subtypes used in the ACE 2005 task (Walker et al., 2006). We aimed at avoiding semantic overlap as much as possible. Four relation types are described in Table 1 and as follow.

- The $LOCATED$ relation captures the physical or geographical location of an entity.
- The $PART - WHOLE$ relation type captures the relationship when the parts contribute to the structure of the wholes.
- The $PERSONAL-SOCIAL$ relations describe the relationship between people.
- The $ORGANIZATION - AFFILIATION$ relation type represents the organizational relationship of entities.
- We do not annotate non-relation entity pairs ($NONE$). These negatives instances need to be self-generated by participated teams, if necessary.

In the case of $PERSONAL - SOCIAL$, an undirected relation type, two entities are symmetric (i.e., not ordered). Other relation types are directed, i.e., their entities are asymmetry (i.e., order sensitive). We restrict the direction of these relation types always come from entity 1 to entity 2. The participated system needs to define which entity mention plays the role of entity 1 and which entity mention plays the role of entity 2.

This task only focused on intra-sentence relation extraction, i.e., we limit relations to only those that are expressed within a single sentence. The relations between entity mentions are annotated if and only if the relationship is explicitly referenced in the sentence that contains the two mentions. Even if there is a relationship between two entities in the real world (or elsewhere in the document), there must be evidence for that relationship in the local context where it is tagged. We do not accept the case of bridging relations (i.e., a relationship derived from two other consecutive relationships), uncertain relations, inferred relations, and relation in the future tenser (i.e., allusion/mean to happen in the future).

A relation is defined by two entities participating in this relationship. In other words, a sentence can contain several different relations if it has more than one pairs of entities. Any qualifying relations must be predicted, even if the text mentions them is overlap or nested with range text of other relations. We do not allow the multi-label cases, i.e., a pair of entities must have only one relationship or no relation. If there is an ambiguity between some relation types, the participated system needs to decide to choose the most suitable label.

Only binary relations are accepted. N-nary relations should be predicted if and only if they can be split into several binary relations without changing the semantic meaning of the relationships.

3 Task Data

3.1 Data Statistics

For the task, we prepared a total of $1,056$ News documents: 506 documents for the training, 250 documents for development and 300 documents in the test set. Of all $1,056$ news documents, 815 documents were selected in a single crawler process. The remaining 241 documents were selected in another crawler process to represent difference features and were incorporated into the test set. We

No	Relation	Agruments	Directionality
1	LOCATED	PER – LOC, ORG – LOC	Directed
2	PART – WHOLE	LOC – LOC, ORG – ORG, ORG – LOC	Directed
3	PERSONAL – SOCIAL	PER – PER	Undirected
4	ORGANIZATION –AFFILIATION	PER – ORG, PER – LOC, ORG – ORG, LOC – ORG	Directed

Table 1: Relation types permitted arguments and directionality.

	Training set	Development set	Test set
Number of documents	506	250	300
LOCATED	612	346	294
PART-WHOLE	1176	514	815
PERSONAL - SOCIAL	102	98	449
ORGANIZATION -AFFILIATION	771	518	205

Table 2: Statistics of the RelEx dataset.

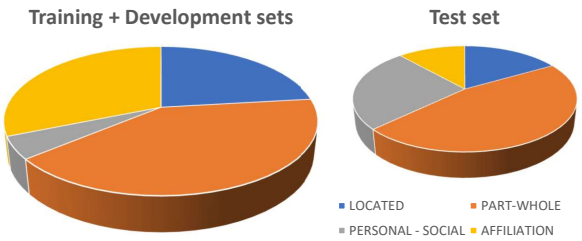

Figure 2: The distribution of relation types in Datasets.

then prepared the manual annotations, Table 2 describes statistics of the RelEx dataset in detailed. Figure 2 show the distribution of relation types in training/development set and the test set. Due to the effect of adding 'strange' data to the test set, the rate is partly inconsistent between training/development and test set.

3.2 Data Annotation

3.2.1 Annotators and Annotation Tool

There are 6 human annotators to participate in the annotation process. An annotation guideline with full definition and illustrative examples was provided. We used a week to train annotators about the markable and non-markable cases in documents. In the following week, annotators conducted trial annotations, then raised some issues that need clarification. An expert then preliminarily assessed the quality of the trial annotation process before started the full annotation process.

We used WebAnno[1] as the Annotation tool. It is a general purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations.

3.2.2 Annotation Process

The annotators were divided into two groups and used their account to conduct independent annotations, i.e., each document was annotated at least twice. The annotation process is described in Figure 3. First, the supervisor separated the whole dataset into several small parts. Each part was given to two independent annotators for annotating. For finding out the agreement between annotators, the committee then calculated the Inter-Annotator Agreement (IAA). Follow (Dalianis, 2018), IAA can be carried out by calculating the Precision, Recall, F-score, and Cohen's kappa, between two annotators. If the IAA is very low, for example, $F1$ is under 0.6, it may be due to the complexity and difficulty of the annotation task or the low quality of the annotation. For the RelEx task, the committee selected the IAA based on $F1$, and chose an acceptable threshold of 0.7. If the IAA between two annotators on a subset was smaller than 0.7, we went through the curation process with a third annotator to decide the final annotation.

4 Challenge Results

4.1 Data Format and Submission

The test set are formatted similarly with the training and development data, but without information for the relation label. The task is to predict, given a sentence and two tagged entities, which of the relation labels to apply. The participated teams must submit the result in the same format with the training and development data.

The participating systems had the following task: Given a documents and tagged entities, predict the semantic relations between those entities and the directions of the relations. Each teams can submit up to 3 runs for the evaluation.

4.2 Evaluation Metrics

The participated results were evaluated using standard metrics of Precision (P), Recall (R) and $F1$. In which, Precision indicates the percentage of

[1] http://webanno.github.io/webanno/

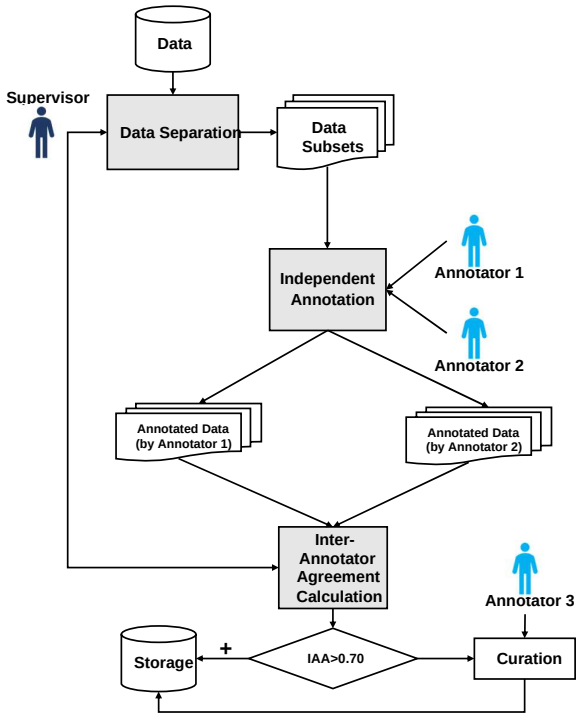

Figure 3: The annotation process.

system positives that are true instances, Recall indicates the percentage of true instances that the system has retrieved. $F1$ is the harmonic mean of Recall and Precision, calculated as follows:

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

We released a detailed scorer which outputs:

- A confusion matrix,
- Results for the individual relations with P, R and $F1$,
- The micro-averaged P, R and $F1$,
- The macro-averaged P, R and $F1$.

Our official scoring metric is macro-averaged $F1$, taking the directionality into account (except $PERSONAL - SOCIAL$ relations).

4.3 Participants and Results

4.3.1 Participants

A total of 4 teams participated in the RelEx task. Since each team was allowed to submit up to 3 runs (i.e., 3 different version of their proposal method), a total of 12 runs were submitted. Table 3 lists the participants and provides a rough overview of the system features. VnCoreNLP[2] and underthesea[3] are used for preprocessing. All proposed model are based on the deep neural network architectures with different approaches, go from a simple method (i.e., multilayer perceptron) to Bidirectional Long Short-Term Memory and more complex architectures (e.g., BERT with entity start). With the application of deep learning models, participated teams use several pre-trained embedding model. In addition to word2vec (Mikolov et al., 2013; Vu, 2016), RelEx challenge acknowledgement several BERT-based word embedding for Vietnamese, including PhoBERT (Nguyen and Nguyen, 2020), NlpHUST/vibert4news[4], FPTAI/vibert (The et al., 2020) and XLMRoBERTa (Conneau et al., 2020).

4.3.2 Results

As shown in Table 4, the macro-averaged $F1$ score of participated teams (only considering the best run) ranges from 57.99% to 66.16% with an average of 62.42%. For reference information, the micro-averaged $F1$ score ranges from 61.84% to 72.06% with an average of 66.99%. The highest macro-averaged P and R is 80.38% and 66.75%, respectively. However, the team with the highest P has quite low R, and vice versa, the team with the highest R has the lowest P. The first and second-ranked teams have the right balance between P and R.

We ranked the teams by the performance of their best macro-averaged $F1$ score. Team of Thuat Nguyen and Hieu Man Duc Trong from Hanoi University of Science and Technology, Hanoi, Vietnam submitted the best system, with a performance of 66.16% of $F1$, i.e., 2.74% better than the runner-up system. The second prize was awarded to Pham Quang Nhat Minh with 63.42% of $F1$. The third prize was awarded to SunBear Team from AI Research Team, R&D Lab, Sun Inc, who proposed many improvements in their model. The detailed results of all teams are shown in Table 5.

4.4 Discussion

4.4.1 Relation-specific Analysis

We also analyze the performance for specific relations on the best results of each team for each relation. $PART - WHOLE$ seems to be

[2] https://github.com/vncorenlp
[3] https://github.com/undertheseanlp
[4] http://huggingface.co/NlpHUST/vibert4news-base-cased

No	Team	Main method	Pre-processing	Embeddings	Additional Techniques
1	HT-HUS	Multi layer neural network	+ VnCoreNLP + Underthesea + Pre-processing rules	+ PhoBert + XLMRoBERTa	
2	MinhPQN	+ R-BERT + BERT with entity start	No information	+FPTAI/vibert + NlpHUST/vibert4news	+ Ensemble model
3	SunBear	+ PhoBert + Linear classification + Multi-layer Perceptron	Underthesea	+ PhoBert	+ Join training Named Entity Recognition and Relation Extraction + Data sampling + Label embedding
4	VC-TUS	Bidirectional Long Short -Term Memory network	VnCoreNLP	+ Word2Vec + PhoBert	+ Position features + Ensemble

Table 3: Overview of the methods used by participating teams in RelEx task.

Team	Macro-averaged			Micro-averaged		
	P	R	F1	P	R	F1
HT-HUS	73.54	62.34	**66.16**	76.17	68.37	**72.06**
MinhPQN	73.32	57.09	63.42	76.83	60.28	67.56
SunBear	58.44	**66.75**	62.09	60.82	**73.29**	66.48
VC-Tus	**80.38**	46.43	57.99	**83.51**	49.09	61.84

Results are reported in %.

Highest result in each column is highlighted in bold.

Table 4: The final results of participated teams (best run results).

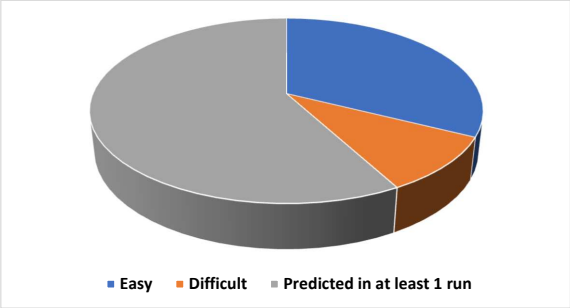

Figure 4: The annotation process.

the easiest relation. Comparing the best runs of teams, the lowest result for this relation is 79.57%, and the highest result was over 84.35%, i.e., the difference is comparatively small (4.78%). $ORGANIZATION-AFFILIATION$ is the relation that has the most difference between the best and worst system (16.73%). The most challenging relation is $PERSONAL-SOCIAL$. It is proved that being a problematic relation for all teams. This note can be clarified from the data statistics, although $PERSONAL-SOCIAL$ is a relation that has many different patterns in realistic, it accounts for only $\sim$ 5% of training and development data. It becomes even more difficult when it takes up $\sim$ 25% of test data. $LOCATED$ follows $PERSONAL-SOCIAL$ in terms of difficulty. Some of its patterns are confused with the $ORGANIZATION-AFFILIATION$ relation, i.e., whether a person is/do something in a particular location or is citizen/resident of a (geopolitical) location. An interesting observation shows that directional relations were not a difficult problem for participated teams. The submission with the most misdirected error failed only 7 examples out of the total number of results returned. Many submission does not have any errors in the directionality.

4.4.2 Difficult Instances

Figure 4 shows the ratio between easy cases (correctly predicted in all runs), difficult cases (did not found by any run), the rest are the number of examples that correctly predicted in at least one run (but not all runs). There were 140 examples ($\sim$ 10%) that are classified incorrectly by all systems. Except for a handful of errors

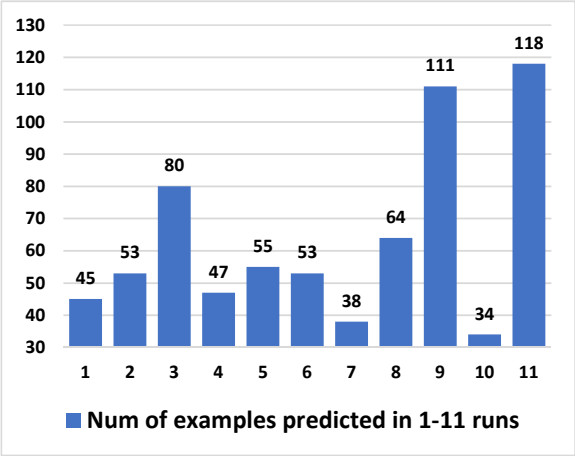

Figure 5: Number of examples predicted in 1-11 runs.

Team/Run	LOC	AFF	P-W	P-S	Macro-averaged			Micro-averaged		
	F1	F1	F1	F1	P	R	F1	P	R	F1
HT-HUS_1	**62.74**	72.33	84.05	40.43	78.76	57.90	64.89	80.49	63.82	71.19
HT-HUS_2	60.70	68.08	**84.35**	44.37	78.17	57.07	64.37	78.68	61.91	69.30
HT-HUS_3	62.50	**74.60**	82.87	44.67	73.54	62.34	**66.16**	76.17	68.37	**72.06**
MinhPQN_1	61.04	65.87	80.77	43.37	72.21	56.78	62.76	75.63	60.08	66.96
MinhPQN_2	62.41	66.38	81.00	43.87	73.32	57.09	63.42	76.83	60.28	67.56
MinhPQN_3	60.40	64.68	80.14	**46.56**	74.36	55.94	62.94	76.87	58.52	66.45
SunBear_1	59.74	67.54	79.57	41.50	58.44	66.75	62.09	60.82	73.29	66.48
SunBear_2	54.43	68.10	76.33	38.83	55.39	64.15	59.42	59.69	70.08	64.47
SunBear_3	49.29	62.10	71.52	31.24	53.11	55.27	53.54	55.91	59.16	57.49
VC-TUS_1	46.37	56.21	74.11	28.68	75.92	40.18	51.34	80.29	44.38	57.16
VC-TUS_2	55.23	57.87	79.70	39.16	80.38	46.43	57.99	83.51	49.09	61.84
VC-TUS_3	54.67	56.96	79.12	38.87	80.83	45.76	57.40	83.38	48.38	61.23

Results are reported in %. Highest result in each column is highlighted in bold.
LOC: LOCATED, AFF: ORGANIZATION-AFFILIATION,
P-W: PART-WHOLE, P-S: PERSONAL-SOCIAL.

Table 5: Detailed results of all submissions.

caused by annotation errors, most of them are made up of examples illustrating the limits of current approaches. We need a more in-depth survey on linguistic patterns and knowledge, as well as more complex reasoning techniques to resolve these cases. A case in point: *"Đừng quên trong tay của HLV Tom Thibodeau vẫn còn đó bộ 3 ngôi sao Karl-Anthony Towns – Andrew Wiggins – Jimmy Buter*." (ID 24527838). In this instance, *[Tom Thibodeau]* is participated in three $PERSONAL - SOCIAL$ relations with *[Karl-Anthony Towns]*, *[Andrew Wiggins]*, and *[Jimmy Buter]*. In which, two relations of *[Tom Thibodeau] - [Andrew Wiggins]* and *[Tom Thibodeau] - [Jimmy Buter]* were not predicted by any team, probably on account of their complex semantics presenting with a conjunction. Another example: *[Hassan được cho là người Iraq , được một cặp vợ chồng người Anh nhận làm con nuôi và cùng sinh sống tại Sunbury , vùng ngoại ô London]* (ID 23352918). Instance *[Hassan] - [Sunbury]* of $LOCATED$ relation is misclassified either as $ORGANIZATION - AFFILIATION$ or as no relation.

Figure 5 gives statistics on how many instances are correctly found in 1 to 11 out of 12 submissions. It shows that the proposed systems of participated teams produce multiple inconsistent results. It also notes the difficulty of the challenge and data.

5 Conclusions

The RelEx task was designed to compare different semantic relation classification approaches and provide a standard testbed for future research. The RelEx dataset constructed in this task is expected to make significant contributions to the other related researches. RelEx challenge is an endorsement of machine learning methods based on deep neural networks. The participated teams have achieved some exciting and potential results. However, the deeper analysis also shows some performance limitations, especially in the case of semantic relations presented in a complex linguistic structure. This observation raises some research problems for future works. Finally, we conclude that the RelEx shared task was run successfully and is expected to contribute significantly to Vietnamese text mining and natural language processing communities.

Acknowledgments

This work was supported by the Vingroup Innovation Foundation (VINIF) under the project code DA137_15062019/year 2019. The shared task committee would like to grateful DAGORAS data technology JSC. for their technical and financial support, and the six annotators for their hardworking to support the shared task.

References

Charu C Aggarwal. 2015. Mining text data. In *data mining*, pages 429–455. Springer.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*, 2:1–15.

Paulo R Cavalin, Fillipe Dornelas, and Sérgio MS da Cruz. 2016. Classification of life events on social

media. In *29th SIBGRAPI (Conference on Graphics, Patterns and Images)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Hercules Dalianis. 2018. Evaluation metrics and evaluation. In *Clinical Text Mining*, pages 45–53. Springer.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 358–365.

Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessieres, and Claire Nédellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 12–22.

Li Dongmei, Zhang Yang, Li Dongyuan, and Lin Danqiong. 2020. Review of entity relation extraction methods. *Journal of Computer Research and Development*, 57(7):1424.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1997–2007.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*, pages 1211–1220. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1037–1042.

Viet Bui The, Oanh Tran Thi, and Phuong Le-Hong. 2020. Improving sequence tagging for vietnamese text using transformer-based neural models. *arXiv preprint arXiv:2006.15994*.

Anjali Thukral, Ayush Jain, Mudit Aggarwal, and Mehul Sharma. 2018. Semi-automatic ontology builder based on relation extraction from textual data. In *Advanced Computational and Communication Paradigms*, pages 343–350. Springer.

Xuan-Son Vu. 2016. Pre-trained word2vec models for vietnamese. https://github.com/sonvx/word2vecVN.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.