AUTOMATIC IDENTIFICATION OF LIGHT VERB CONSTRUCTIONS: A REVIEW

(Date received: 13.01.2022/Date accepted: 03.04.2022)

Kathleen Swee Neo Tan[1*], Tong Ming Lim[2], Chi Wee Tan[1], Wei Wei Chew[3]

[1] *Faculty of Computing and Information Technology*

[2] *Centre for Business Incubation and Entrepreneurial Ventures*

[3] *Faculty of Social Science and Humanities, Tunku Abdul Rahman University College,*
Kampus Utama, Jalan Genting Kelang, 53300 Wilayah Persekutuan Kuala Lumpur, Malaysia

**Corresponding author: tansn@tarc.edu.my*

ABSTRACT

Light verb constructions (LVC) are complex predicates that are present in many languages. They belong to the Multiword Expression (MWE) category known as verbal MWEs and has the canonical form of verb+noun. Examples of LVCs include give help, make decisions, and take walks. LVC identification is essential for many natural processing (NLP) applications such as machine translation, sentiment analysis, and information extraction. However, the task of LVC identification is challenging due to its characteristics such as variability, discontinuity, and ambiguity. This paper presents a review of recent work, discusses the gaps that still exist, and proposes some future work that may contribute significant progress in LVC identification.

Keywords: light verb constructions, multiword expressions, computational linguistics, natural language processing

1.0 INTRODUCTION

Multiword Expressions (MWE) are expressions that contain two or more words that are used together to convey a certain meaning. MWE identification is important for NLP tools such as part-of-speech (POS) taggers, semantic parsers and syntactic parsers, as well as downstream applications such as machine translation, emotion analysis, and question answering systems (Constant *et al.,* 2017). Recently, there has been a growing interest in Light Verb Constructions (LVC), which is a type of verbal MWE (Cordeiro & Candito, 2019; Nagy T. *et al.,* 2020). LVCs are complex predicates that have the canonical form of verb+noun. Examples of LVCs include *give help, make decisions,* and *take walks*. There are two particularly interesting characteristics of LVCs. Firstly, the verbal component of the LVC does not contribute much to the meaning of the LVC and are not interpreted in the literal sense. Consider the ‘heavy’ usages or literal meaning of the verbs used in these non-LVC examples: make implies the act of creating something (as in *make a cake*), take is an action that results in the possession of the object (e.g., *take the plate from the cupboard*), and give is the act of transferring an object to be in the possession of a subject (e.g., *gave a bouquet of flowers*). The heavy usage of these verbs is also known as *productive verbs* as their use indicate that they effect or produce some results (e.g., the creation of an object or transfer of possession). Secondly, although it is more efficient to use the synthetic verb counterpart in a sentence (e.g., the LVC *‘make a review’* can actually be more efficiently replaced by its synthetic verb counterpart ‘review’), there is a greater tendency to use the LVC instead due to the ease in which LVCs may be modified for greater expressiveness (Bonial & Pollard, 2020). Table 1 illustrates the use of several LVCs and their synthetic verb counterparts in sentences.

Table 1: Examples of the use of LVCs and synthetic verbs in sentences

With LVC	With synthetic verb counterpart
We *give a review* of light verb constructions in computational linguistics.	We *review* light verb constructions in computational linguistics.
Jon *took a brisk walk* around the college this morning.	Jon *walked briskly* around the college this morning.
They will *make a decision* on the new product next week.	They will *decide* on the new product next week.

The interest in LVCs has been demonstrated by work in the field of linguistics (Bonial & Pollard, 2020; Gilquin, 2019; Ong & Rahim, 2021) as well as computational linguistics in various languages (Klyueva *et al.,* 2017; Maldonado *et al.,* 2017; Nagy T. *et al.,* 2020). The importance of LVC identification in NLP applications can be observed from recent work in the development of multilingual annotated corpora for the automatic identification of verbal MWEs which includes LVCs (Ramisch *et al.,* 2018, 2020; Savary *et al.,* 2017).

As an extension of our previous work (Tan *et al.,* 2021b), the two main approaches in which the LVC identification task may be framed and the types of evaluation used are presented. The remainder of this paper is organized as follows. In the next section, we present recent work related to LVC identification, followed by a discussion on observed gaps and future work that could bring novel contributions. The final section concludes the paper.

2.0 LVC IDENTIFICATION METHODS

LVC identification is the task of automatically detecting instances of LVCs in running text. The task of LVC identification is complicated by the fact that LVCs have a number of challenging characteristics which includes discontinuity (i.e. gaps between the verbal and nominal parts of the LVC such as she ***gave*** *five interesting* ***lectures***), variability (e.g., the passive form the ***lecture*** *was* ***given***), and ambiguity (e.g., the phrase *they will make the decision known to the employees* does not contain any LVCs) (Savary *et al.,* 2017).

The LVC identification task may be addressed as either a classification or a sequence labeling task. For the evaluation of all MWE-types, Savary *et al*. (2017) differentiated between MWE-based evaluation and token-based evaluation. MWE-based evaluation implements strict matching in which all components of an MWE have to be correctly predicted whereas for token-based evaluation, any correctly predicted component of the MWE is counted. However, not all papers explicitly indicate the type of evaluation (i.e., MWE- or token-based) that was used. In addition, some papers provide the per-language evaluation scores while others report either the micro-average or macro-average scores across all languages. A summary and comparison of the studies on the automatic identification of LVCs in the recent years are shown in Table 2, while a chronological view of the work is shown in Figure 1. For the year 2021, the related work on LVC had been on aspects supporting the LVC identification task such as the development of annotation guidelines for LVC (Bonial, 2021), the investigation of properties for the aspectual variant of LVCs (Fotopoulou *et al.,* 2021) and the study of systematic patterns for LVC families (Fleischhauer, 2021).

2.1 Classification-Based Approaches

When LVC identification is framed as a classification task, there are two main steps that need to be carried out. Firstly, LVC candidates have to be extracted. Secondly, binary classification of the extracted LVC candidates is performed using machine learning algorithms.

In the system proposed by Waszczuk (2018), a dependency tree was constructed for each sentence. For each node in the tree, the system predicted whether or not that node was an LVC based on local contextual information which included word forms, POS tags, dependency labels, lemmas, and so on. Next, segmentation to determine the LVC boundaries was carried out by constructing a hypergraph which represented all traversals of the dependency tree. The hypergraph contained a distinct hyperpath for each traversal. Using features extracted from each traversal's hyperpath, a multiclass logistic regression model was then used to determine the hyperpath with the highest probability to find the globally optimal labeling for the given dependency tree. As future work, they suggested improving the LVC segmentation part by considering the incorporation of lexicons and word embeddings in their system.

Cordeiro & Candito (2019) extracted LVC candidates based on syntactic patterns that considered LVC variations such as morphosyntactic variations, complex nominal components, and other language-specific characteristics. By extracting language-specific morphosyntactic patterns comprising the POS tag and syntactic relation between components from the LVCs present in the training dataset, the frequently occurring patterns were used to identify LVC candidates in the datasets. Binary classification of the LVC candidates was then carried out using support vector machine (SVM) and feed-forward neural network (FFN). They suggested to investigate the use of contextualized word embeddings as future work.

Nagy T. *et al*. (2020) used a decision-tree to perform classification of LVC candidates. First, dependency parsers were used to produce dependency representations of the corpus. LVC candidates were then extracted from raw text based on the syntactic relations between the verbal and nominal components of LVCs. A rich feature set that included both language-independent (i.e., statistical, lexical, morphological, syntactic, and orthographic features) and language-dependent (auxiliary verbs, gender, and agglutinative morphology) features were constructed. They found that the performance of their method depended on the quality of the dependency parsers used for LVC candidate extraction.

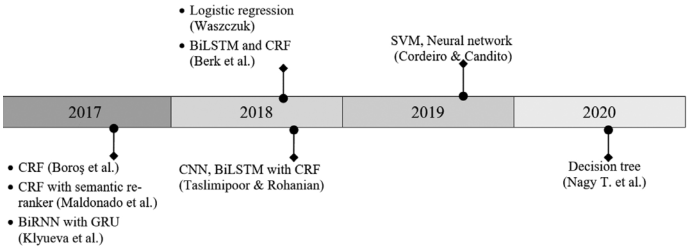
• Logistic regression (Waszczuk)
• BiLSTM and CRF (Berk et al.)
SVM, Neural network (Cordeiro & Candito)
2017
2018
2019
2020
• CRF (Boroş et al.)
• CRF with semantic re-ranker (Maldonado et al.)
• BiRNN with GRU (Klyueva et al.)
CNN, BiLSTM with CRF (Taslimipoor & Rohanian)
Decision tree (Nagy T. et al.)

Figure 1: Chronological view of work on automatic LVC identification

2.2 Sequence Labeling Approaches

Some studies approached LVC identification as a sequence labeling problem where each word in a sentence is tagged using the Begin-Inside-Outside (BIO) tagging scheme (Ramshaw & Marcus, 1999). A word would be tagged with *B* to indicate that it is the beginning word of the LVC, *I* if the word is a component word of the LVC (i.e., inside the LVC), and *O* if the word is outside of any LVCs. The BIO tagging scheme has the benefit of being able to handle discontinuous LVCs which have components that are not adjacent to each other such as *the* ***decision*** *that was* ***made***. The use of the BIO tagging scheme is illustrated in Table 3.

Boroş *et al*. (2017) used Conditional Random Fields (CRF) to predict the transition between labels. For each word in a sentence, the lemma and POS tags for the window of words surrounding the current word served as the features. The identification of LVCs involved two steps. First, head labeling to identify the head word of the LVC (in this case the light verb) was carried out using a window size of 2. The second step was tail labeling to identify that nominal component of the LVC using a window size of 4. The authors found that compared to the single-step detection of LVC, their two-step approach increased precision by 9%.

Table 2: Summary of Papers on LVC Identification

Author(s)/Year	Approach	Method	Languages	F1-scores
Boroş *et al.* (2017)	Sequence labeling	CRF	CS, DE, EL, ES, FR, HU, IT, MT, RO, SL, SV, TR	MWE-based: 5.84% (MT) – 86.27% (RO)
Maldonado *et al.* (2017)	Sequence labeling	CRF with semantic re-ranker (SEM)	CS, DE, EL, ES, FR, HU, IT, MT, PL, PT, RO, SL, SV, TR	CRF MWE-based: 1.22% (SL) – 46.24% (PT) CRF Token-based: 4.30% (SL) – 57.08% (PT) CRF with SEM MWE-based: 1.19% (SL) – 52.67% (PT) CRF with SEM Token-based: 3.97% (SL) – 56.83% (PT)
Klyueva *et al.* (2017)	Sequence labeling	BiRNN with GRU	BG, CS, DE, EL, ES, FR, HE, HU, PL, PT, RO, SL, TR	MWE-based: 0% (BG, DE, SL) – 37% (PT) Token-based: 1% (SL) – 49% (PT)
Waszczuk (2018)	Classification	Logistic regression	BG, DE, EL, EN, ES, EU, FA, FR, HE, HI, HR, HU, IT, LT, PL, PT, RO, SL, TR	MWE-based: mAvg: 46.03
Berk *et al.* (2018)	Sequence labeling	BiLSTM and CRF	BG, DE, ES, FR, HU, IT, PL, PT, RO, SL	MWE-based: 28.55% (ES) – 74.48% (HU) Token-based: 35.66% (ES) – 81.86% (RO)
Taslimipoor & Rohanian (2018)	Sequence labeling	Convolutional, BiLSTM with CRF	BG, DE, EL, EN, ES, EU, FA, FR, HE, HI, HR, HU, IT, LT, PL, PT, RO, SL, TR	MWE-based: 6.25% (EN) – 86.15% (RO) Token-based: 11.25% (EN) – 87.41% (RO)
Cordeiro & Candito (2019)	Classification	SVM, Neural network	BG, DE, EL, EN, ES, EU, FA, FR, HE, HI, HR, HU, IT, LT, PL, PT, RO, SL, TR	MWE-based: SVM: 26% (EN) – 81 % (HU), µAvg = 63% FFN: 21% (HE) – 78% (HI), µAvg = 56%
Nagy T. *et al.* (2020)	Classification	Decision tree	DE, EN, ES, HU	50.64% (DE), 52.90 (ES), 64.72% (HU), 65.35% (EN)

F1-scores: mAvg-macro-average score across all languages; µAvg: micro-average score across all languages.
For papers with more than 4 languages, only the lowest and highest F1-scores are indicated for brevity.
Some papers did not indicate whether the evaluation was MWE-based or token-based, and some papers only reported the average F1-score across all languages.

Language codes: BG-Bulgarian, CS-Czech, DE-German, EL-Greek, EN-English, ES-Spanish, EU-Basque, FA-Farsi, FR-French, HE-Hebrew, HI-Hindi, HR-Croatian, HU-Hungarian, IT-Italian, LT-Lithuanian, MT-Maltese, PL-Polish, PT-Portuguese, RO-Romanian, SL-Slovene, SV-Swedish, TR-Turkish

Table 3: Example of a sentence with BIO tags identifying an LVC

Sentence	He will make a very difficult decision later						
Word	He	will	**make**	a	very	**difficult**	decision	later
Tag	*O*	*O*	*B*	*O*	*O*	*O*	*I*	*O*

Maldonado *et al*. (2017) used a CRF model which exploited syntactic dependency features and also included an optional semantic re-ranker as a post-processing step. Instead of developing feature sets for each of the 14 languages, the authors created a feature set for each language family based on the assumption that the morphosyntactic relationships among closely related languages would be similar. Experiments on semantic re-ranking using a regression model trained on semantic vectors was conducted for 12 languages and showed improvement in 7 of the languages. For future work, the authors plan to focus on language-specific features. In addition, they suggested looking into word embeddings as a possible way to improve the performance of the model.

Berk *et al*. (2018) proposed a bidirectional Long Short-Term Memory (LSTM)-CRF model in which the inputs to the model consisted of the POS tags and dependency relation tags. In the BiLSTM layer, the forward LSTM unit enabled the previous words to be used as features whereas the backward LSTM unit enabled the future words to be used as features. The CRF layer enabled the decoding of the sequence labels using the gappy, 1-level variant of the BIO tagging scheme proposed by Schneider *et al*. (2014).

Taslimipoor & Rohanian (2018) proposed a model comprising two convolutional layers that serve as n-gram detectors, a BiLSTM for handling long distance relationships between words, and an optional CRF layer to process dependencies among the output tags. They used pretrained Wikipedia word embeddings (Bojanowski *et al.,* 2017) and binary word shape features to indicate whether the token started with an uppercase letter, was entirely in uppercase, had a # or @ as the first character, was a URL, contained a number, or was a digit. They reported that the pre-trained embeddings achieved the best performance and that the additional CRF layer did not necessarily improve the performance of the model.

The model proposed by Klyueva *et al*. (2017) was based on a bidirectional recurrent neural network (RNN) with gated-recurrent units (GRUs) that was trained using linguistic, morphological and syntactic features. Each input word was represented as a concatenation of embeddings of the word's form, lemma, and POS tag. They found that discontinuous LVCs were often not tagged. Their model did not consider embedded or overlapping LVCs.

3.0 DISCUSSION

Despite significant progress in LVC identification, there are several gaps yet to be addressed.

Large variation in the evaluation results for different languages: Table 2 shows that there is a large difference between the lowest and highest per-language F1-scores for almost all the papers that reported per-language results for the LVC identification task. One reason for this was the training corpus size differences for the languages (Berk *et al.,* 2018; Cordeiro & Candito, 2019) - a larger training corpus would inevitably result in better results. Secondly, the training corpus for certain languages had lower average occurrences of LVCs (Waszczuk, 2018). This means that the machine learning models would encounter fewer examples of LVCs, which may be a hindrance to the learning process (Berk *et al.,* 2018; Nagy T. *et al.,* 2020). Thirdly, languages with dependency parsing tools of lower quality would inadvertently affect the quality of predictions. The difference in results across languages may also be due to the distribution of *seen* and *unseen LVCs* in the training and evaluation datasets. *Seen LVCs* refer to LVCs that occur at least once in the training dataset while those that are *unseen* were not present in the training dataset at all but appeared in the evaluation dataset (Cordeiro & Candito, 2019). Therefore, one important future work would be to explore methods for improving predictions for unseen LVCs.

Lack of tools and resources for under-resourced languages: LVCs may appear in a variety of forms due to morphosyntactic variations – for example, the nominal part of the LVC may be a complex noun phrase instead of a single noun or may even be further discontinuous whereby the verbal and nominal components are separated by many words (e.g., *the **walk** in the beautiful and dense rainforest that he had **taken***). This makes it particularly challenging especially because most existing work rely on the use of POS-taggers and dependency parsers to capture morphosyntactic variations. For under-resourced languages, the lack of such tools and LVC-annotated corpora provides the motivation to explore methods for LVC identification that do not depend on the use of such tools or that require a smaller annotated corpus.

Code-mixed LVCs: One growing challenge is the identification of code-mixed LVCs which are commonly used in social media. Code-mixed text includes words from two or more languages – an increasingly common phenomenon in recent times as more people are bilingual. There has been a growing interest in sentiment analysis of code-mixed text (Lo *et al.,* 2017; Sasidhar *et al.,* 2020; Wang *et al.,* 2017) and even in Malay-English code-mixed text (Abu Bakar *et al.,* 2020; Tan *et al.,* 2020). The impact of code-mixed MWE identification on the emotion detection task (Tan *et al.,* 2021a) and use of code-mixed LVCs (Alexiadou, 2017; González-Vilbazo & López, 2011) have also been the focus of research. To our knowledge, there has not been work on code-mixed LVC identification and therefore, this needs to be addressed to enable LVCs to be treated as a single semantic unit and avoid loss of contextual meaning that arise from the individual words of the LVC being considered as separate features.

Use of word embeddings: To overcome the problems faced in resource-poor languages and code-mixed LVC identification, the use of word embeddings should be further explored. In particular, the use of word embeddings with character n-grams to represent out-of-vocabulary (OOV) words (Bojanowski *et al.,* 2017) which are often either intentionally/unintentionally misspelt words or slang words that are prevalent in social media can potentially help as they are able to reflect the semantic relationships between words. Two interesting ideas to consider is to train word embeddings using a code-mix corpus as was done for POS tagging by Bhattu *et al*. (2020), and to include LVCs and other MWEs as single tokens in embeddings.

4.0 CONCLUSION

LVCs are particularly challenging to identify due to their flexibility, ambiguity, and discontinuity. This study explored recent trends in LVC identification, a task which plays an important role in downstream text processing tasks such as emotion analysis, machine translation, and question answer systems. In addition, existing gaps were discussed, and several promising future works were identified including investigation of methods that do not require the use of dependency parsers and POS taggers, training embeddings that represent each LVC as a single token, as well as the use of code-mixed word embeddings to facilitate the identification of code-mixed LVCs. ■

REFERENCES

[1] Abu Bakar, M. F. R., Idris, N., Shuib, L., & Khamis, N. (2020). Sentiment Analysis of Noisy Malay Text: State of Art, Challenges and Future Work. IEEE Access, 8, 24687–24696. https://doi.org/10.1109/ACCESS.2020.2968955

[2] Alexiadou, A. (2017). Building verbs in language mixing varieties. Zeitschrift Fur Sprachwissenschaft, 36(1), 165–192. https://doi.org/10.1515/zfs-2017-0008

[3] Berk, G., Erden, B., & Güngör, T. (2018). Deep-BGT at PARSEME Shared Task 2018: Bidirectional LSTM-CRF Model for Verbal Multiword Expression Identification. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), 248–253.

[4] Bhattu, S. N., Nunna, S. K., Somayajulu, D. V. L. N., & Pradhan, B. (2020). Improving Code-mixed POS Tagging Using Code-mixed Embeddings. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(4). https://doi.org/10.1145/3380967

[5] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[6] Bonial, C. (2021). Précis of Take a Look at This! Form, Function, and Productivity of English Light Verb Constructions. Colorado Research in Linguistics, 209. https://search.proquest.com/docview/1651531484?accountid=13375

[7] Bonial, C., & Pollard, K. A. (2020). Choosing an event description: What a PropBank study reveals about the contrast between light verb constructions and counterpart synthetic verbs. Journal of Linguistics, 56, 577–600. https://doi.org/10.1017/S0022226720000109

[8] Boroş, T., Pipa, S., Mititelu, V. B., & Tufiş, D. (2017). A data-driven approach to verbal multiword expression detection. PARSEME Shared Task system description paper. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), 121–126. https://doi.org/10.18653/v1/w17-1716

[9] Constant, M., Eryigit, G., Monti, J., Plas, L. van der, Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword Expression Processing: A Survey. Computational Linguistics - Association for Computational Linguistics, 43(4), 837–892. https://doi.org/10.1162/COLI

[10] Cordeiro, S. R., & Candito, M. (2019). Syntax-based identification of light-verb constructions. The 22nd Nordic Conference on Computational Linguistics (NoDaLiDa 2019).

[11] Fleischhauer, J. (2021). Light Verb Constructions and Their Families - A Corpus Study on German stehen unter-LVCs. Proceedings of the 17th Workshop on Multiword Expressions, 63–69. https://doi.org/10.18653/v1/2021.mwe-1.8

[12] Fotopoulou, A., Laporte, E., & Nakamura, T. (2021). Where Do Aspectual Variants of Light Verb Constructions Belong? Proceedings of the 17th Workshop on Multiword Expressions, 2–12. https://doi.org/10.18653/v1/2021.mwe-1.2

[13] Gilquin, G. (2019). Light verb constructions in spoken L2 English: An exploratory cross-sectional study. International Journal of Learner Corpus Research, 5(2), 181–206. https://doi.org/10.1075/ijlcr.18003.gil

[14] González-Vilbazo, K., & López, L. (2011). Some properties of light verbs in code-switching. Lingua, 121, 832–850. https://doi.org/10.1016/j.lingua.2010.11.011

[15] Klyueva, N., Doucet, A., & Straka, M. (2017). Neural Networks for Multi-Word Expression Detection. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), 60–65.

[16] Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. Artificial Intelligence Review, 48, 499–527. https://doi.org/10.1007/s10462-016-9508-4

[17] Maldonado, A., Han, L., Moreau, E., Alsulaimani, A., Chowdhury, K. D., Vogel, C., & Liu, Q. (2017). Detection of Verbal Multi-Word Expressions via Conditional Random Fields with Syntactic Dependency Features and Semantic Re-Ranking. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), 114–120. https://doi.org/10.18653/v1/w17-1715

[18] Nagy T., I., Rácz, A., & Vincze, V. (2020). Detecting light verb constructions across languages. Natural Language Engineering, 26, 319–348. https://doi.org/10.1017/S1351324919000330

[19] Ong, C. S. B., & Rahim, H. A. (2021). Nativised structural patterns of make light verb construction in Malaysian English. Concentric, 47(1), 93–112. https://doi.org/10.1075/consl.00024.rah

[20] Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Mititelu, V. B., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Escartín, C. P., … Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG 2018), 222–240. https://aclanthology.org/W18-4925

[21] Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020). Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, 107–118. https://aclanthology.org/2020.mwe-1.14

[22] Ramshaw, L. A., & Marcus, M. P. (1999). Text Chunking Using Transformation-Based Learning. Proceedings of the Third ACL Workshop on Very Large Corpora, 157–176. https://doi.org/10.1007/978-94-017-2390-9_10

[23] Sasidhar, T. T., Premjith, B., & Soman, K. P. (2020). Emotion Detection in Hinglish(Hindi+English) Code-Mixed Social Media Text. Procedia Computer Science, 171, 1346–1352. https://doi.org/10.1016/j.procs.2020.04.144

[24] Savary, A., Ramisch, C., Cordeiro, S. R., Sangati, F., Vincze, V., Qasemizadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), 31–47. http://multiword.sf.net/

[25] Schneider, N., Danchik, E., Dyer, C., & Smith, N. A. (2014). Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. Transactions of the Association for Computational Linguistics, 2, 193–206. https://doi.org/10.1162/tacl_a_00176

[26] Tan, K. S. N., Lim, T. M., & Lim, Y. M. (2020). Emotion Analysis Using Self-Training on Malaysian Code-Mixed Twitter Data. Web Based Communities and Social Media 2020, 181–188. https://www.elearning-conf.org/wp-content/uploads/2020/07/01_202008L022_F048.pdf

[27] Tan, K. S. N., Lim, T. M., & Tan, C. W. (2021a). A Study on Multiword Expression Features in Emotion Detection of Code-Mixed Twitter Data. 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 1–5. https://doi.org/10.1109/IICAIET51634.2021.9573850

[28] Tan, K. S. N., Lim, T. M., Tan, C. W., & Chew, W. W. (2021b). Review on Light Verb Constructions in Computational Linguistics. International Conference on Digital Transformation and Applications 2021 (ICDXA 2021), 153–160.

[29] Taslimipoor, S., & Rohanian, O. (2018). SHOMA at Parseme Shared Task on Automatic Identification of VMWEs: Neural Multiword Expression Tagging with High Generalisation. ArXiv Preprint ArXiv:1809.03056. http://arxiv.org/abs/1809.03056

[30] Wang, Z., Lee, S. Y. M., Li, S., & Zhou, G. (2017). Emotion Analysis in Code-Switching Text with Joint Factor Graph Model. IEEE/ACM Transactions on Audio Speech and Language Processing, 25(3), 469–480. https://doi.org/10.1109/TASLP.2016.2637280

[31] Waszczuk, J. (2018). TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model. Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018).

PROFILES

KATHLEEN TAN SWEE NEO received her BS (Computer Science and Physics) in 1992 from Campbell University, US and MSc in Management of Information Technology in 2003 from University of Sunderland, UK. She is a principal lecturer in the Department of Computer Science and Embedded Systems at TAR UC and is currently a PhD candidate in the Faculty of Computing and Information Technology, TAR UC. Her research interests include multiword expression identification and emotion analysis on code-mixed social media text, as well as big data analytics.
Email address: tansn@tarc.edu.my

DR TAN CHI WEE received BCompSc(Hons) and PhD degrees in year 2013 and 2019 respectively in Universiti Teknologi Malaysia. Currently, he is a Senior Lecturer cum Programme Leader at Tunku Abdul Rahman University College and actively involved in the Centre of Excellence for Big Data and Artificial Intelligent (CoE) and become the research group leader for Audio, Image and Video Analytics Group under Centre for Data Science and Analytics (CDSA). Dr Tan's main research areas are Computer Vision (CV), Image Processing (IP) and Natural Language Processing (NLP) and Artificial Intelligence (AI). He is an enthusiastic researcher experienced in conducting and supporting research into Image Processing. Being a meticulous and analytical researcher with Train-The-Trainer certificate of many years of educational and hands-on experience, he was invited to Université d'Artois (France) under Marie Skłodowska-Curie Research and Innovation Staff Exchange (RISE) programme for collaborative research between European countries with Southeast Asian countries on motion detection and computer vision and being involved in industry project as professional consultant.
Email address: chiwee@tarc.edu.my

PROFESSOR LIM has about 10 years of industry experiences in the design, development, implementation and maintenance of commercial software from 1989 to 1999 after departing from TARC where he spent his early days with TARC as an IT lecturer from 1987 to 1989 after returning from Mississippi State University USA with a Master of Computer Science degree. He is currently the Director for CBIEV at TAR UC, Professor at FOCS at TAR UC and Head for Big Data Analytics Centre. His research interest involving Natural Language Processing, Sentiment Analysis and Code-Mixed language analysis. In the last 15 years, his work has consistently focused on organizational knowledge sharing and technology acceptance, social media analytics and social influence maximization in Sunway University and Tunku Abdul Rahman University College (TAR UC). Professor Lim has graduated more than 20 master and 2 PhD students while he was with Monash, UTAR and Sunway University.
Email address: limtm@tarc.edu.my

THE LATE DR CHEW WEI WEI obtained her BA (Honours) in Malay Studies in 1997 and Diploma in Translation in 2000 from the University of Malaya, and her MA in Translation in 2005 and PhD in English Language Skills in 2014 from Universiti Sains Malaysia. Besides her 20 years of teaching experience, she was involved in research focused on multilingualism and the use of code-mixed Malay-English in social media. In addition, she was a Course Leader for Nation Building and Languages (2013 - 2016), Chairperson for the Centre for Social Integration and Social Skills Research (2017-2019) as well as co-author for the textbooks Pendidikan Moral (2011) and Civic Consciousness and Volunteerism (2019) at Tunku Abdul Rahman University College.