Rule-based Text Extraction for Multimodal Knowledge Graph

Idza Aisara Norabid, Fariza Fauzi
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia

***Abstract*—Textual information is widely integrated in visual tasks such as object/scene detection and image annotation. However, the textual information is not fully exploited, overlooking the wide background knowledge available for Web images. This work proposes a multimodal knowledge graph (KG) to represent the knowledge extracted from unstructured Web image surrounding text and to integrate the relationship between image and text entities. Existing multimodal KG works have mainly focused on advanced visual processes for extracting entities and relations from images, and only employed standard text processing techniques such as tokenization, stop word removal, and part-of-speech (POS) tagging to capture nouns only or basic subject-verb-object from text in the semantic enrichment process. Adversely, neglecting other rich information in the text. Thus, the proposed approach attempts to address this as an automatic relation extraction (RE) problem to extract all possible triples from the text information from simple to complex sentences, in constructing the multimodal KG which eventually can be used as a training seed for visual tasks. A linguistic analysis is performed on a set of Web news articles consisting of news images and their related text. The dependency relations and POS information obtained are used to formulate a set of domain-agnostic entity-relation extraction rules. A triple extractor incorporating these rules, is developed to extract the triples from a news articles dataset and construct the proposed MKG. The Precision and Recall metrics are used to evaluate the extractor's performance. The evaluation results show that the proposed approach can extract entities and relations in the dataset with the precision score of 0.90 and recall score of 0.60. While the results are promising, the extraction rules can still be improved to capture all the knowledge.**

Keywords*—Relation extraction; knowledge graph; multimodal knowledge graph; dependency relations; object/scene detection

I. INTRODUCTION

The web consists of valuable images with surrounding textual information (also referred to as contextual information) that have been used in various computer vision tasks such as object/scene detection, image annotation, image clustering, image understanding, etc. This information is in which the surrounding texts are related to the image, rich in high-level semantic concepts and contains both direct and indirect information about the image [1-2].

While contextual information has been long used to improve computer vision task performance, there is still opportunity for improvement. According to [3], there is a gap between textual and visual analysis for image understanding. Existing algorithms, such as deep neural networks, mainly focus on specific features in the image itself, typically ignoring the extensive background knowledge of the real world [4] that can be found in the contextual information.

In [3], [5] and [6], both visual and text are merged in the analysis for object detection and image retrieval, but these works focused more on the visual part rather than the text. The descriptions used are obtained from expert, hence, the image-text correlation is very high, however annotations from experts have the disadvantage of being time consuming and also expensive.

Many textual descriptions contain extensive background knowledge especially in news articles and blogs. The news carries variety of domain such as sport, education, entertainment and more. This implies that the knowledge is not restricted to a single area, but rather covers a broad range of topics. Hollink et al. [7] also states that news articles have a natural relationship between text and images. Works of [7] and [8] have proven the importance of images to be in their natural textual context, where they created corpus from online news articles. Typically, a news article will include a headline, the article and an image (or more) that is relevant to the news. Also, the image comes with a description (i.e., caption), where the description serves as context for the image and provides knowledge about it. By simply looking at the image, one may not fully comprehend the scene depicted. The thorough explanation of the scene is given in the description as well as in the headline and article, as illustrated in Fig. 1. Compared to manual captions from MSCOCO dataset in Fig. 2, the texts in these news articles are rich with high-level knowledge (abstract rather than object-level semantics), relevant to the article images, and readily available.

Fig. 1. An Image and its Headline and Caption from a News Article.

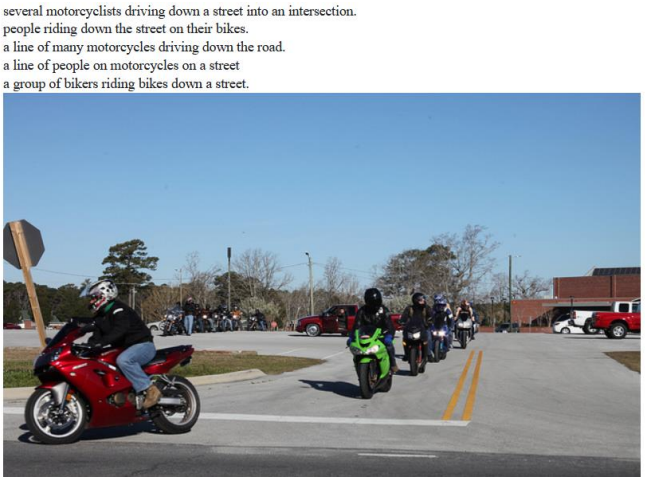


Fig. 2. A Sample Image and its Captions from MSCOCO.

If the background knowledge in the Web news articles can be acquired in a structured manner, then a knowledge base can be generated, which can then be applied in visual tasks. Pan et al. [9] worked on methods that can improve the task of question answering (QA) and similarly, Wu et al. [10] proposed method to improve performance for answering visual questions. Both studies prove that it is important to incorporate external knowledge into machines to propose how humans handle such tasks. However, this vast knowledge that relates to the images is in the unstructured textual form. Thus, the suitable technique for extracting knowledge in a more structured way from large amounts of text and images and for describing the diversity of entities and concepts that exist in the real-world is required.

Li et al. [11] propose knowledge graph to learn knowledge for social image understanding. Knowledge graphs (KG) have long been used to represent knowledge and real-world events as graphs with nodes (entities), edges (relations) and labels. A KG is a data structure capable of capturing real-world concepts/entities and their relationships from unstructured text, transforming them into a structured graph. It reveals relationships between entities found in the text. Two entities and the relation between them are called triple (i.e., the basic building block for KG).

Gong and Wang [12] have produced a KG in their work where the graph is a multimodal KG (i.e., a KG consisting of text and images). The method used still has room for improvement as it only captures nouns or noun phrases, ignoring other rich information such as the verbs or the adjectives that explain about the image, hence, not fully utilizing the vast image background knowledge that can be acquired from the unstructured text.

To build a KG from text, the task of relation extraction (RE) is applied to extract the relationship. The development of this relation extraction algorithm has several techniques. The two general categories are rule-based and machine learning-based [13]. Rule-based relation extraction is performed by using linguistic knowledge and domain knowledge to build pattern based on words, parts of speech (part-of-speech) or semantics in collaboration with domain experts and then the relationship is extracted according to the set of rules. Several works have used this approach in achieving their objectives [14,15,16].

Next, machine learning-based relation extraction methods use large amounts of labeled data for training and have shown good results in some instances [13]. There are also several open-source information extraction systems that use this approach such as NELL [17] and OLLIE [18]. However, problems can arise, such as lack of training data and poor generalized performance. Thus, a rule-based relation extraction is the best option for extracting relationships as it does not require prior training data and the set rules will extract a more overall result.

In conclusion, the issues discovered are that unstructured background knowledge (text) needs to be organized in a structured way while still maintaining the natural image-text relationship found in news articles. Second, the use of simple text (nouns) in the construction of multimodal knowledge graphs causes the overlook of other rich information available from the text. Therefore, the aim of this work is to build a multimodal knowledge graph by using the images accompanied by textual information in news articles, as well as linguistic-based relation extraction techniques to overcome the stated problems.

This paper contributes to the development of a set of rules for extracting entities and relationships from text and image in news articles. The rules are based on linguistic analysis of simple to complex sentences, factoring in the image from the news article. Grammar dependency relationships are utilized as they are syntactic rather than semantic, and thus, not restricted to any domain, and relationship between words is preserved as well. In addition, two new rules are introduced to preserve the inherent relation between the news image and text. A triple extractor is implemented using these rules. The resultant extracted triples are then used to construct a multimodal KG that represents the news image and its background knowledge, where the main entities and their relationships are clearly illustrated.

This paper is divided into several sections. Section 2 highlights issues related to multimodal KG and current linguistic-based relation extraction techniques. Section 3 discusses in depth the proposed framework to build the multimodal KG automatically from a corpus of news articles. The framework consists of three main phases: Phase 1: Pre-process datasets, Phase 2: Extract entities and relationships, and Phase 3: Build a multimodal KG. The triple extractor which produced the multimodal KG that describes the relationship between texts and images is evaluated based on precision and recall metrics and the findings are shown and discussed in Section 4 and finally, Section 5 concludes this paper.

II. Related Work

This section provides an overview of KG and multimodal KG, with a focus on how text are processed and used in building multimodal KG as well as relation extraction (RE) techniques specifically the rule-based approach and dependency relationships.

A. Multimodal Knowledge Graphs (KG)

A knowledge graph (KG) is a directed labelled graph consisting of nodes and edges. Nodes are real-world concepts and apart from text, images can also be used as nodes. Edge connects a pair of nodes and shows the relationship between nodes [19]. Nodes can also be known as entities and the edges that connect these two entities are known as relations or relationships. Two entities and the relation between them are referred to as a triple, which are the basic building block for KG. Many existing KGs have been built such as Google Knowledge Graph, DBPedia, Wordnet, ConceptNet. These existing graphs have been used in many real-world applications including computer vision tasks such as object detection, visual question answering tasks, image classification and more.

Wu et al. [20] have built a KG from text in news articles. Since this work focuses on summarize output, the resulting graph is a summary KG even though the original input is a long sentence. The essence of the sentence has been captured; however, it may not capture other information found in a long sentence. For example, given the following two sentences:

"Two more young black men join in the beating, which is caught on cameras."

"Two men who are young black and join fight."

The first sentence is the original long input sentence which is summarized to the second shorter input text. Triples from the summarized sentences are used to build the KG. Hence, in the example, the output does not capture the phrase "caught on cameras" which can also be the entity and relation for another triple. In another example, "Alice and Bob took the train to visit the zoo. They saw a baby giraffe, a lion, and a flock of colorful tropical birds." is summarized to "Alice and Bob visited the zoo and saw animals and birds". The words in bold in the original longer sentence are the text that are dropped and summed up as "animals and birds". These nouns can be important entities in computer vision tasks and the adjectives of the entities can be used to describe them. To obtain the main gist of a piece of information, the summarized sentences are sufficient, and the additional information may seem trivial. However, this trivial information can be considered particularly useful in object detection tasks even if it is only some descriptive text for certain entities.

Gong et al. [21] have produced multimodal learning approach for information extraction in which entities involve not only text but can include images or audio and relationships that connect entities either within or across modalities. The authors state that multimodal information such as text, pictures or audio are usually interrelated and complement to each other. Text and image modalities are focused due to the high availability of information.

Likewise, Gong and Wang [12] propose a multimodal learning algorithm to integrate textual information into visual knowledge extraction. While they have linked both the visual and textual parts in their proposed multimodal information extraction method, only nouns and noun phrases are used to tag image (or object in the image) with the "has-tag" relationship linking the image (or image object) to the image tags (i.e. the nouns or noun phrases), leaving out other rich information available from text, for example, in the sentence "A girl is playing with a sleeping dog in a room", the bold text "playing" and "sleeping" which give the actions for the nouns "girl" and "dog", respectively, are not captured.

Attribute is generally used to describe an object. According to [22], attributes allow to describe, compare and categorize objects easily. Researchers such as [23] and [24] have proven that with the addition of these attribute, there have been and improvement in the visual tasks. Hence, the present multimodal KGs can still be further improved by filling in more information that is available from text into multimodal KG.

B. Relation Extraction (RE)

As mentioned in Section 1, to build a KG from text, it is very important to understand the text before extracting the relationship. Thus, leading to the task of relation extraction (RE). RE is a major sub-task of information extraction [25] and is also utilized for the detection and classification of semantic relationships between entity pairs [26].

Among the techniques in RE, [14] is one of the works that used a rule-based approach. The authors used this approach for mapping a predicate of a triple to an identical predicate in a KG. However, the generated rules cannot cover all possible patterns in open domain because of the sparsity of unstructured text. Similarly, in [15] use a rule-based approach but with the addition of a similarity-based approach to achieve their objective. In which, the resulting rules are able to cover all possible patterns which result in more complete triples.

In [16] has conducted a study to build an open information extraction system for Indonesian language with rule-based approach. This author has proven that by only using rule-based still can formulate a generalized rule that can capture triples in a wide, open domain. Thus, this work is referenced in terms of the method used to extract relationships between entities. They use part of speech (POS) tagging such as noun, verb, etc. and dependency relationships to extract relationships. The authors concluded that the method used was to identify the relationship based on the VERB POS tag in the extraction of the single verbs. Moreover, the ADVMOD (adverbial modifier) dependency relationship was used alongside VERB POS to obtain a more complete relationship. Extracting entities for both subject and object produced a complete triple. However, the author does not consider the syntactic relationship that exists between the texts which describes a word i.e., an adjective to a noun. By taking in consideration this type of relation, most of the relationships that exist between texts will be captured.

Overall, the reviewed RE methods perform well for simple sentences but poorly for complex sentences. This study attempts to consider adverbial phrases with the extraction of verbs and prepositions in addition to the extraction of single verbs and utilizes the adjectival modifier (AMOD) dependency relation where this relationship describes the nature of an entity; therefore, leveraging on the available text resources.

C. Dependency Relationships

The dependency-based parser labels the relationship that is dependent on the key word in order to get a sense of the

predicate-argument relationship [27]. The task of the dependency parser is to take the input text and apply the proper set of dependency relationships to it [28]. A dependency parser helps to create a dependency tree, which is a tree model based on dependency relationships, by parsing words or sentences.

TABLE I. RELATIONS IN CLAUSE PREDICATE CATEGORY

Clause Predicate	*Description*
Nsubj	Nominal subject
Nsubjpass	Passive nominal
Csubj	Clausal subject
Csubjpass	Clausal passive subject
Dobj	Direct object
Iobj	Indirect object
Ccomp	Clausal complement
Xcomp	Open clausal complement

TABLE II. RELATIONS IN NOUN DEPENDENTS (MODIFIER) CATEGORY

Noun Dependents	*Description*
Amod	Adjectival modifier
Advmod	Adverbial modifier
Nmod	Nominal modifier
Nummod	Numeric modifier
Appos	Appositional modifier
Det	Determiner
Compound	Compound

Based on Universal Dependencies (UD) [29], there are 42 relationships that can be grouped into nine categories: (1) clausal predicates, (2) Non-core dependents of clausal predicates, (3) clausal dependents, (4) Noun dependents, (5) Coordination, (6) Compounding and unanalyzed, (7) Case-marking, prepositions, possessive, (8) Loose joining relations dan (9) others. According to [27], frequently used relationships focused on only two of the nine UD categories. The two categories are clausal predicates and noun dependents (modifiers). Table I and Table II are the examples of the list of universal dependency relationships that have existed. Some of these relationships, mainly subject (nsubj, nsubjpass, csubj, csubjpass), object (dobj and iobj), modifiers (amod, advmod) and compound relations, will be investigated in the analysis process of defining the rules.

III. PROPOSED METHOD

In this section, a detailed explanation of the framework for building a multimodal KG is given. The framework for this study has three main phases. Phase 1: pre-process dataset, Phase 2: extract entities and relationships and Phase 3: build a multimodal KG. Fig. 3 shows the flow of how a multimodal KG is built from a news article.

This study contributes to the extraction of entities and relationships from real-world sentences found in news articles that can be simple short sentences up to long and complex sentences. Simple sentence is the article headline, mainly a short sentence that only have one verb per sentence while complex sentence is the caption which is a long sentence that can consist of multiple verbs in a sentence. This study aims to extract the relationship between two entities or known as triple from the text. The dependency relationship technique is applied to maintain the relationship of each word. Linguistic analysis is performed on each sentence to obtain grammatical dependencies, where a set of dependency relationships will be identified. This is an initial step to detect consistent pattern to formulate relation extraction rules. Once the rules set have been formulated, this set will be used to extract triples of the text. Finally, the extracted triple is used to construct a multimodal KG.

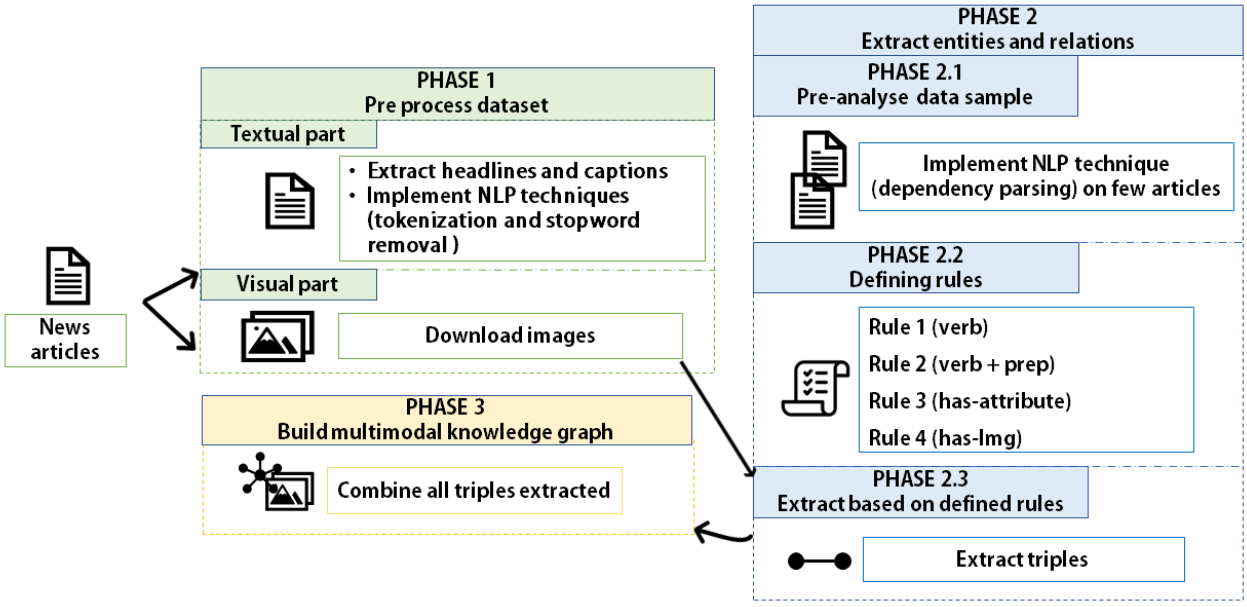


Fig. 3. Framework for Building a Multimodal KG.

A. Phase 1 (Pre-process Dataset)

Starting with the first phase, the Phase 1 input is a collection of online news articles. Each article consists of an image and a text article. Here, the focus is on the image and the descriptive text that accompanies the image. Therefore, this study only considers article titles, captions and images. The text source is derived from the title and caption of the image while the visual source is from the article image.

A web news article consisting of news text and images related to the news. News articles need to be cleaned and filtered because news articles have a lot of information gathered (headlines, captions, date published and more). However, in this study, only a small portion of this rich information will be used. This is because the title and caption are sufficient to describe the image. NLP techniques such as tokenization and stop word removal will be applied to titles and captions. Both techniques will be modified so that the output produced is consistent with the study.

To briefly described, tokens can be formed in individual words or phrases but usually, one word is detected as one token. Even so, in some cases, there are words that cannot be considered a token. Tokenization is customized in a way that can combine several words as a single token. This customization will collect all words that have a hyphen (-) and combine them into one token. The next pre-processing step is stopword removal. The original stopword list should not be used in this study. Words such as prepositions (at, in, of, etc.) can be used as examples of why the original list cannot be used. This is because words prepositional words will be used later to extract relationships. Thus, the stopword list will be self-determined to suit this study. Specifically, words like 'left, right, above, center, below'. These words are generally used in the caption to describe the position of a particular object in an image and are removed because it affects the performance of dependency-based parser which makes it less accurate.

For the visual part, each news article usually has an image to support the news. Each of these images will be downloaded via URL. It is then stored to be paired with the has-Image relationship which will be explained further in phase 2.2. This is to indicate that the caption is related to the image.

In conclusion, the output from this phase is the extracted text referring to the title and caption, and the image that has been downloaded. Finally, the dataset, D, represents the set of the entire news article.

$$D = \{T, V\} \quad (1)$$

where T and V are equivalent to the text and visual parts, respectively. For the set, T, consisting of h and c. h represents the title while c represents the caption.

$$\{h, c\} = T \quad (2)$$

For set V, which consists of only one element, img, which is the downloaded image.

$$\{img\} = V \quad (3)$$

B. Phase 2 (Dependency-based Entity Relation Extraction)

The next phase is an extension of the textual part which is extracting entities and relationships. It will be divided into three sub-phases namely Phase 2.1: pre-analyse data, Phase 2.2: defining rules and Phase 2.3: extract based on the defined rules. Briefly, in Phase 2.1, several articles consisting of simple and complex sentences were selected and parsed through a grammatical parser. In Phase 2.2, the output of the grammar parser which is the dependency tree of each article will be analysed. This is an initial step to detect consistent pattern to formulate relation extraction rules. Once the rules have been determined, the triples (entities and relationships) will be extracted in Phase 2.3. Also, the downloaded images earlier on will be used as one of the entities for the triple set. Typically, a sentence with only a single verb consists of one triple. However, in some cases, there can be one sentence consisting of several triples. For instance, long sentences that have multiple verbs will produce more than one set of triples.

1) Phase 2.1 (pre-analyse data: dependency parsing analysis): In this sub-phase, the pre-analysis is carried out to identify rules for entity extraction. 10 sentences (headlines and captions) are randomly selected from several news articles together with the accompanying images and are analyzed manually which information supposed to be extracted. Then, it is parsed through a parser to show its grammatical structure (POS tag, dependency relations) to detect a pattern. The 10 sentences consist of five simple sentences and five complex sentences. Five more sentences are analyzed to ensure that no new patterns emerge. Hence, these 10 sentences are sufficient for the pre-analysis because of the nature of English sentences to have a similar grammar pattern, thus the result will be much alike.

Firstly, the set of sentences are manually examined where words that describe an image are identified and marked in yellow as shown in the Fig. 4 and 5. These marked words are considered as information that should be extracted. The sentences are put through a grammar dependency parser to obtain the parse tree structure, list of dependency relations and POS tag. Each word that has been marked is viewed in its grammatical structure that is the dependency relationship and POS tag as shown in Fig. 6. Based on the grammatical outputs produced, there are several dependency relationships that are often present on the marked words. The Visitation: Glasgow City Council pay family over Nazi-looted artwork. The crew of an Emirates Airline Boeing 777 prepares for passengers ahead of a demonstration flight in Dubai in 2007. Emirates announced plans Thursday to begin flying a Dubai to Panama City route on a 777, which will be the longest in the world.

Table III shows the frequently found dependency relations together with their explanations by [30,31].

Fig. 4. Example of Simple Sentence.

Fig. 5. Example of Complex Sentence.

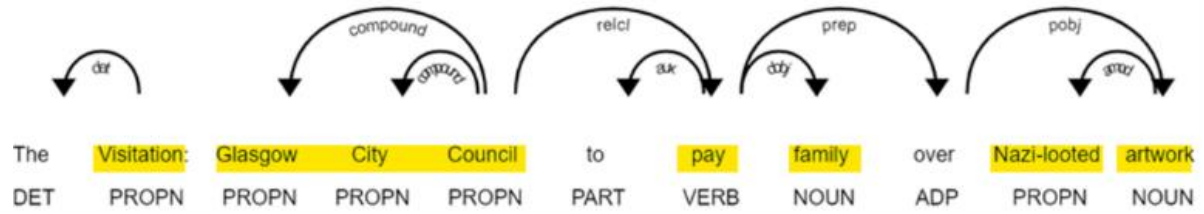


Fig. 6. Example of Dependency Tree.

TABLE III. TYPES OF DEPENDENCY RELATIONSHIPS

Dependency Relationships	Description
ROOT	the main topic (verb) for the sentence
Compound	nouns that modify the head of a noun phrase
Nsubj	nominal subject (noun phrase)
Dobj	direct object
Pobj	prepositional object
amod	adjective phrases that change the meaning of a noun phrase
Prep	prepositional phrases that modify the main meaning

According to [27], frequently used relationships are focused on two categories of relationships as mentioned in the previous section which is the clause predicate (Table I) and Noun Dependent (Table II). In Table III, these are the list of consistent relationships generated by the parser which fall into both of those relationship categories. The relationships are Compound, Nsubj, Dobj and amod. The prep relationship needs to be considered in order to capture the verb as a whole. Pobj in turn connects object entities to prep relationships.

According to [31], ROOT is a special label in the dependency tree that is usually on the main verb of a sentence. For some cases, if a phrase is processed or in other words is not a full sentence, ROOT is assigned to the noun of the head of the phrase. It can be observed that ROOT can be either a verb or a noun, in some sentences; and when the ROOT relation and POS noun are found in a sentence, the noun entity is highly related to the image, or to a particular entity in the image. Hence, this is used to specify a new relation between the text (noun entity) and image (i.e., a text-visual relationship).

Apart from the dependency relationship is the POS tagging. This is the process of labelling punctuation in a language based on class classification. In other words, POS tagging indicate parts of speech to each word such as nouns, verbs, adjectives and more. Table IV shows the most common labels found on words that have been identified by the parser.

Having observed and analyzed the output of dependency relationships and POS tagging for these data samples, Tables III and IV are consistent patterns generated by the parser of the marked text. All these dependency relations and POS tag are used as the basis for the entity and relations extraction rules.

TABLE IV. POS TAG LABEL

POS tag	Description
ADP	Preposition
ADJ	Adjective
DET	Determiner
NOUN	Noun
PROPN	Proper nouns
VERB	Verb

Based on the output in Fig. 6, basically for a relationship between texts, the text to be captured as a relationship has a POS tag labelled VERB and the entity has a POS tag labelled NOUN or PROPN. But the proposed method will also use the dependency relationship that has been generated by the

dependency parser to extract the entities more accurately. The pattern for entity extraction based on dependency relationships is that text labeled Nsubj, Dobj or Pobj will be captured as an entity. The dependency relationship derived from the dependency tree also denotes a word that is either a subject or an object.

To extract the overall relation of the verb, the text having the 'prep' dependency relationship was also combined with the main verb. With this, making the extracted relationship was more ideal. Furthermore, texts describing an entity are extracted based on 'amod' dependency relationships making full use of all naturally occurring relationships between texts. In contrast to the text-image relationship, this relationship is pre-defined and uses the ROOT label to extract the appropriate entity.

Consistent pattern of dependency relationships and POS tag are used as the basis for rules for extracting entities and relationships. Also, relationships will be categorized into two namely 1) text-text and 2) text-visual. Category 1 is an extracted relationship that exists naturally between texts, for example, verb-based relationships, verb-based relationships and prepositions and relationships based on 'amod' dependency relationships but are named as has-Attribute while Category 2 is a relationship set to link text with images.

Once the entity and relation are successfully extracted completely, then the triple, R that is completely extracted will be produced and the relation (h), subject (s), and object (o) are arranged in the following order.

$$R = \{h, s, o\} \quad (4)$$

2) Phase 2.2 (Defining rules): As noted earlier, this study focuses on rule-based relation extraction and covers not only text-text relationships but also text-visual relationships. In this subphase, four types of rules are set for extracting relationships as shown in Fig. 7. Therefore, the development of the rules is explained in more detail. After the analysis in Phase 2.1 is performed, the following rules are defined for extracting the relationship:

- Rule 1 (based on verb relationships).
- Rule 2 (based on verb + preposition relationship).
- Rule 3 (based on has-Attribute relationship).
- Rule 4 (based on has-Image relationship).

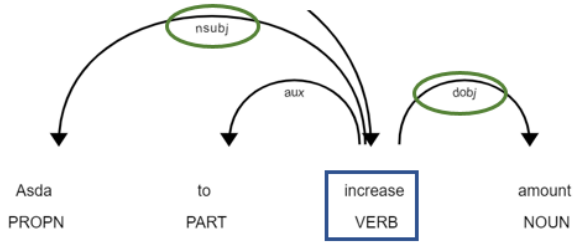
For Rule 1, the relation is captured first before the entity, so the first step is to identify the verb. The token having the ROOT dependency label and verb POS tag (VERB) will be captured as its relation. Next is to find subjects and objects as entities.

As for Rule 2, in addition to the verb itself, the relationship of verb + preposition is also considered in this study; similar to Rule 1, but with the addition of a prep dependency relationship as in Fig. 7.

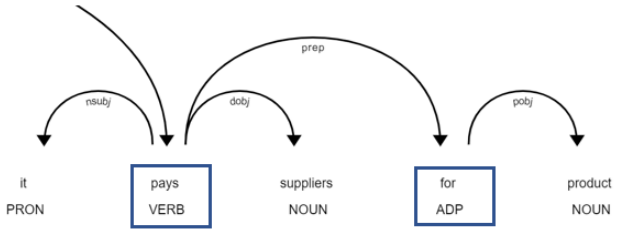
Next, this Rule 3 is based on an amod-dependency relationship renamed as 'has-Attribute' relationship to indicate the characteristics of a particular word, which is mostly taken from an adjective word. Other than the previous rule, this Rule 3 will identify the object first by identifying the word with the amod dependence.

Since the has-Image relationship has been predefined so Rule 4 is defined to extract the entities for the relationship that link the image and text. Thus, it represents a text-visual relationship.

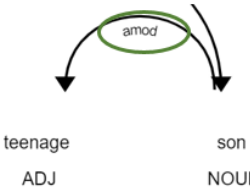
3) Phase 2.3 (Extract based on rules): Once the rules are determined, then the relations and entities will be extracted according to the rules as described in Phase 2.2. The processed data will go through an algorithm and a triple list will be generated. It will be arranged according to the respective articles. In this way it is clearer that there are some sentences that have more than one triple especially complex sentence.



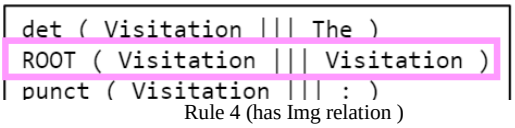
Rule 1 (Verb only)



Rule 2 (Verb + preposition)



Rule 3 (has Attribute relation)



Rule 4 (has Img relation)

Fig. 7. List of Rules.

C. Phase 3 (Build a Multimodal KG)

In this phase, the relationship of texts and text-visual will be described. The extracted triples will be combined and produce a multimodal knowledge graph which can be formulated as follows. G represents a multimodal knowledge graph for an article that contains a combination of several subgraphs or triples, R_n where n is the total number of triples extracted.

$$G = \{R_1, R_2, \dots R_n\} \quad (5)$$

The multimodal KG can be visualized as in Fig. 8 where the graph has two different types of modalities namely text and image. From the graph, there are other subgraphs. A subgraph produced represents a triple.

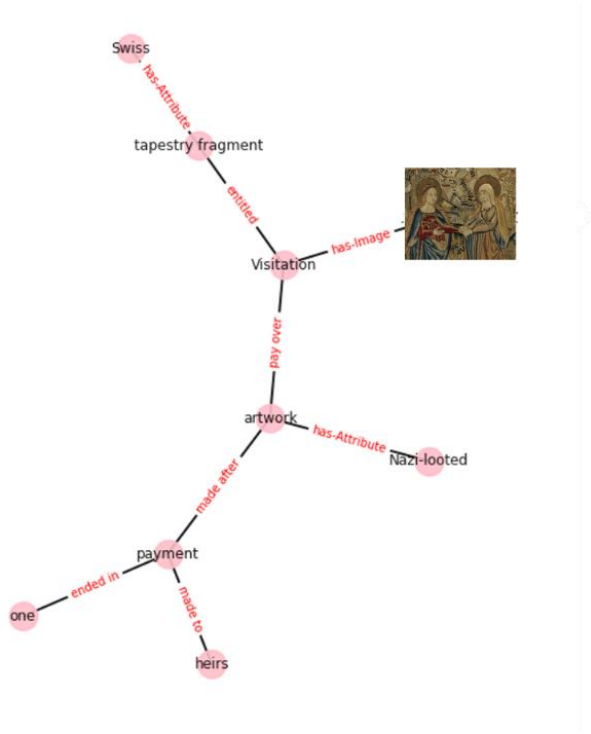


Fig. 8. Multimodal Knowledge Graph.

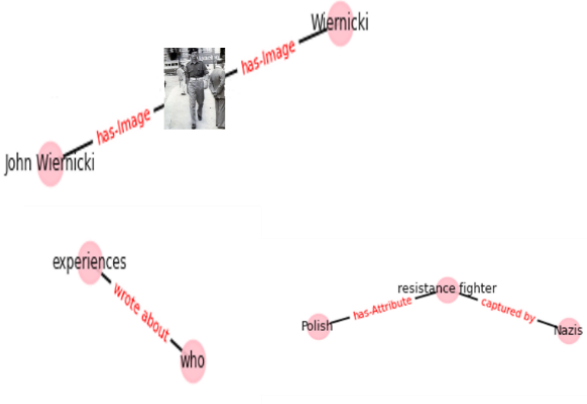


Fig. 9. Hanging Sub-graph.

However, after the experiment was done it was found that there was a subgraph that hung alone as in the following example of Fig. 9. For human-like mindset, it can be concluded that the hanging subgraph still has some relation with the image. This relationship can be labelled as an indirect relationship. Thus, the addition of the rule (Rule 5) for the indirect relations will be applied to the subgraph. This relationship will be called has-Bg-Kg where Bg represents Background while Kg represents Knowledge. This relationship indicates that the subgraph describes the background knowledge of an image in addition to creating a stronger relationship between other subgraphs and also the image-text relationship.

IV. Experiment and Discussion

A. Experimental Setup

In this study, the dataset used in this experiment is a sub-dataset from ION Corpus (Hollink et al., 2016) consisting of 60 news articles published on five newspaper websites, collected from August 2014 and August 2015. This dataset contains a collection of sentences with different difficulty level such as simple sentence which mainly the headline of a news article and the complex/long sentence which are the captions from the images.

The original news articles had to be cleaned for having a lot of information collected (title, caption, date of publication and more). The data that has been extracted from news articles primarily are headlines and captions. These text needs to go through a process of tokenization and stopword removal. The library that will be used for the NLP techniques implemented in this experiment is the Spacy Library. Both processes need to be modified as needed to produce a suitable output.

With the result from the pre-processing phase of the dataset, the texts will go through the dependency parser and a dependency tree will be produced as. Several sentences will be expressed as a dependency tree as an initial step in making rules for the rule-based relation extraction. A total of 60 articles were used to extract relationships and entities to produce triples.

To evaluate the quality and quantity of the generated triples for the multimodal KG, the evaluation is performed by manually extracting triples from the 60 articles. These sentences from the articles are examined for relevant triples and then compared to the triples extracted by the algorithm.

B. Comparison

This section will discuss the comparisons between the two methods from other works such as Gong & Wang [12] and Romadhony et al. [16]. Table V shows the summary of both the comparisons. Gong & Wang's method [12] will be compared to the way they extract only nouns to link between text as well as visuals. Romadhony et al.'s method [16] extracts the relationships that can be found from the text itself. The rules they set are from the verb and also the ADVMOD dependency relationship where it changes the predicate or verb.

TABLE V. SUMMARY OF COMPARISON

Proposed method	Gong & Wang [12]	Romadhony et al. [16]
Extract from verb and also consider preposition	only extract nouns	Extract from the verb and the ADVMOD dependency relationship
Additional attribute relationship	-	-
Has relation between image and text but also extract mostly information available	Has link between text as well as visuals but only simple text	Only extract the relationships that can be found from the text

The proposed method is an improvement of the following two works. This is because, for the method of Romadhony et al. [16], the authors focus to texts only. For [12], the method used only focuses on the image-text relationship which indirectly ignores other information that can be obtained from the text. This proposed method not only succeeds in extracting the relationship of texts, but also the relationship of image-texts also makes it very full of information to be filled in the multimodal KG to be built.

In addition, the proposed method can also produce triples of a long and complex sentence. The has-Attribute relationship makes the extracted relationship more adequate and less neglect of information from the text. This is because all the information that can be extracted from the text can be used to the fullest.

$$P = \frac{X\ correctly\ extracted\ triples\ by\ algorithm}{Total\ triples\ extracted\ by\ algotithm} \quad (5)$$

$$R = \frac{X\ correctly\ extracted\ by\ algorithm}{Total\ relevant\ triples\ in\ the\ corpus} \quad (6)$$

This proposed method is evaluated with the formula precision (P) and recall (R). Based on Table VI, here we can see that P = 0.90 and R = 0.60 for this study is the highest when compared to the other two methods. Thus, it can be concluded that this proposed method succeeds in extracting more triples and importantly accurate triples. The value of accuracy (P) here is high because among the 209 triples produced, there are 190 triples produced correctly. However, the recovery value (R) has only a value of 0.6 where only 190 triples are produced out of 319 triples that are theoretically capable of being produced.

TABLE VI. PRECISION AND RECALL SCORE

Total/ Method		***Proposed method***	***(Gong & Wang, 2017)***	***(Romadhony et al., 2018)***
Triples should be extracted (319)	The triple is extracted by the algorithm	209	314	71
	Triple the extracted accurately	190	147	60
	Precision score (P)	0.90	0.47	0.85
	Recall score (R)	0.60	0.46	0.20

This proposed method is highly dependent on parser dependency performance making it a challenge because when the performance of this parser is low as it affects the results. Generally, parser performance deteriorates with complex sentences. Another issue faced with complex sentences is the longer the sentence, the more clauses it contains, making it harder to trace back to the subject (entity) in the main clause, which resulted in the extraction of some incomplete triplets with insufficient entities, as described in the previous paragraph. Another reason for the incomplete triplets is because of co-referencing (pronouns) problem. If co-referencing resolution is performed and parser performance for complex sentences can be improved, then the value of R can also be increased.

V. CONCLUSION

The proposed multimodal KG has successfully extracted the Web image background knowledge from unstructured texts and organized in a structured graph while still maintaining the image-text relationship. A set of rules based on repeated patterns of dependency relations and POS information, can correctly extract from simple to complex sentences regardless of the domain (sport, education, world, etc.). Two additional rules are included to take into consideration the inherent correlation between the Web image with the news headline and image caption. Hence, capturing the background knowledge for Web images that are much needed by researchers in the computer vision field. This multimodal KG can be used as training data for machine learning approaches such as graph embeddings.

REFERENCES

[1] Chan, C. S., Johar, A., & Hong, J. L. “Contextual information for image retrieval systems.” Proceedings - 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2013, 2013, pp. 863–867.

[2] Tiwari, P. “A Survey on Image Context Extraction Method.” International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(11), 2014, pp. 85–93.

[3] Wang, B., Lin, D., Xiong, H., & Zheng, Y. F. “Joint Inference of Objects and Scenes with Efficient Learning of Text-Object-Scene Relations.” IEEE Transactions on Multimedia, 18(3), 2016, pp. 507–520.

[4] Fang, Y., Kuan, K., Lin, J., Tan, C., & Chandrasekhar, V. “Object Detection Meets Knowledge Graphs.” 2017, pp. 1661–1667.

[5] Chu, T. H., Huang, H. H., & Chen, H. H. “Image recall on image-text intertwined lifelogs.” Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019, 2019, pp. 398–402.

[6] Vondrick, C., Oktay, D., Pirsiavash, H., & Torralba, A. “Predicting Motivations of Actions by Leveraging Text.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2997–3005. https://doi.org/10.1109/CVPR.2016.327.

[7] Hollink, L., Bedjeti, A., van Harmelen, M., & Elliott, D. “A Corpus of Images and Text in Online News.” Lrec, 2016, pp. 1377–1382.

[8] Ramisa, A., Yan, F., Moreno-noguer, F., & Mikolajczyk, K. “BreakingNews: Article Annotation by Image and Text Processing.” IEEE Trans Pattern Anal Mach Intell. 40(5), 2018, pp. 1–21.

[9] Pan, X., Sun, K., Yu, D., Chen, J., Ji, H., Cardie, C., & Yu, D. “Improving Question Answering with External Knowledge.” 2019, pp. 27–37.

[10] Wu, Q., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. “Ask me anything: Free-form visual question answering based on knowledge from external sources.” Proceedings of the IEEE Computer Society

Conference on Computer Vision and Pattern Recognition, 2016-Decem, 2016, pp. 4622–4630.

[11] Li, Z., Tang, J., & Mei, T. "Deep Collaborative Embedding for Social Image Understanding." IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(X), 2018. pp. 1. exist in the real-world is by building a knowledge graph.

[12] Gong, D., & Wang, D. Z. "Extracting visual knowledge from the web with multimodal learning." IJCAI International Joint Conference on Artificial Intelligence, 0, 2017, pp. 1718–1724.

[13] Li, K., Zhang, J., Yao, C., & Shi, C. "Automatic relation extraction from text: A survey." Proceedings - 2016 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI 2016, 2018-Janua, 2016, pp. 83–86.

[14] Exner, P., & Nugues, P. "Entity extraction: From unstructured text to dbpedia rdf triples." CEUR Workshop Proceedings, 906, 2012, pp.58–69.

[15] Kertkeidkachorn, N., & Ichise, R. "T2KG: An end-to-end system for creating knowledge graph from unstructured text." AAAI Workshop - Technical Report, WS-17-01-, 2017. pp. 743–749.

[16] Romadhony, A., Purwarianti, A., & Widyantoro, D. H. "Rule-based Indonesian Open Information Extraction." ICAICTA 2018 - 5th International Conference on Advanced Informatics: Concepts Theory and Applications, 2018, pp. 107–112.

[17] C Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M." Toward an architecture for never-ending language learning." Proceedings of the National Conference on Artificial Intelligence, 3, 2010, pp.1306–1313.

[18] Mausam, Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. "Open language learning for information extraction." EMNLP-CoNLL 2012 - 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference, July, 2012, pp. 523–534.

[19] Chaudhri, V. K. "Knowledge Graphs: What is a Knowledge Graph?." 2021, https://web.stanford.edu/class/cs520/2020/notes/What_is_a_Knowledge _Graph.html [accessed on 10 July 2021].

[20] Wu, P., Zhou, Q., Lei, Z., Qiu, W., & Li, X. "Template Oriented Text Summarization via Knowledge Graph." ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing, 2018, pp. 79–83.

[21] Gong, D., Wang, D. Z., & Peng, Y. (2017). Multimodal learning for web information extraction. MM 2017 - Proceedings of the 2017 ACM Multimedia Conference, 288–296. https://doi.org/10.1145/3123266.3123296.

[22] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., & Fei-Fei, L. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." International Journal of Computer Vision, 2016, pp. 123, 32-73.

[23] Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. "Describing objects by their attributes." In IEEE conference on computervision and pattern recognition, CVPR 2009, 2009, pp. 1778–1785.

[24] Goering, C., Rodner, E., Freytag, A., & Denzler, J. "Nonparametric part transfer for fine-grained recognition." In 2014 IEEE conference on computer vision and pattern recognition (CVPR), 2014, pp. 2489–2496.

[25] Miao, F., Liu, H., Miao, B., & Liu, C. "Open domain news text relationship extraction based on dependency syntax." Proceedings of 2018 IEEE International Conference of Safety Produce Informatization, IICSPI 2018, 2019, pp. 310–314.

[26] She, H., Wu, B., Wang, B., & Chi, R. "Distant Supervision for Relation Extraction with Hierarchical Attention and Entity Descriptions. " Proceedings of the International Joint Conference on Neural Networks, 2018-July, 2018, pp. 1–8.

[27] Cao, Q., Liang, X., Li, B., Li, G., & Lin, L. "Visual Question Reasoning on General Dependency Tree." Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 7249–7257.

[28] Yusuf, A. A., Nwojo, N. A., & Boukar, M. M. "Basic dependency parsing in natural language inference." 2017 13th International Conference on Electronics, Computer and Computation, ICECCO 2017, 2018-Janua, pp. 1–4.

[29] De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. "Universal stanford dependencies: A cross-linguistic typology." Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014, 2014, pp. 4585–4592.

[30] Choi, J. D. "ClearNLP Dependency Labels." https://github.com/clir/clearnlp-guidelines/blob/master/md/specifications/dependency_labels.md, 2015, [accessed on 20 October 2021].

[31] A Altinok, D. "Mastering spaCy (1st ed.)." Packt Publishing. https://www.perlego.com/book/2742862/mastering-spacy-pdf, 2021, [accessed on 20 October 2021].