ISSN 2007-9737

Multimodal Learning Based Spatial Relation Identification

Sandeep Kumar Dash[1], Y. V. Sureshchandra[1],Yatharth Mishra[1],
Partha Pakray[2], Ranjita Das[1], Alexander Gelbukh[3]

[1] National Insitute of Technology, Mizoram,
India

[2] National Insitute of Technology, Silchar,
India

[3] Instituto Politécnico Nacional, CIC,
Mexico

sandeep.cse@nitmz.ac.in, vipulyadav150@gmail.com, yatharthmishra01@gmail.com,
partha@cse.nits.ac.in, rdas@nitmz.ac.in, www.gelbukh.com

Abstract. Spatial Relation identification is one of the integral parts of Spatial Information Retrieval. It deals with identifying the spatially related objects in view of their physical orientation or placement with respect to each other. The concept is widely used in many fields such as Robotics, Image Caption Generation and many more such areas. In this work the focus is to gather information from multiple modalities such as Image and its corresponding Text so as to strengthen the learning process for the identification of Spatial Relation pairs from a given text. Two different multimodal approaches are proposed in this work. In the first approach, information is explored as a sequential learning process where the individual Spatial Roles are identified as connected entities, which makes the Spatial Relation retrieval easy and efficient enough. To counter the small size of the dataset along with necessity to avoid overfitting, an efficient backward propagation based Neural Network was used to classify the candidate roles and the relations. The feature selection was different for all the classification tasks. Building on the selected feature from the first approach, the second approach uses a transfer learning method that utilizes an existing image caption generation model to retrieve the vital topic based information from image which is then used for the task. Thereby both approaches used information from two modalities which are further used to train the system in the respective approach. The model achieves state-of-the-art performance in terms of Precision for two of the Spatial Roles identification. This validates the advantage of using multimodal learning when compared with other partial-multimodal processes.

Keywords. Spatial role labeling, spatial relation identification, multimodal learning, multi layer perceptron.

1 Introduction

Multimodal learning involves learning from two or more source of information so as to utilize both implicit and explicit information available in them for a given problem. The different modalities are text,image,voice,video etc. The advantage of utilizing multiple modalities for a given problem, if it can be applied, is relating the implicit information with the solution space. This not only bolsters the efficiency of the method but also provides a subtle solution. Since Spatial information retrieval is one of the key aspects in utilizing multimodal learning for identifying the relational information about real world objects, this particular work involves image and its corresponding description in text as the two modalities under consideration. This helps in increasing the efficiency of model which needs object based relational information such as path navigation models in robotics, caption generation models for describing the objects which should be in focus in an image and many more.

In this prospective, the task of Multimodal Spatial Role Labeling (mSpRL) [8] proposed as part of CLEF 2017 [4] holds much importance as the sub

ISSN 2007-9737

tasks in this clearly builds upon on the spatial information present in the text.

1.1 Multimodal Spatial Role Labeling

Multimodal Spatial Role Labeling has been built upon the SemEval 2012 task of Spatial Role Labeling (SpRL) [5] and follows the same nomenclature.

The task deals with extracting Spatial Information about objects referenced in a text with supporting information identified from its corresponding image. The proposed research touches upon the first and second major sub-parts of the task. The following sub-tasks are part of them.

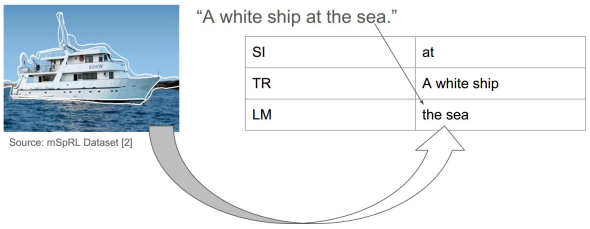

Fig. 1. Identification of Spatial Roles

1.1.1 Identifying Spatial Roles

This deals with retrieving the following three candidate roles, as phrases, from the sentence:

- **Spatial Indicator(SI)**: This is the candidate role which ensures presence of spatial information in the text.
- **Trajector(TR)**: These roles depict the entities about which the spatial information is being retrieved.
- **Landmark(LM)**: These are the reference entities which describes the location of Trajector.

Fig. 1 explains utilization of Multimodal information for identifying the candidate roles. The feature extracted from image bolsters the presence of sea as a 'Landmark' entity.

1.1.2 Identifying Spatial Relation

The sub-task deals with relating the identified spatial roles as a triplet, which retrieves the relative spatial information about the real world entities. It is represented as $\langle SI, TR, LM \rangle$.

For the above example $\langle at, awhiteship, thesea \rangle$ is the Spatial Relation triplet. The Spatial Relation triplet may also be ambiguous as the participating valid roles may not be linked to each other. As in the following example, multiple triplets are possible for a given sentence, but not necessarily all are correct:

S1: *there are posters and maps on the wall in the background.*
1st triplet : $\langle postersandmaps, on, thewall \rangle$
2nd triplet : $\langle postersandmaps, in, thebackground \rangle$
3rd triplet : $\langle thewall, in, thebackground \rangle$

The first and third triplets out of the above are only correct. This is an important aspect of the Spatial Information as it validates both the Spatial Roles present in it and the relation among them.

The ambiguity behind this is stemmed with in the Natural Language text, where the different candidate phrases are often not attached to their spatially related entities which makes it difficult to classify the individual roles in a spatial relation.

This is where the image information can help. The feature extracted from images can explicitly deduce the position of the objects mentioned in the text from where the model can take the necessary information. For example, consider the following sentence:

S2: *picture and two lamps on the wall above the bed*

There are four objects present in the sentence: picture, lamps, wall, bed. Here, the dependency analysis will attach the object 'wall' with 'the bed' however the correct association is of 'picture and lamps' with both 'wall' and 'bed' respectively.

The research direction is limited by the non-availability of dataset sufficient enough for building a proper model in terms of multimodal learning. Traditional classification approaches have been overcoming this, with transfer learning

based training of their models, so as to either train binary classifiers to identify individual roles or multiclass classifiers to identify all the roles through it. The proposed model is built upon two different approaches namely an unique progressive classification approach of the individual Spatial Roles and a transfer learning approach.

In the first approach it initially trains a backward propagation based neural network to classify 'Landmark' along with the preposition it is dependent upon as one of the main features. This establishes efficient association of the landmark noun phrase with the preposition as such noun phrases are associated with only one preposition.

The second approach utilizes the information contained in the Topic Vectors for an image for classification of Spatial relations along with the other feature extracted in the first method. The topic vectors are extracted from the images by mapping the images on to a caption generation model which is trained on the images of MSCOCO 2014 dataset. For the first approach, consider the following example sentence:

S3: *a man is sitting at a table on the balcony*

There will be two candidate landmarks, 'a table' and 'the balcony'. These are trained along with the associated preposition 'at' and 'on', respectively. Further the Spatial Indicator, which is a preposition is also subjected to the same neural network classifier with its associated noun phrase. This stems from the fact that the Spatial Indicator and Landmark are syntactically and semantically dependent upon each other. Finally the Trajector, which is also a noun phrase is subjected to the same neural network classifier along with the pre-obtained SI and LM pairs.

However syntactically not all such noun phrases are linked with the SI and LM pairs. Therefore to obtain a system with better precision rather than better recall such pairs where the Noun Phrase is not associated with any of the Prepositional phrase in the tagged output is not considered for the classification process. Finally the same approach is followed for binary classification of the candidate triplets into Spatial Relation, based on the identified roles.

The main contribution of this work is two fold. The first one is implicating that the Spatial Information identification is a sequential process. That is finding the candidate roles are efficiently done in relation to another. It highlights related classification is more efficient than joint classification for extracting Spatial Roles. This is reflected in the state-of-the-art precise result produced for Multimodal Spatial Role identification process. Secondly the proposed model builds up on the shortcoming of low size of the dataset without following any transfer learning approach in one of the models and in another model it utilizes the transfer learning approach to produce fairly high precise Spatial Relation pairs.

The rest of the paper is organized as follows. Section 2 highlights related works in the direction. Section 3 outlines the proposed model detailing the methodology. Section 4 describes the experimental implementation of the models. Results Analysis is done in Section 5. Section 6 outlines the conclusion and future direction.

2 Related Work

Spatial Role Labeling [8] is an important aspect of Natural Language Processing. The Spatial information emerging out of it has variety of applications in real world scenarios. Due to lack of ample amount of annotated data, it becomes difficult to build a high accuracy model. Moreover use of a single modality such as text limits the extraction of all possible spatial information out of it. The CLEF 2017 workshop had thus coined the idea of Multimodal Spatial Role Labeling to utilize the missing link of information from the given text accompanied by Image.

The very first Multimodal Spatial Role labeling model was proposed in [8]. Their model shows inclusion of image feature vector with the word vector which improves the efficiency of the Spatial Role Labeling model. They have used an annotation scheme on MSCOCO's [7] images, in terms of the relative positions of the center point of the bounding boxes containing visual objects. The annotation style was based mainly on three predicates, alone(v1), below(v1; v2) and beside(v1; v2), where v1 and v2 were

visual objects. The image information was thus encoded in terms of these predicates which was combined with the text vector to be fed into a Multi Layer Perceptron for multiclass classification of the Spatial Roles. Further the Spatial Relation was also classified by combining the image information with the identified spatial roles as triplet. However in our proposed approach the image information is the feature vector of the entire image which lets the model learn the relative object position automatically without the need of a separate annotation model. Moreover the classification approach for Spatial Roles is not individually designed. It is rather a progressive approach as the initially identified Spatial Role is used further to identify other Spatial Roles.

In another approach followed in [11] the authors have used transfer learning techniques to train inter-modality alignment model as well as visual classifiers to build a global inference model that incorporates these components to a structured output prediction model for spatial role and relation extraction. One of their visual models is trained on ImageCLEF [15] for word-segment alignment and another one trained on Visual Genome dataset [6] for link disambiguation. Unlike their approach the proposed model neither use any externally trained system component nor any other dataset for classification of the Spatial Role and Relation, and achieves better result.

The concept of predicate and its role identification as in Semantic Role Labeling is followed in [9]. The authors have adapted a pre-existing system based on Convolutional Neural Network that was used for the task of Semantic Role Labeling. In the classification model therein the word vectors are combined with the lexical position based distance of the candidate role from the predicate. Their results have been reported on the Spatial Relation identification task only, with the corpus provided in SemEval 2015 [10]. However it was not a multimodal model.

A high-recall based model from [12] follows a heuristics of identifying objects capable of being a part of Spatial Relation. However, it utilizes a predefined set of prepositions. All possible combinations of the triplets are subjected to binary Support Vector Machine (SVM) based classification. The model is based on around 100 lexical features out of which some were removed using greedy method. Following their method the proposed model also selects nouns which are attached to a Prepositional Phrase which increases the model accuracy. However it does not form the Spatial Relation triplets based on fixed lexicons of prepositions rather follows classification approach to predict them leading to a more generic model.

3 Proposed Model

3.1 Model Description

3.1.1 Progressive Classification Approach

Our approach casts Spatial Information retrieval as a progressive classification approach. The Spatial Roles are identified initially with the heuristic that each of the roles are represented with relation to another. The model assumes that there is a Landmark for every Trajector and the Landmark is suitably described through a Spatial Indicator.

Trajectors and Landmarks are mostly noun phrases whereas Spatial Indicators are preposition phrases or prepositions. The model follows an individual and stepwise approach for classifying the spatial roles. Initially the classification of Landmarks is done with the feature vector formed with combination of (Iv, Nv, Dv, Sv), where:

- Iv: Image embedding,
- Nv: Embedding of head word of the candidate noun which has a preposition as its dependency head in the dependency tree of the sentence Vector,
- Dv: Embedding of the preposition on which the candidate noun is dependent,
- Sv: Embedding of the sentence.

For the sentence ‘He is standing behind the table.’, only ‘table’ qualifies for the candidate Landmark as can be found from the dependency analysis of the parse tree of the sentence. The prepositions similarly are paired with the nouns which are the identified landmarks in the previous stage. Thus the feature vector for classification of Spatial Indicators is represented as combination of (Iv, Pv, Nv, LI, Sv), where:

Iv: Image embedding,

Pv: Embedding of the candidate preposition,

Nv: Embedding of the noun,

LI: Indicator showing the noun is Landmark or not (value 1 or 0),

Sv: Embedding of the sentence.

Although both Trajectors and Landmarks are nouns they are syntactically different. The identification of candidate Trajectors is projected as a NP-PP attachment problem. This implies that the noun phrase attached prior to the preposition phrase (PP) often is the lexical level from where Trajectors can be deduced.

Thereby the model uses a chunking rule to identify those nouns which are head tokens of such noun phrases which is followed by a PP in the parsed tree representation of the sentence. Consider the following sentence:

S4: *trees in the background*

In this sentence, the chunking rule identifies the following chunks:

(NPPP (NP trees/NNS) (PP in/IN (NP the/DT background/NN)))

Here, NPPP represents NP followed by PP. In this out of the two identified NPs, the first NP is selected as a candidate for the Trajector role. The feature vector combines the candidate role along with the preposition and noun found in the attached PP. It also associates the information about whether the preposition and noun found in the PP are identified earlier as Spatial Indicator and Landmark or not.

Thereby it ensures that the spatial role is retrieved in coordination with other spatial roles. Thus the feature vector for classification of Trajector is represented as the combination of (Iv, Nv, Pv, Lnv, Sv), where:

Iv: Image embedding,

Nv: Embedding of the candidate noun,

Pv: Embedding of the preposition,

Lnv: Embedding of the noun in preposition phrase,

LI: Indicator showing the noun is Landmark or not (value 1 or 0),

SI: Indicator showing the preposition is Spatial Indicator or not (value 1 or 0),

Sv: Embedding of the sentence.

As the candidate roles are identified it becomes easier to form the triplets of Spatial Relation candidates. The candidates in this case are taken using the same NPPP chunk that were generated during Trajector role identification. The feature vector for classification of Spatial Relation is represented as the combination of (Iv, SIv, TRv, LMv, SI, TRI, LMI, Sv) where:

Iv: Image embedding,

SIv: Embedding of the preposition,

TRv: Embedding of noun before Preposition phrase,

LMv: Embedding of noun in preposition phrase,

SI: Indicator showing the preposition is Spatial Indicator or not (value 1 or 0),

TRI: Indicator showing the noun before preposition phrase is Trajector or not (value 1 or 0),

LMI: Indicator showing the noun in preposition phrase is Landmark or not (value 1 or 0),

Sv: Embedding of the sentence.

3.1.2 Transfer Learning Approach

As part of the Transfer Learning approach, an image caption generation model proposed in [2] was used. The approach followed in the model was to extract topic vector from the images and used it as one of the input feature to the caption generation process. The topic vector was extracted using the LDA model and was comprised of relevant concepts or semantics of the image.

The said model was trained on MSCOCO 2017 dataset. In this work, MSCOCO 2014 dataset was used to build the topic vector retrieval model. Initially following their approach LDA model [1] was used to extract topics from the description present in the CLEF mSpRL dataset. The topic model as proposed in their work was fed with the image feature extracted using a CNN architecture, which predicts the topics for the input image.

The topic words are considered to be the object categories present in MSCOCO 2014 dataset. These corresponded to the probability of occurrences of different topics related to the image. The extracted topic vector for each image is then used along with the vector composed of feature components as mentioned in the progressive classification approach for spatial relation identification.

3.2 Classification Framework

The model uses the back-propagation algorithm of a classical feed forward neural network. The main aim of selecting the network is its self learning ability. The network weights are updated by means of the back-propagating gradient vector where each element is defined as the derivative of an error measure with respect to a parameter. The error signal at the output of the neuron j for n-th iteration is defined as follows:

$$e_j(n) = d_j(n) - y_j(n), \tag{1}$$

where d_j is the desired output for neuron j and y_j is the actual output for neuron j calculated by using the current weights of the network at iteration n.

The Spatial Role labels and Spatial Relations follow the binary classification approach for identification of the Spatial information.

3.3 Dataset Description

The dataset is from the Multimodal Spatial Role Labeling track of CLEF 2017, which has two components: image and text. The image portion contains various sub-components, such as images and corresponding segmented images with features from each segment along with its label, among others.

The text portion contains descriptions of each image: an image may have one or more descriptions; however, the Spatial Information may not be present with each image

The training set contains 600 sentences, corresponding to 275 images. The test set contains 613 sentences, corresponding to 340 images. Each sentence in the training set has four types of annotations: the three types of Spatial Roles, and the Spatial Relation between the roles.

4 Implementation Details

The multimodal model utilizes feature from both Image and Text to bolster the classification result as the vital Spatial information link, which may not be available through a sentence alone is obtained through the low level feature from the image. The proposed approaches utilizes Neural network model with and without transfer learning process, to learn the relevant spatial information. It does not utilize the baseline model by [8], as is the case in other such systems. The detailed implementation of the model is described as under.

4.1 Finding Candidate Spatial Role Labels

As mentioned before, two of the Spatial Role Labels, Trajectors and Landmarks, are nouns and Spatial Indicators are Prepositions. For retrieving these part-of-speech tags for the words, BLLIP parser [1] is used. The process starts with identification of preposition PP, out of which the preposition and noun are considered as two major features, as explained in the previous section.

Thereafter, the other two candidate role labels are identified following the approach explained before. The chunking rule for extraction of NPPP phrase is written using the regular expression grammar.

4.2 Forming Candidate Spatial Relation Pairs

The identified sets of Spatial Role Labels for each sentences are paired together to form the candidates for Spatial Relations. However since each sentence can have multiple Spatial Indicators, thereby to get proper pairs only those candidate combinations are selected which followed dependency among the involved Spatial Role Labels. As mentioned before the candidate Landmarks are found from the dependency on prepositions from the preposition phrase.

For any sentence this pairing of Spatial Indicator and Landmark was formed initially which was further combined with the Trajector nouns. As the Trajector nouns are not having any direct syntactic dependency on either the Spatial Indicator nor the Landmark the pairing was formed with all the identified noun phrases for the sentence except the

Table 1. Performance in Spatial Role Labeling

	TR			LM			SI		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
VIEW[8]	0.513	0.755	0.611	0.354	0.555	0.432	0.727	0.757	0.741
VATE	0.302	0.152	0.202	0.229	0.123	0.160	–	–	–
Proposed Model	**0.761**	0.406	0.529	**0.760**	0.537	0.629	0.549	0.651	0.596

Table 2. Performance in Spatial Relation Identification

Model Name	**P**	**R**	**F1**
View	0.228	0.242	0.235
Visually Guided Spatial Relation Extraction from Text	**0.716**	**0.661**	**0.687**
Proposed Model (with transfer learning)	0.600	0.620	0.610
Proposed Model (without transfer learning)	0.539	0.496	0.516

Landmark nouns. Further to identify more possible pairs as per the training set which includes 'on the right', 'on the left', 'at the back', 'in front of' etc as Spatial Indicators such chunking rules were also incorporated.

4.3 Embedding Image and Text

Image feature of each individual images of the dataset used in the first approach are extracted through last layer of ResNet50 CNN architecture [3].

InceptionV3 [14] was used for extracting image feature for the second approach. Similarly text vector for the candidate roles as well the entire sentence in the first approach were generated through Spacy [13].

4.4 Extracting Topic Vectors from Images

MSCOCO 2014 training dataset was used to implement the topic model. Initially, it is required to obtain image categories of each image in the CLEF mSpRL dataset. These image categories are obtained in order to map the images onto MSCOCO dataset.

Therefore, an image is assigned the category of that image from MSCOCO that is at lowest vector distance from it (the distance was calculated as cosine similarity).

4.5 Feed Forward Neural Network

The feed forward neural network also known as Multi Layer Perceptron utilized here is initialized with random numbers in the range of 0 to 1 for its weights. Sigmoid Activation function was used as Transfer function. The evaluation of the algorithm was done using k-fold cross-validation with 5 folds. The network had 5 neurons in the hidden layer and 2 neurons in the output layer. The network was trained with a learning rate of 0.3.

The same network parameters were used for all four classification instance with different number of epochs for each of them. The Spatial Role labels were trained using 300, 200 and 100 epochs respectively followed by 100 epochs for Spatial Relations. The mean classification accuracy for each of them was more than 97%.

5 Result and Performance Analysis

The evaluation script provided in the SpRL track was used to generate the result which generates the scores in terms of Precision, Recall and F1. The proposed model is compared with other Multimodal Spatial Role Labeling systems which is shown in Tables 1 and 2. The model mentioned in the first row is considered as the baseline model which uses the Textual features and visually informed embedding of word trained through a different dataset named MSCOCO.

Out of the other results reported by them this result is compared with the proposed model as it follows similar approach of structured output that is prediction of Spatial Role labels initially and thereafter composing the words into triplets of (SI,TR,LM). The second model is built upon the baseline model's Role Labels, thereby no result has been mentioned for the role labels. The third model uses WordNet based multimodal alignment for finding the first two Spatial roles.

The proposed model is comparable to only these models as other models have not utilized the multimodal information. The precision score obtained in proposed model for two of the Spatial roles are higher than other models. Moreover there is no approach to increase the precision or recall for SI identification nor any attempt to hard code from a list of prepositions which are frequently found in the training set. The sole reason behind this was to demonstrate that the Roles are better extracted when they are relatively retrieved. This reduces the result of the SI identifier. However two of the prepositions which have no spatial information linked with it such as 'with' and 'by' are removed from consideration. This increases the precision of the Landmark significantly which in turn improved the precision of Trajector as well. The additional chunking rules utilized for identifying the valid candidate Spatial Relation pairs increased the recall of the model as the pairing of the Spatial Role Labels in Spatial Relation pair was not random rather was dependency based. The transfer learning approach shows better overall performance than the approach without it. This is because transfer learning approach pulls out more information from the image.

6 Conclusions and Future Work

Spatial Role Labeling approach followed in this work utilizes both forms of modality, texts and corresponding images. The efficient feed forward based neural network overcomes the unavailability of large annotated corpora and generates an efficient Spatial Information extraction system. It provides a new direction of progressive classification for identification of Spatial information.

In the second approach the transfer learning methodology also bolsters the idea of using multimodal information improves the retrieval of vital information that would not have been possible from a single modality. As future work the feature from individual objects in the image will be used alongside the topic vector for the image. This may further improve the efficiency of the retrieval process.

References

1. **Charniak, E. & Johnson, M. (2005).** Coarseto-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 173–180.
2. **Dash, S. K., Acharya, S., Pakray, P., Das, R., & Gelbukh, A. (2019).** Topic-based image caption generation. *Arabian Journal for Science and Engineering*, pp. 1–10.
3. **He, K., Zhang, X., Ren, S., & Sun, J. (2016).** Deep residual learning for image recognition. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, IEEE, pp. 770–778.
4. **Jones, G. J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., & Ferro, N.**, editors **(2017).** *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings*, volume 10456 of *Lecture Notes in Computer Science*. Springer.
5. **Kordjamshidi, P., Bethard, S., & Moens, M. F. (2012).** SemEval-2012 task 3: Spatial role labeling. *Proceedings of the First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 365–373.
6. **Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L. J., Shamma, D. A., & Bernstein, M. S. (2017).** Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73.

7. **Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollr, P., & Zitnick, C. L. (2014).** Microsoft COCO: Common objects in context. *Proceedings of European Conference on Computer Vision*, pp. 740–755.
8. **Ludwig, O., Liu, X., Kordjamshidi, P., & Moens, M. F. (2016).** Deep embedding for spatial role labeling. Preprint.
9. **Mazalov, A., Martins, B., & Matos, D. (2015).** Spatial role labeling with convolutional neural networks. *Proceedings of the 9th Workshop on Geographic Information Retrieval*, pp. 1–7.
10. **Pustejovsky, J., Kordjamshidi, P., Moens, M. F., Levine, A., Dworman, S., & Yocum, Z. (2015).** SemEval-2015 task 8: SpaceEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, pp. 884–894.
11. **Rahgooy, T., Manzoor, U., & Kordjamshidi, P. (2018).** Visually guided spatial relation extraction from text. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2*, Association for Computational Linguistics, pp. 788–794.
12. **Roberts, K. & Harabagiu, S. M. (2012).** UTD-SpRL: A joint approach to spatial role labeling. *Proceedings of the First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, pp. 419–424.
13. spaCy. Industrial-strength natural language processing in Python. [Online], `https://spacy.io`.
14. **Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016).** Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on Computer vision and Pattern Recognition*, pp. 2818–2826.
15. **Tsikrika, T., Popescu, A., & Kludas, J. (2011).** Overview of the Wikipedia image retrieval task at ImageCLEF 2011. *Proceedings of CLEF*, pp. 1–17.

Article received on 27/05/2020; accepted on 12/08/2020.
Corresponding author is Sandeep Kumar Dash.