Intergrating a Minimal Differentiator Expressions Approach into CBR for Linguistic Pattern reuse in Crime Relation: Proposed Method

IOP Conf. Series: Materials Science and Engineering **864** (2020) 012078 doi:10.1088/1757-899X/864/1/012078

Intergrating a Minimal Differentiator Expressions Approach into CBR for Linguistic Pattern reuse in Crime Relation: Proposed Method

M. Ikhwan Syafiq[1,1], M. Shukor Talib[2] , Naomie Salim[3] , Zuriahati Mohd Yunos[4] and Habibollah Haron[5]

[1,2,3,4,5]School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia.

E-mail: shukor@utm.my

Abstract. The relation extraction of crime news can help the monitoring specialists to accelerate the crime investigation. However, constructing patterns or designing templates manually requires domain experts. Also the built patterns do not guarantee complete differentiation among different relation instances. The automatic detection of crime entities and relationship among entities can help the regulatory authorities to accelerate the crime investigation and decision support instead of being reliant on manual process. This study aims to increase the effectiveness of the extraction of crime entities and relationship among entities based on the determination of crime lingusitic pattern using Minimal Differentiator Expressions (MDEs) that represent the cases that will be used by the CBR classifier. The proposed extraction methods can help in compiling a highly accurate and machine-understandable crime knowledge bases which can support the regulatory authorities' investigation. This paper conducted on our proposed MDEs algorithm for linguistic pattern reuse in CBR approaches.

1. Introduction

There are many pattern-based models proposed previously for relation extraction task. The patterns used in these models can be divided into two types. The first type of patterns are manually designed patterns such as [1][2][3] models which they used heuristic rule-based patterns, or lexical patterns designed manually, or shallow parsing for analysing the text to extract related entities. However, building patterns manually whether they are based on headwords or keywords or syntactic structure requires labour intensive effort and cost. If the knowledge grows; modifying these patterns can become very complex. The second type of patterns are automatic learned patterns such as [4][5] models which used co-occurrence approach or syntactic structure between prior known related entities pairs to learn the patterns; and then used the learned patterns to extract new potential related pattern-based models for relation extraction ; it was found that most of the existing models tend to focus on very strict linguistic or syntactic patterns or templates.

In the proposed model [6], the patterns have been learnt automatically from the related headwords or collected context to identify entities and ignoring some irrelevant contexts with the purpose without balance between generalizability and specificity of patterns to other domain. The propose method tend to be more specific which are more strict patterns. However, our purpose methods aim to employ a large

number of linguistic patterns that are more flexible than specific patterns. At the same time, it aims to preserve the precision since it is based on syntactic structural information of sentences and does not depend upon specific patterns which are more strict. Our propose model will utilize the Minimal Differentiator Expressions to exploit the ability of sustained learning. Thus, the model can accept new cases which is linguistic pattern without the need to re-train the model like machine-learning-based models. In addition, the MDEs has the potential to be further enhanced to gradually improved its performance without depend upon on specific patterns

1.1. CBR for pattern reuse

In the proposed relation extraction work, the task of identifying the crime relations between entity pairs considered as a classification problem, whereby the relation can be classified into one of the following: Cause, Action, or No Relation. The cases are represented as linguistic patterns for crime relation learnt from a set of sentences contain previously known relations between crime relation pair entities, and then saved in a case base. During use of the CBR model to solved new case (i.e., identify the relation between input crime entity), the system searches the case base for case(s) most similar to the case problem (i.e., retrieve the relevant linguistic pattern). The proposed method incorporates the adaption of the standard CBR algorithm that is tailored to the classification task.

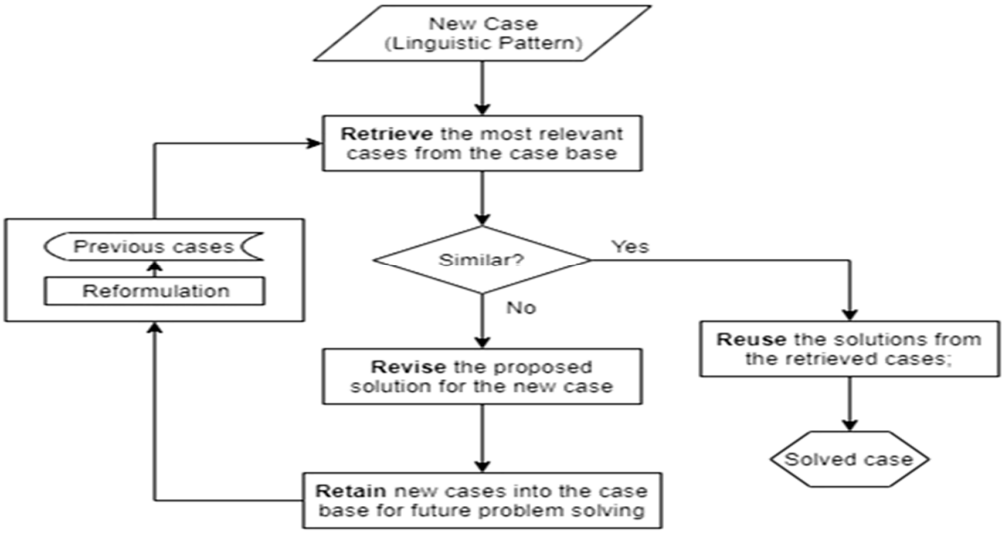

Fig. 1 The general cycle of CBR

The general cycle of CBR consists of four main sub processes, namely Retrieve, Reuse, Revise, and Retain that are linked to a knowledge repository called the case base as display in Figure 1. When a new case (problem) received, the CBR model will first retrieve one or more previously experienced similar cases from the case base. Secondly, reuse the solution from the retrieved cases for the new case; thirdly revise the solution for the new case and finally retain the revised solution by incorporating it into the existing case base [7][8].

2. MDEs for Linguistic Retrieval

2.1. Overview

In this work we are concerned in extracting dynamically, i.e. without predefined syntactical structures, patterns that might depict frequent thoughts expressed by users talking about a given subject. These patterns could be used for instance, to mine online text news and opinions about name of suspect, type of weapon, organizations, location, among others. Likewise, in this study, each case in the designed case base represents an example of <PER, PER>, <PER, GPE>, <GPE, PER>, <PER, TIME>, <TIME, PER> pair will be labelled by linguistic pattern(s) or expression(s). Figure 2 shows with example how the cases are formed from the sentence.

2.2. Generate Linguistic Pattern

Firstly, every sentence will be first pre-processed, i.e. by removing punctuation marks and lowercasing. Then it undergo Web Crawl program for tagging the names entities in crime domain. Followed by parsing the tagged sentences and removing any crime entities existing in the DRPs. Secondly, extract a part of dependency relation path (DRP) that connect every entity pairs to produce the representative tokens according to the linguistic pattern for each sub DRP are generated automatically by Algorithm 1-5. Thus, linguistic pattern represent the cases that will be used by the CBR classifier.

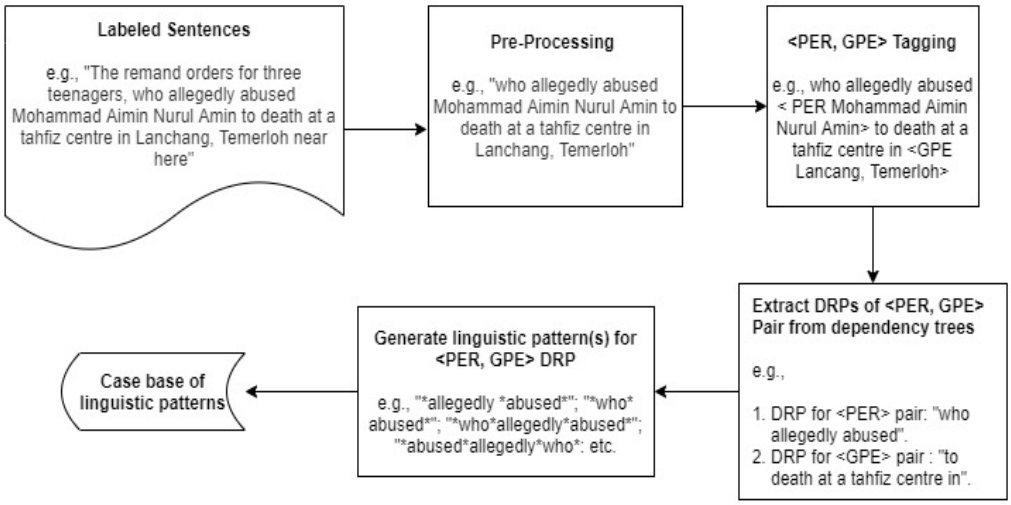

Fig. 2 Example of extracted information; NER and Relation extraction task

2.3. Minimal Differentiator Expression (MDEs) for Linguistic Retrieval

The Minimal Differentiator Expressions (MDEs) algorithm proposed in [9] for Frequency Answered Question (FAQ) retrieval will be adapted in this work to automatically generate a set of linguistic patterns (expressions). Thus, the MDEs is used for looking up for relevant or expressions. In adapting MDE algorithm, the following definition should be considered: Definition (1): "Expression: Being S a sentence, e is said to be an Expression of S, noted as "e exp S", if:

(i) It is composed by a subset of the words in S arranged in the same order.
(ii) Not all its words are stop words.

$$e \text{ exp} \leftrightarrow S \begin{cases} Words\ (e) \subseteq Words(S) \Lambda\ Order\ (S,e) \\ \Lambda \\ Words\ (e) \nsubseteq StopWords \end{cases} \quad (1)$$

Where Words () is the set of words that compose an expression or a sentence, and Order () is the Boolean expression that verifies that shared words are arranged in the same order. The set of all the expressions of a given sentences S is denoted as:" [9].

$$E(S) = \{\, e | e \exp S \,\} \tag{2}$$

An example of correct expressions consist of a subset of words in the sentence:

"The remand orders for three teenagers, who allegedly abused Mohammad Aimin Nurul Amin to death at a Tahfiz Centre at Lanchang Temerloh near here."

Are {*who*}; {*allegedly*}; {*abused*}; {*death*}; {*who*allegedly*}; {*who*abused*}; {*allegedly*abused*}; where the character "*" denoted for zero or more word. But the expression {*to*a*}; is not correct because all of its words are stop words despite of they are arranged in the same order as in the sentence. At the same time, the expression {*who*abused*allegedly*}; is not correct expression because its words are not in the correct order.

In representing the cases, the MDEs (i.e., which will formed as a regular expressions) will be built from parts of DRPs connect every i.e., <PER, PER> pair(s) with crime relation in each sentence. Thus each sentence will compose at least one sub sentence equivalent to S

Definition (2): "Differentiator Expression: An expression e is a Differentiator Expression (DE) of the sentence S if it unambiguously distinguishes S from the rest of sentences in the case B (not necessarily from the reformulations contained in its own case C). DE is defined as follows:

$$e\ exp_{DE}\ S\ wrt\ B \leftrightarrow \begin{cases} e \in E\ (S) \\ \Lambda \\ \forall \acute{S} \in B, \acute{S} \notin C, e \notin E(\acute{S}) \end{cases} \tag{3}$$

Where *wrt* is "with respect to", and $e\ expDES \in$ C, C represents that e is a differentiator expression of S. Finally, the set of all differentiator expressions of S with respect to B is denoted as:

$$DE\ (S, B) = \{e\ |e\ exp_{DE} S\ wrt\ B\} \tag{4}$$

Definition (3): "Minimal Differentiator Expression: Let e be a differentiator expression; a Minimal-Differentiator-Expression (MDE) is a differentiator expression e which does not contain any other differentiator expression é" [9]. All MDEs in a sentences S with respect to a case base B is denoted as:

$$DE\ (S, B) = \{e\ |e\ exp_{DE} S\ wrt\ B\} \tag{5}$$

From the above example of expressions; if the expression {*abused*} can differentiate the sentence from other sentences in case base; the expression {*who*abused*}; which has fewer number of words and allow differentiation. Thus the advantage of DEs is the prevention of linguistic interferences and overlapping between templates or expressions which is difficult to detect manually.

The MDEs algorithm calculates the $MDE(S,B)$ set for each sentence S in a case base B. As mentioned above, in this work, the sentence is represented as a part of DRP connects entity pair with specific relation. Thus, each sentence is considered as a prototype in the supervised learning algorithm [9].

3. Conclusions

This propose method is to employ a large number of linguistic patterns that are more flexible than specific patterns. At the same time, it aims to preserve the precision since it is based on syntactic structural information of sentences and does not depend upon specific patterns which are stricter. Therefore, this combination has a trade-off between the generalizability and strictness of patterns. At the same time, the proposed model will utilize the minimal differentiator expressions to exploit the ability of sustained learning. Thus, the model can accept new cases without the need to re-train the model like machine-learning-based models. In addition, the CBR has the potential to be further enhanced to gradually improved its performance.

Acknowledgement

This work was financially supported by a UTMER grant, cost center no 17J47 and Faculty of Engineering, Universiti Teknologi Malaysia (UTM).

References

[1] P. Das and A. K. Das, *Advanced Computational and Communication Paradigms*, vol. 475. Springer Singapore, 2018.

[2] X. Wang, X. Jiang, M. Liu, T. He, and X. Hu, "Bacterial named entity recognition based on dictionary and conditional random field," *Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017*, vol. 2017-Janua, pp. 439–444, 2017.

[3] C. Giannone, R. Basili, C. Del Vescovo, P. Naggar, and A. Moschitti, "Kernel-based relation extraction from investigative data," p. 93, 2009.

[4] S. K and P. S. Thilagam, "Crime base: Towards building a knowledge base for crime entities and their relationships from online news papers," *Inf. Process. Manag.*, vol. 56, no. 6, p. 102059, 2019.

[5] P. Das and A. K. Das, "Graph-based clustering of extracted paraphrases for labelling crime reports," *Knowledge-Based Syst.*, vol. 179, pp. 55–76, 2019.

[6] O. Chergui, A. Begdouri, and D. Groux-Leclet, "Integrating a Bayesian semantic similarity approach into CBR for knowledge reuse in Community Question Answering," *Knowledge-Based Syst.*, vol. 185, p. 104919, 2019.

[7] A. Hassanien, *Advances in Intelligent Systems & Computing 1058 Proceedings of the International Conference on Advanced Intelligent Systems & Informatics*. 2019.

[8] E. C. Lopes and U. Schiel, "Integrating context into a criminal case-based reasoning model," *2nd Int. Conf. Information, Process. Knowl. Manag. eKNOW 2010*, pp. 37–42, 2010.

[9] A. Moreo, M. Navarro, J. L. Castro, and J. M. Zurita, "A high-performance FAQ retrieval method using minimal differentiator expressions," *Knowledge-Based Syst.*, vol. 36, pp. 9–20, 2012.