Chemical-Induced Disease Detection Using Invariance-based Pattern Learning Model

Neha Warikoo[123], Yung-Chun Chang[4] and Wen-Lian Hsu[3]*

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei, 112, Taiwan
[2]Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 115, Taiwan
[3]Institute of Information Science, Academia Sinica, Taipei 115 Taiwan
[4]Graduate Institute of Data Science, Taipei Medical University, Taipei 106, Taiwan

Abstract

In this work, we introduce a novel feature engineering approach named "algebraic invariance" to identify discriminative patterns for learning relation pair features for the chemical-disease relation (CDR) task of BioCreative V. Our method exploits the existing structural similarity of the key concepts of relation descriptions from the CDR corpus to generate robust linguistic patterns for SVM tree kernel-based learning. Preprocessing of the training data classifies the entity pairs as either related or unrelated to build instance types for both inter-sentential and intra-sentential scenarios. An invariant function is proposed to process and optimally cluster similar patterns for both positive and negative instances. The learning model for CDR pairs is based on the SVM tree kernel approach, which generates feature trees and vectors and is modeled on suitable invariance based patterns, bringing brevity, precision and context to the identifier features. Results demonstrate that our method outperformed compared approaches, achieved a high recall rate of 85.08%, and averaged an F_1-score of 54.34% without the use of any additional knowledge bases.

1 Introduction

Causality or association determination between target entities, especially those involved in diseases, has quickly become the topic of interest within the area of biomedical text mining. Such studies have created a large number of information pools that enables clinicians to make diagnoses more effectively. A pertinent example is the prediction of chemical-disease interactions based on biomedical text, which if used to its fullest potential can revolutionize the way precision medicine and drug testing is conducted. The idea is to preemptively identify any associations between a drug and subsequent physiological responses for subjects accepting treatment for a disease (Wei et al. 2015). Most of the physiological responses emerge as secondary disease symptoms and often as adverse drug reactions. If studied in appropriate context, these events may contain information of unwanted damages to the patients. Any side effects or adverse drug reactions can be avoided for patients participating in clinical trials if similar trials have had invoked deleterious responses in its participants, which can be heuristically implied by textual and statistical evidence presented in scientific publications and other approved research materials.

Recently, BioCreative V introduced the task of chemical-induced disease (CID) relation extraction from PubMed abstracts, focusing on identifying chemical and disease entities acting in a "cause and effect" mannerism in a binary association. We have adapted the same task guidelines to shape our objective of identifying chemical-induced diseases via pattern-based learning. Our approach for the CID task attempts to capture the commonalty in structured patterns used to describe such relations. Corresponding positive and negatives instances from abstracts are processed as vector representations converged into signature patterns that can be accessorized as identifiers for the nature of the relationship. The generated patterns are then learned by using the convolution tree kernel (CTK) to classify potential entity pairs.

Unlike other relation extraction tasks, the vague context of associating entities in the sentences generated by intra-sentential and inter-sentential association pairs increases the complexity of this dataset. In an intra-sentential scenario, the chemical-induced drug response is explicitly given within a sentence. An example is the relation between "cocaine" and "myocardial infrac-

* Corresponding author

Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), pages 57–64,
Taipei, Taiwan, November 27, 2017.

tion and bundle branch block" shown in Figure 1 (a). As for inter-sentential cases, the association statements can span across several sentences. Figure 1 (b) indicates the specific effects of audiovisual toxicity caused by "desferrioxamine" can only be established by parsing multiple statements describing the secondary links identified through the perception of "audiovisual defects".

(a): Intra-sentential

Electrocardiographic evidence of myocardial injury in psychiatrically hospitalized *cocaine* abusers..... Eleven of the *cocaine* abusers and none of the controls had ECG evidence of significant myocardial injury defined as *myocardial infarction*, ischemia, and *bundle branch block*.

(b): Inter-sentential

Ocular and auditory toxicity in hemodialyzed patients receiving *desferrioxamine*. During an 18-month period of study 41 hemodialyzed patients receiving *desferrioxamine* (10-40 mg/kg BW/3 times weekly) for the first time were monitored for detection of audiovisual toxicity..... Visual toxicity was of retinal origin and was characterized by a tritan-type dyschromatopsy, sometimes associated with a loss of visual acuity and *pigmentary retinal deposits*. Auditory toxicity was characterized by a mid- to high-frequency neurosensorial hearing loss and the lesion was of the cochlear type..... Please make both abstracts as a figure.

Figure 1: Intra-sentential and inter-sentential cases from the corpus.

To resolve this complexity, we developed a preprocessing module to identify sentences based on permutations of all possible entities. In addition, linguistic patterns were learned from biomedical literatures based on the concept of Algebraic Invariant. These patterns are provided to SVM based on convolution tree kernel as features for supervised learning. Depending on the characteristics captured by the patterns, the classifier aims to differentiate instances involving related and unrelated entity pairs.

2 Related Work

Since its inception, multiple learning approaches were employed with and without Knowledge Base (KB) to simplify the CID task. Zhou et al. (2015) used a shortest dependency tree-based method for relation extraction in the CDR corpus. They experimented with flattened features, structured features, and structured phrases and reported a F_1-score of 55.05% with a combination of all of the features. The approach of Pons et al. (2015) for the same task is based on their feature set established on a prior graph database for chemical-disease interaction along with separate sets of statistical and lexical features. Over a dozen lexical and dependency path based features were exploited by Gu et al. (2015) to demonstrate the effectiveness of intra-sentential and inter-sentential level classification using a Maximum Entropy model. Xu et al. (2015) utilized a knowledge base-targeted method in learning the relation patterns. In addition, they also employed context-based features along with some auxiliary features to short list the number of relations for sentence level and ultimately document level classifier. Le et al. (2015) applied a pipeline model based on co-reference resolution and intra-sentential relations. Based on the entity pairs recognized by the model, token dependent features, n-gram word features and graph-based features SVM classifiers were used for relation identification. Chemical-disease relations identified in the CTD[1] database were incorporated in lexical feature vectors by Alam et al. (2015) to add higher confidence value to significant features based on their collective mentions in the database. Moreover, Zhou et al. (2016) used variants of the neural network method to obtain performances ranging from 47.2 with a convolution neural network model to 61.3 with a hybrid model of tree-kernel based SVM, LSTM, and a post-processing module. As an extension of the neural network approach for this task, Gu et al. (2017) introduced another model based on their previous effort (Gu et al. 2015) in which they used ME to determine intra-sentential relations and a convolution neural network model for inter-sentential relation recognition. A post-processing module that removes redundancies and adjusts hypernyms was implemented to enhance the model.

Our model is KB independent with a SVM tree kernel learning method. It focuses on customizing the context of the learning tree to application relevance through our novel algebraic invariant pattern generation approach.

3 Method

The task of CID identification mandates pre-annotation of chemical and disease entities throughout the text. The organizers have used manual annotation along with *tmChem* (Leaman et al. 2015) and *DNorm* (Leaman et al. 2013) for chemical and disease term identification. In order

[1] https://toxnet.nlm.nih.gov/newtoxnet/ctd.htm

to focus on relation extraction, we decided to use the pre-annotations given in the training, development, and test datasets for generating possible relation entity pairs in each respective set. To develop a classifier for recognizing related entity pairs, we divided our pattern learning effort into three different stages. The first stage is the preprocessing of biomedical text followed by candidate sentence selection. With the help of these candidate sentences, relevant context based patterns are exhumed from the original text. Values based on these patterns are used as coefficients variables in invariant polynomial function to cluster similar ranking patterns. Similar ranking patterns are aligned and restructured into a more generic form. Each of these patterns is used to generate feature file for SVM based tree kernel, in which everything except regional matches to context-based patterns are pruned. SVM classifier predicts the corresponding labels for the hence generated candidate instance based trees to determine the relation between the entity pairs. Each stage is illustrated in details in the subsequent sections.

3.1 Candidate Instance Generation

The abstract data in its initial form contains multiple entities associated with either intra-sentential or inter-sentential relations, thereby increasing the difficulty of this task. Moreover, other existing sentences may become noises as they do not co-relate or attribute in any form in determining entity pair relations. Candidate Instance Generation entails screening for relation-oriented sentences, which are referred to as "Instances" henceforth. Prior to candidate instance generation, we proceeded with generic tasks of natural text preprocessing via Sentence Detection (Apache Open NLP)[2], Entity Class Labeling (In-built Module), and part-of-speech (POS) tagging (Genia Tagger)[3]. Moreover, we resolved duplicate adjacency entity labels (In-built Module), which are often observed in biomedical literature. For example, although "plasma renin activity (PRA)" is annotated with two separate labels "plasma renin activity" and "(PRA)", but they both correspond to the same bio-entity as the bracketed acronym mentioned in adjacency is a duplicate label. Resolving such duplicates optimizes the pair-based instance generation task.

We choose to generate candidate instances from POS tag-labeled sentences since they are more appropriate in depicting the skeletal similarity of relation expressions in contrast to natural text. Therefore, following the preprocessing, the POS-tagged data was drafted into candidate instances based on entity pairs (one chemical and one disease mention per sentence per pairwise iteration) relabeling to indicate the primary Chemical and Disease pair. The verb implying the relation (proximal verb) was also assigned a prominent identifier as shown in Figure 2.

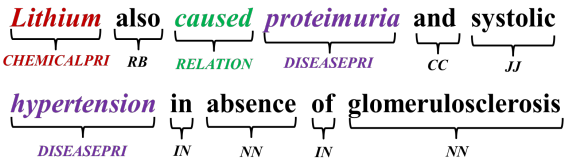

Figure 2: Candidate sentence tagging.

There can be more than one entity pair relations within a sentence. Therefore, for each pair set, a duplicate instance of the sentence highlighting the relevant pair is generated. Other than key entity pairs, we also identified and annotated the proximal verb with a third term "Relation". It entails a non-basal form verb nearest to the current entity pair set. The rationale of using this verb form is that in most sentences describing bio-entity relations, causal relations are asserted in a smaller frame within the sentence. Even in complex sentences, subject and the acted object are often linked by non-basal form verbs in close vicinity to the actors. Given the related entity pairs in the training data, positive and negative instances are generated and later processed by the successive feature-engineering module.

3.2 Invariance-based Feature Engineering

In our approach, we propose that different candidate instances show similarity in subject inference even if they are structurally diverse when relevant contexts are provided as reference points. There are multiple ways to communicate the same idea in a language, whether by direct implication or at times with additional context or compound references. However, in each of these cases, the skeleton of some reference points stays the same across different sentence structures. Our idea is to demonstrate the invariance or lack of change in the nature of such descriptive sections from the

[2] http://opennlp.sourceforge.net/models-1.5/en-sent.bin

[3] http://www.nactem.ac.uk/tsujii/GENIA/tagger/geniatagger-3.0.2.tar.gz

text, and exploit this characteristic in generating more robust features while limiting the degree of evaluation function.

The idea is heavily drawn on Algebraic Invariance to show that two separate sentences are similar in their inferential meaning if their invariant function does not vary. Such a function can be represented as follows:

$$I(q_{n0}\cdots q_{0n}) \equiv \Delta^{W} * I(p_{n0}\cdots p_{0n}) \qquad (1)$$

where *I(q)* and *I(p)* indicate the invariant function , Δ is the determinant of the representational polynomial undergone transformation, and *W* is the invariant weight. Any object/element can be represented in the Euclidean system using a polynomial function $P(x,y) = \sum p_{ij}x^{i}y^{j}$. Upon transformation "*T*", the same polynomial can be represented by another polynomial $Q(u,v) = \sum q_{ij}u^{i}v^{j}$, bound in relation *(u,v) = T(x,y)* with original form.

In order to restructure the invariance concept in a natural text paradigm, we used an assumed homogenous polynomial function based on three key referential groups viz. Entity1 (chemical), Relation (proximal verb), and Entity2 (disease) to project every instance in the Euclidian space. Since our primary goal is to identify chemical-induced diseases, we limited our function to a second order polynomial based on each variable set as given below:

$$P(x,y) = p_{20}x^{2} + p_{11}x^{1}y^{1} + p_{02}y^{2} \qquad (2)$$

where *x* and *y* are representational binary association variables indicative of the "Entity1~Relation" and "Entity2~Relation" set, respectively. p_{20}, p_{11}, and p_{02} are coefficients of the representative polynomial evaluated by the maximum value from a five-frame adjacency matrix vector for each of the corresponding variable pairs. Our algorithm treats each candidate instance polynomial as a transformed version of all other instance polynomials. According to the concept of invariance, if the invariant functional of the current candidate polynomial is equal to the invariant functional of other instance polynomials, then the current instance is considered similar to each of those instances, thereby reducing the dimensionality of screening space for pattern generation and keeping context-specific similarities. In order to determine the invariant function, we assume rotation (ϕ = 0) as transformation for our polynomial to calculate the corresponding invariant function for the given second order polynomial (Keren (1994)). The equation for calculating invariant polynomial in the assumed case is given below:

$$I(q_{n0}\cdots q_{0n}) \overset{\text{def}}{=} I(p_{no}\cdots p_{0n}) = \left[p_{20}^{2} + \left(\frac{p_{11}^{2}}{2}\right) + p_{02}^{2}\right] \qquad (3)$$

where *I(q)* and *I(p)* are the invariant functions for the transformed instance polynomial *Q(u,v)* and original instance polynomial *P(x,y),* respectively. p_{20}, p_{11}, and p_{02} are the coefficients of the original polynomial function *P(x,y)*. Every candidate instance is screened for each of the three key referential groups as shown by candidate instances 1 and 2 in Figure 3, in which each underlined portion conforms a group of context patterns. Based upon their polynomial correspondence, the two candidate instances can be represented on coordinate space as demonstrated in Figure 4. The instances are also construed as proximal or non-proximal in structure depending upon the invariant scores. If they are similar, a generic context pattern can be obtained from both of them through alignment for each group.

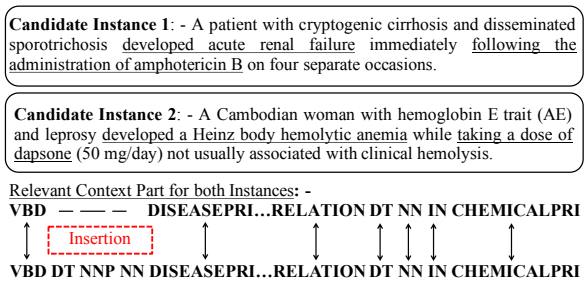

Figure 3: Context identification and prospective alignment of candidate instances.

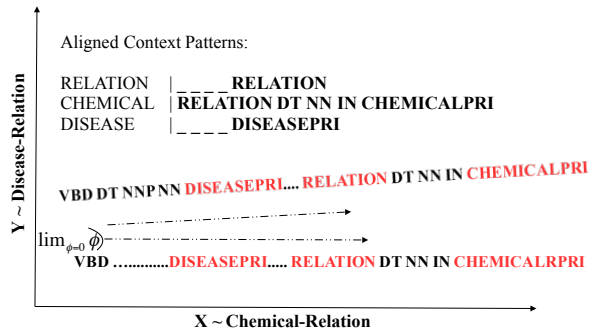

Figure 4: Vector representation of candidate instances on coordinate space.

For every referential group, 5 POS-tagged contextual frames with a range of 5 are generated by shifting the window frame iteratively over the instance, moving group index from 1 through 5. Then for each frame size, an adjacency matrix is generated per referential group by matching identical context patterns across instances to provide a statistical significance value for every instance as displayed in Figure 5. In addition to the repetitive count of contextual frames, each frame is individually scored to evaluate its significance. The context is scored using n-gram probabilistic model where *n* is the index of the referential entity group.

Since the index for reference group varies as per the frame being used, we have slightly modified the formula to accommodate the significance of whole patterns over the sub-patterns in the equation below. The modified formula takes into account all of the variant n-gram patterns both succeeded and preceded by the current referential group context. In this manner, it is able to attribute a more accurate representational value of the particular pattern from the entire context sample space.

$$\rho_{e_c k} = \begin{cases} \left(\frac{\sum P(x_0 \cdots x_{e_c})}{\sum P(x_0 \cdots x_{e_c-1})}\right) + \left(\frac{\sum P(x_0 \cdots x_n)}{\sum P(x_0 \cdots x_{e_c})}\right) \\ \forall\, e_c < 5, n = 5 \\ \left(\frac{\sum P(x_0 \cdots x_{e_c})}{\sum P(x_0 \cdots x_{e_c-1})}\right) + \delta_v \approx 0.0000000001 \\ \forall\, e_c = 5, n = 5 \end{cases} \quad (4)$$

where e_c is the index of the current referential group and n is the total size of frame. $\rho_{e_c k}$ is the score for each cell with frame size e_c and candidate instance k. $\sum P(x_0 \ldots x_n)$ is the number of times the current extracted POS frame of size 5 has occurred across all of the contexts generated from all instances. δ_v indicates the fringe value used in case the referential group has a terminal index. It is introduced to avoid attributing excess weight for standard n-gram patterns.

As indicated in Figure 5, since the variables in our polynomial equation (2) are based on binary association between entities, therefore the scores generated for each referential group are summed up with their corresponding pair variable score from the equation to evaluate the conjugate coefficient. Corresponding coefficient values from the homogenous representation equation (2) are substituted in equation (3) to obtain the invariant

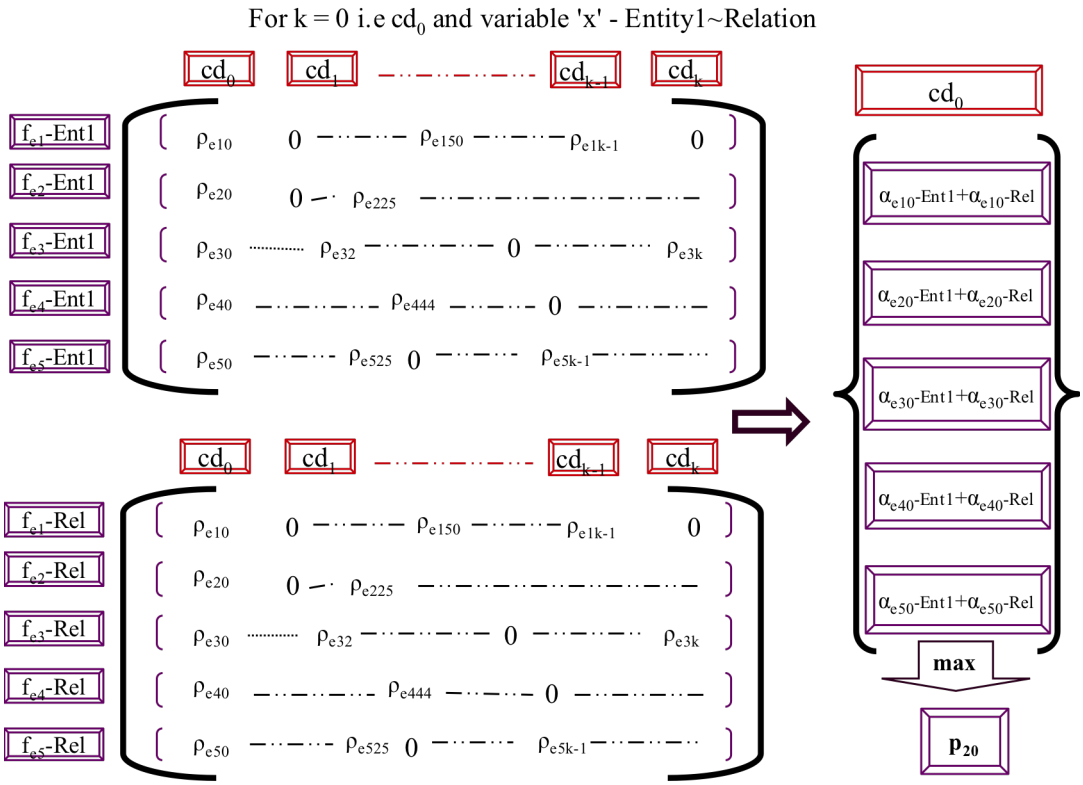

Figure 5: Adjacency matrix for calculation of representational polynomial coefficient

function score $I(p)_k$ in which k is the current candidate instance ID. The instances are then ranked in the descending order based on the calculated scores. According to equation (1) (set Δ=1.00 and W~1), each candidate instance is compared with its successor. If the approximation of values is similar, then the instances are considered as structurally similar and clustered together to form a feature attribute for classification. Otherwise, they are diversified into different pattern groups as illustrated in Figure 6. The clustered instances were used in pattern generation. Individual alignments were performed between instances for each of the referential groups (i.e. "Entity1-Chemical", "Relation-Proximal Verb", and "Entity2-Disease") to generate a triple context set-based pattern. The alignment is based on the highest scoring path obtained from the substitution matrix delivered by the recurrence relation:

$$sim(i,j) = \begin{cases} sim(i,j-1) + \\ \lambda(\_,j) \approx -2 \\ sim(i-1,j-1) + \\ \lambda(i,j) \approx \begin{vmatrix} 1 \,\forall\, i = j \\ -1 \,\forall\, i \neq j \end{vmatrix} \\ sim(i-1,j) + \\ \lambda(i,\_) \approx -2 \end{cases} \quad (5)$$

where *sim(i,j)* is the similarity score of the i^{th} row and j^{th} column of the substitution matrix. $\lambda(i,j)$ indicates the penalty function scoring insertion, deletion, match or mismatch depending on the token comparison of respective indexes.

3.3 Tree Kernel Induced Learning

The stringent context patterns retrieved from invariance functions are mapped against the candidate instances, and the optimal match of each case is selected to define a feature tree for learning. Parse tree is generated using the Stanford parser (Chen and Manning 2014; Socher et al. 2013) and the selected context-based pattern determines which leaf nodes are to be pruned to refine the context of the tree. The tree is decorated through highlighting the instances with CID by prefixing a node for such positive instances. Along with the parse tree, a feature vector corresponding to each candidate instance is also maintained to examine the similarity of phrase structures (both simple and complex). Each phrase structure is characterized with an "ID", and the feature vector maintains the count of corresponding phrase structures per instance. A combination of the parse tree and feature vector is used in developing and testing the model. To classify the phrase structures according to the similarity index, Convolution Tree Kernel is employed to compare the substructures across parsed instances. SVM-Light-TK-1.5[4] toolkit was used in both the learning and classification modules (Moschitti 2004, 2006).

Algorithm 1: Invariance Pattern Generation

INPUT:
context$P$: $P_{Relation|Chemical|Disease}(x_0....x_n)_k$ triplet pattern for all candidate instances
ordered$I(P)$: Invariant functional score $I(P)_k$ for all candidate instances in descending order
BEGIN
1: set seed$I(P)$ = ordered$I(P)_0$
2: set seed$P$ = contextP_0
3: ***FOR*** k=0 : size(ordered$I(P)$)
4: curr$I(P)$ = ordered$I(P)_k$
5: curr$P$ = contextP_k
6: invarQuotient = (seed$I(P)$/ curr$I(P)$)
7: ***IF*** invarQuotient == 1.0
8: reset seed$P$ = ***align*** seed$P$ with curr$P$
9: *remove(* ordered$I(P)_k$ *)* & k = k-1| k!= 0
10: ***ELSE***
11: ***IF*** seed$P$ exists in InvariancePatterns
12: *remove(* ordered$I(P)_k$ *)* & k = k-1
13: ***ELSE***
14: *add(*seed$P$*)* to InvariancePatterns
15: seed$I(P)$ = curr$I(P)$
16: seed$P$ = curr$P$
17: END ***FOR***
18: *add(* seed$P$*)* to InvariancePatterns
OUTPUT: InvariancePatterns
END

Figure 6: Algorithm for Invariance Based Pattern Identification.

4 Experiments

4.1 Experiment Setup

We chose to adapt the CDR corpus released for BioCreative V – Track 2 to evaluate our method. The corpus comprises of 1500 PubMed abstracts in total, out of which 1400 abstracts were selected from the CTD-Pfizer collaboration corpus, with the remaining ones as new curations. The

[4] http://disi.unitn.it/moschitti/TK1.5-software/download.html

abstracts were equally distributed among the training, development, and test sets. Chemical and disease mentions were annotated and normalized to the corresponding MESH IDs (Li et al. 2016). Known chemical induced disease relations, determined from both the title and abstract text, were appended with each document ID. We did not conduct additional Named Entity Recognition (NER), and simply performed our analysis on the entities predefined in the dataset. Statistics on the entities and relation pairs within the corpus is displayed in Table 1. We evaluated the performance of relation detection in terms of the precision, recall, and the F_1-score. The F_1-score is the harmonic mean of the precision and recall, and is often selected to determine the overall effectiveness of a system.

Dataset	#Chemical	#Disease	#Relation
Train (500)	4182	5203	1038
Dev. (500)	4244	5347	1012
Test (500)	4424	5385	1066

Table 1: CDR Corpus Statistics

4.2 Results and Discussion

The performance of our method was compared with different approaches used for CID detection. Systems developed by Xu et al. (2015), Alam et al. (2015), and Pons et al. (2015), (Table 2) were based on using external KBs for relation pair identification. Xu et al. (2015) coupled the relation pair information from CTD, MEDI, and SIDER along with context-based features to optimize the learning and obtained F_1-score of 57.03%. Alam et al. (2015) developed a binary feature vector set model based on various characteristics exhibited by the entity pairs in abstracts. They utilized statistically significant relation pairs from the CTD database as one of the signal features to augment the confidence value for such feature sets. They achieved a high recall of 81.03% and averaged about 52.77% on F_1-score. Pons et al. (2015) employed the graph database (BRAIN) to screen out candidate relation pairs which were directly or indirectly associated with each other.

Le et al. (2015) and Li et al. (2015) both used SVM for classification based on various sets of lexical features. In one of the recent attempts on this task, Gu et al. (2017) applied a hybrid CNN and ME based model to handle multi and single sentence level relation pairs to acquire a performance of 61.30%. Although their model did not involve any KB-based refining, but post-processing strategies for filtering the relation pairs were employed. Our approach is also a ML dedicated approach in which the extracted patterns were used to develop SVM features based on the convolution tree kernel for learning. In contrast to all of the other methods, our approach achieved the highest recall rate of 85.08%, signifying that the features generated by our Invariance approach can identify positive association pairs with a higher specificity. Our F_1-score averaged at 54.34%, which is an overall satisfactory score but still falls short against the better systems based on KB and NN. The gap in performance can be overcome by improving the precision through additional resources to normalize the negative features. Holistically, our method as a feature-engineering tool is concise, more precise and feature flexible in comparison to other metrics used for feature generation. It can accommodate multiple features to adjust the size of polynomial and reduce the complexity in the evaluation of classifiers to make them more feasible, including simpler linear classifiers as well. The flexibility and power of this approach makes it an efficient choice for implementation with any learning algorithm.

Method	Precision	Recall	F_1-score
Li et al.	54.46	33.21	41.26
Le et al.	53.41	49.91	51.60
Alam et al.	39.12	81.03	52.77
Pons et al.	51.30	53.90	52.60
Xu et al.	55.67	58.44	57.03
Gu et al.	**55.70**	68.10	**61.30**
Our method	39.92	**85.08**	54.34

Table 2: Comparative Assessment of the CID task

5 Conclusion

This paper describes a novel method of feature engineering based on algebraic invariance, which in conjunction with SVM tree kernel-based approach is effective in identifying relation pairs for the CID task. Comparative analysis demonstrates that the method is powerful enough to identify diverse patterns/features within any corpus set without auxiliary resources. Therefore, we conjecture that our method as a feature generation tool can be highly effective and easily adoptable in various application scenarios.

For further enhancement in the future, we plan to use deep learning methods for model de-

velopment in conjunction with our approach to gauge the variability introduced by the various learning models in context of our method. Furthermore, we also plan to enlist context-specific knowledge bases to optimize our feature sets and improve the overall performance.

Acknowledgments

We are grateful for the constructive comments from three anonymous reviewers. This work was supported by grant MOST106-3114-E-001- 002 and MOST105-2221-E-001-008-MY3 from the Ministry of Science and Technology, Taiwan.

Reference

Danqi Chen and Christopher D Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP 2014*

Leonard Eugene Dickson. 1914. Mathematical Monongraphs Algebraic Invariants, No.14. *John Wiley*, New York.

Daniel Keren. 1994. Using Symbolic Computation to Find Algebraic Invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16, No 11*, Nov 1994.

Jinghang Gu, Longhua Qian, Guodong Zhou. 2015. Chemical-induced Disease Relation Extraction with Lexical Features. *Proceeding of the fifth BioCreative challenge evaluation workshop*, 2015

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2016. Chemical-induced disease relation extraction via convolutional neural network. *Database (2017)* Vol. 2017

Leaman R, Wei C-H, Lu Z. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*. 2015;7(Suppl 1):S3. doi:10.1186/1758-2946-7-S1-S3.

Robert Leaman, Rezarta Islamaj Doğan, Zhiyong Lu; DNorm: disease name normalization with pairwise learning to rank, *Bioinformatics*, Volume 29, Issue 22, 15 November 2013, Pages 2909–2917

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer, A. Schein and L. Ungar. 2004. Integrated Annotation for Biomedical Information Extraction, *HLT/NAACL 2004 Workshop: Biolink 2004, pp. 61-68*.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J.Mattingly, Thomas C. Wiegers, Zhiyong Lu. 2015. Annotating chemicals, diseases and their interactions in biomedical literature. *Proceedings of the fifth BioCreative challenge evaluation workshop*, 2015.

Alessandro Moschitti. 2006, Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. *Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany, 2006.*

Alessandro Moschitti. 2004. A study on Convolution Kernels for Shallow Semantic Parsing. *Proceedings of the 42-th Conference on Association for Computational Linguistic (ACL-2004), Barcelona, Spain, 2004.*

E. Pons, B.F.H. Becker, S.A. Akhondi, Z. Afzal, E.M. van Mulligen, J.A. Kors. 2015. RELigator: Chemical-disease relation extraction using prior knowledge and textual information. *Proceeding of the fifth BioCreative challenge evaluation workshop*, 2015

Richard Socher, John Bauer, Christopher D. Manning and Andrew Y. Ng. 2013. Parsing With Compositional Vector Grammars. *Proceedings of ACL 2013*

Yoshimasa Tsuruka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text, *Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005*

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data, *Proceedings of HLT/EMNLP 2005, pp. 467-474*

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Jiao Li Thomas C. Wiegers and Zhiyong Lu. 2015. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database,* 2015.

Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Ruiling Liu, Qiang Wei, and Hua Xu. 2015. UTH-CCB@BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text . *Proceeding of the fifth BioCreative challenge evaluation workshop*, 2015

Huiwei Zhou, Huijie Deng, Jiao He. 2015. Chemical-disease Relations Extraction Based on The Shortest Dependency Path Tree. *Proceeding of the fifth BioCreative challenge evaluation workshop*, 2015