VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

LE HOANG QUYNH

PHENOTYPE-DISEASE SEMANTIC RELATION EXTRACTION IN BIOMEDICAL TEXT

DOCTORAL DISSERTATION SUMMARY

Hanoi, 2020

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

LE HOANG QUYNH

PHENOTYPE-DISEASE SEMANTIC RELATION EXTRACTION IN BIOMEDICAL TEXT

Major: Information Systems

Code: 9480104.01

DOCTORAL DISSERTATION SUMMARY

SUPERVISOR:
1. Prof. Dr. Nigel Collier
2. Dr. Dang Thanh Hai

Hanoi, 2020

Chapter 1

Introduction to Biomedical Relation Extraction

1.1 Problem statement

For the concept of Relation Extraction that we focus on, two text mining tasks are specifically relevant: Named Entity Recognition (NER) and Relation Classification.

The former, **named entity recognition** (NER, entity tagging), is an intermediate step for relation extraction. It refers to locating and classifying named entities in text into predefined categories. In other words, NER is the problem of finding entity mentions such as diseases, chemicals, genes, proteins, or organisms in natural language literature, then tagging them with their location and type.

The latter, **relation classification** (RC), go after NER to find the semantic relations between the corresponding entities. Biomedical relation classification often tries to classify the relationship between pairs of biomedical entities to relations such as drug-drug interaction, chemical-induced disease, bacteria live-in location, or tag them as 'none' if we can not find any relationship between them.

1.1.1 Biomedical named entity recognition

A named entity (NE) (also called entity mention) is a continuous sequence of words that designates some real world entity such as *'AIDS'*, *'Apple Inc.'* and *'Cambridge'*. The task of **Named Entity Recognition (NER)** seeks to locate NE from free-form text and classify them into a set of predefined categories/types such as person, organization, location, expressions of times, quantities, monetary values, percentages or 'none-of-the-

above'. In other words, NER is the problem of finding the mentions of entities in natural language text and labelling them with their location and type.

Named Entity Recognition in Biomedical Domains:

Biomedical named entities (biomedical NE) are phrases or combinations of phrases that denote important concepts in biomedicine. They can be chemicals, diseases, anatomies, pathways and genes/proteins, etc. that are named in biomedical literature, which has been growing at an unprecedented speed.

Named entity recognition's formal definition:

Named entity recognition is typically modeled as a sequence labeling problem, which try to assign labels to each elements of a sequence. It is defined formally as follows:

Given a sequence of input tokens $X = (x_1, \ldots, x_n)$, and a set of labels L, determine a sequence of labels $Y = (y_1, ..., y_n)$ such that $y_i \in L$ for $1 \leqslant i \leqslant n$.

We would like to assign a label y_i to each observation x_i. While one may apply standard classification to predict the label y_i based solely on x_i, in sequence labelling, it is assumed that the label y_i depends not only on its corresponding observation x_i but also possibly on other observations and other labels in the sequence. Typically this dependency is limited to observations and labels within a close neighbourhood of the current position i.

1.1.2 Biomedical relation classification

Relation classification (RC) typically follows NER in the relation extraction system.

Relation extraction is *'the task of discovering semantic connections between entities. In text, this usually amounts to examining pairs of entities in a document and determining (from local language cues) whether a relation exists between them.'*

We take the pairwise approach for the task of relation classification. I.e., after NER, we considered all pairs of recognized NERs as potential candidates, and give them as the input to the relation classification system. The relation classification system then classifies these candidates to assign them to a pre-defined relation type or 'None-of-above' (i.e., the negations).

Relation classification in Biomedical Domains:

Biomedical relation extraction concerns the detection of semantic relations between biomedical named entities or noun phrases. Recently, there has been considerable interest in relation extraction and relation classification with a variety of relationship. The common biomedical relations includes Drug-drug interaction, chemical-disease relation, Protein-protein interaction and many others. With a multitude of possible relation types,

it is critical to understand how systems will behave in a variety of settings. In biomedical domain, relation classification is useful in many fact extraction applications ranging from identifying adverse drug reactions to major life events. It is also important in tasks such as Question Answering and Knowledge Acquisition.

Relation classification's formal definition:

As treated as a classification problem, relation classifier can be defined as a real-valued function f_R that decides whether the corresponding entities are in a relation or not. Formally,

$$f_R(T(d, e_1, e_2)) = \begin{cases} +1 \; if \; e_1 \; and \; e_2 \; are \; related \; according \; to \; relation \; R; \\ -1 \; if \; otherwise \end{cases} \tag{1.1}$$

In which,

e_1 and e_2 are two entities that create a candidate for relation classification.

d is a document which includes corresponding entities e_1 and e_2. d can be a sentence, a paragraph or a document depending on the scope of relationships.

$T(d)$ is the information that is extracted from d.

1.2 Literature review

1.2.1 Literature review of biomedical named entity recognition

Many works on biomedical NER uses **statistical feature–based machine learning methods** which are often more robust in terms of system performance. ***Supervised machine learning*** (or learning from labelled data) utilizes large annotated corpus and the pre-defined feature set for inferring optimal prediction functions by training the model and then use it to predict the labels to new data. ***Conditional Random Fields (CRF)*** is the most popular discriminative machine learning model that alternative to the previous for sequence labelling. In addition to CRF, supervised machine learning methods that can be used for NER are extremely abundant with many variants, such as Hidden Markov Model (HMM), semi-markov model, Maximum Entropy Markov Model (MEMM), Support Vector Machine, decision tree, transition-based model, and more.

In the past few years, the advent of **deep neural networks** with the capability of automatically feature engineering even from noisy data has leveraged the development of NER models. A variety of deep learning methods and architecture have used in the field of NLP in general and biomedical NER in particular. In which, the most typical deep

neural networks (DNNs) are the Convolutional Neural Network (CNN), the Recurrent Neural Networks (RNNs) and their variants LSTM.

1.2.2 Literature review of biomedical relation extraction

Since the co-occurrence method often has low precision and rule/pattern-based methods are labour-intensive but not generalized, **statistical machine learning approaches** are currently one of the top choices for relation extraction. ***Supervised machine learning methods*** are data-driven, i.e., based on domain-specific manually annotated corpora. In the biomedical domain, these approaches are widely used since they can take advantages of various annotated biomedical corpora which are free availability but bring potential performance. In this dissertation, we only focus on the feature-based methods. The popular feature-based supervised machine learning algorithm is *Support Vector Machines* (SVM). Feature-based SVM was used for extracting chemical-induced disease relation, Live-in event, drug-drug interaction, protein-protein interaction, protein-organism-location relation and many other biomedical relations. In addition to SVM, machine learning methods that applied for biomedical relation extraction are abundant, such as Conditional Random Fields, Naive Bayes, maximum entropy, logistic regression.

Recent successes in **deep learning** have stimulated interest in applying neural architectures to the task of relation classification. They are extremely good at automatically feature engineering from noisy data, thus, not requiring a handcrafted feature set but still yielding good performances. ***Recurrent Neural Network*** and ***Convolutional Neural Networks*** (CNNs) are among early approaches to be applied successfully to biomedical relation classification problem and yields the state-of-the-art results.

1.3 Related resources

We use the BioCreative V Chemical-Disease relation (BC5 CDR) corpus as a baseline data for our experiments in this dissertation. In addition, many other datasets are used, depending on the purpose of the proposed model verification. The detailed information of all dataset are shown in Table 1.1 and Table 1.2.

Table 1.1: Information about the BC5 CDR, NCBI and FSU-PRGE corpora for NER.

Corpus	**Subset**	**Articles**	**Disease**		**Chemical**		**Gene/Protein**	
			Mentions	**Uniques**	***Mentions***	***Uniques***	***Mentions***	***Uniques***
BC5 CDR	Training	500	4182	1965	5203	1038		
	Development	500	4244	1865	5347	1012		
	Test	500	4424	1988	5385	1066		
NCBI Disease	Training	593	5145	1710				
	Development	100	787	368				
	Test	100	760	427				
FSU PRGE	Whole corpus	3309					59365	16683

Table 1.2: Information about the BC5 CDR, BB3, DDI and Phenebank corpora for relation classification.

#	Corpus	IAA	Size	Entity	Relation	% of negatives	Cross-sentence	Directed	Undirected	SDP length
1	BioCreative V CDR	-	1000 (500)	2	1	61.4 %	✓	✓	–	6.8 (24)
2	DDI-2013	D: 0.84 M: 0.62	730 (175)	4	4	85.3 %	–	–	✓	9.0 (66)
3	BB3	0.47	95 (51)	3	1	61.4 %	✓	✓	–	7.5 (25)
4	Phenebank	0.56	1000 (500)	9	5	77.0 %	✓	✓	✓	6.2 (26)

IAA: the Inter-annotator Agreement score; Size: training set size (test set size in the brackets) in terms of the number of documents; Entity: the number of entity types; Relation: the number of relation types; % of negative: the distribution of positive and negative instances; inter-sentence: if there are inter-sentence relations; Directed: if there are directed relations in the corpus; Undirected: if there are undirected relations in the corpus; SDP length: the averaged (max in brackets) length of the SDPs in the corpus.

Chapter 2

A Pipeline End-To-End Model For Biomedical Relation Extraction

In this Chapter, we describe an extension of UET-CAM system, which participated in the BioCreative V CDR track and was ranked 4^{th} among 18 participating systems by the track committee.

2.1 Distant learning with silverCID corpus

Distant-supervision learning, that takes advantages of both supervised and unsupervised learning, successfully applied in several researches on relation extraction. In this works, we tend to apply the distant-supervision learning for chemical-induced disease relation classification problem. It use a silver standard CID corpus (SilverCID) that constructed using the CTD database and PubMed according to five steps: Relation filtering, Collecting, Overlap removal, Annotating and Sentence filtering. This data set contains $38,332$ sentences, 1.25 million tokens, $48,856$ chemical entities ($1,196$ unique chemical entities), $44,744$ disease entities ($2,098$ unique disease entities) and $48,199$ CID relations ($12,776$ unique CID relations).

2.2 Proposed model

The overall architecture of our proposed system is described in Figure 2.1 based on the integration of several machine learning techniques to maximize their strengths and overcome the weaknesses. Pre-processing steps include sentence splitting, tokenization, abbreviation identification, stemming, POS tagging and dependency parsing (Stanford [1]).

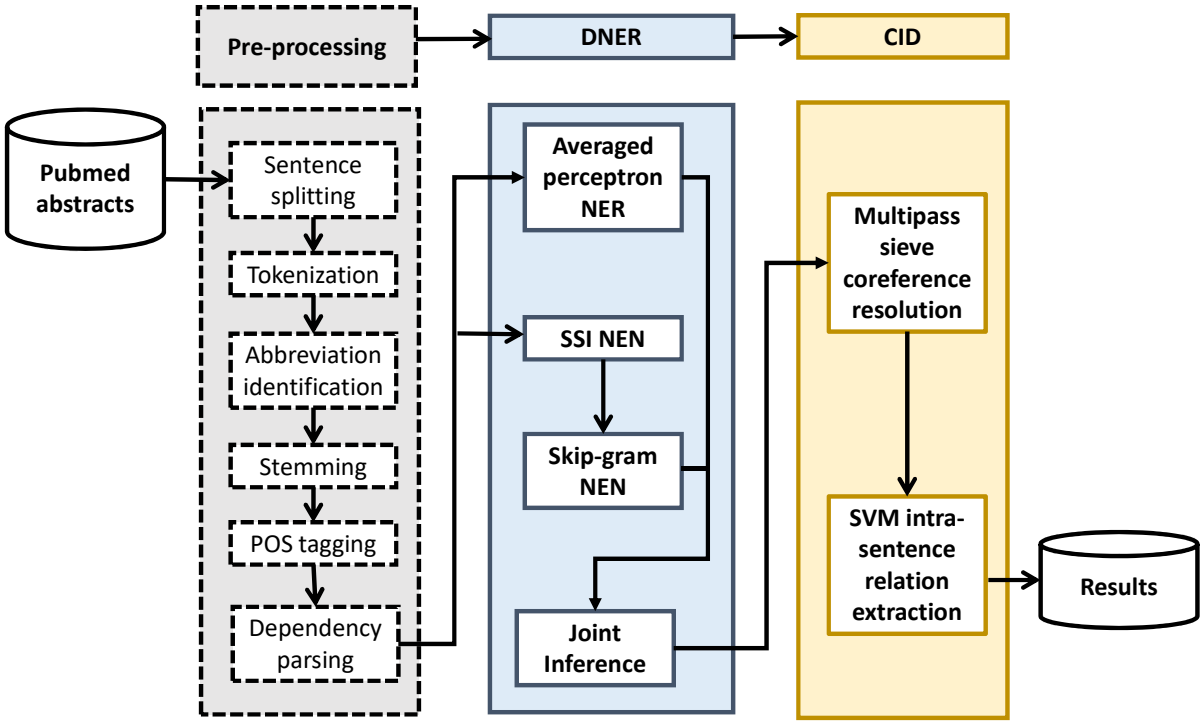

Figure 2.1: Architecture of the proposed UET-CAM system.
Boxes with dotted lines indicate pre-processing modules, which are done by available public tools.

2.2.1 Joint model of named entity recognition and normalization

Traditionally, NER and Named Entity Normalization (NEN) were treated as two separate tasks, in which, NEN took the output of NER as its input in a pipeline manner. Several studies have pointed out the limitations of this pipeline approach, i.e. causing cascading errors from NER to NEN, and limiting the ability of the NER system to exploit the lexical information provided back by the normalization directly.

– We employ a structured perceptron model for NER.

[1] Stanford Dependencies: `http://nlp.stanford.edu/software/stanford-dependencies.shtml`

– The NEN module is a combining model that consists of supervised and unsupervised word embedding methods for named entity normalization in biomedical text. In which, supervised semantic indexing is a supervised WE methods, and skip-grams is an unsupervised WE methods.

– Our DNER system was a joint decoding model, which used a modified beam search for decoding. In this model, we trained two separate models for NER and NEN and then decoded them simultaneously. We also proposed a new scoring function for Beam search decoding.

2.2.2 Coreference resolution

Our proposed system employed the coreference module that was based on a multi-pass sieve model.We first processed each abstract by noun phrase (NP) chunking and then created a set of NPs pairs for each abstract. These pairs of NPs were then passed through the sieves. Those that were not kept in each sieve were passed through the next sieve to the end. Any sieve kept pairs were considered as co-referent pairs, There were nine sieves used, each corresponding to a set of rules.

2.2.3 Support vector machine intra-sentence relation classification

Our work was based on knowing that if a noun phrase and an entity are co-referent, the noun phrase can be considered as an entity of that type. The intra-sentence relation extraction module received sentences that contain a disease - chemical pair as input and classified whether this pair had the CID relation or not. The intra-sentence relation extraction module was based on Support Vector Machine (SVM) – one of the most popular machine learning methods which have been successfully applied for biomedical relation extraction. We used the Liblinear[2] tool to train a supervision binary SVM classifier ($L2-$ regularized and $L1-$ loss) on the CDR track training/development data set and our SilverCID corpus.

[2] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`

2.3 Experimental results and discussion

2.3.1 Named entity recognition and normalization results

The experimental results of the DNER phase on the CDR track testing data set are shown in Table 2.1. Note that only disease entities were evaluated.

Table 2.1: Disease named entity recognition results on BC5 CDR corpus of UET-CAM system.

Method		**Precision** (%)	**Recall** (%)	**F1** (%)
BioCreative benchmarks*	Dictionary look-up	42.71	67.46	52.30
	DNorm	81.15	80.13	80.64
	Average result	78.99	74.81	76.03
	Rank no. 1 result	89.63	83.50	86.46
UET-CAM DNER		**73.20**	**79.98**	**76.44**
+ SilverCID corpus		**79.90**	**85.16**	**82.44**
NER-NEN pipeline		78.26	83.17	80.64

**Provided by the BioCreative 2015 organizer.*

2.3.2 CID relation classification results

Table 2.2 shows the results of our system on the CID task.

Table 2.2: Relation classification results on BC5 CDR corpus of UET-CAM system.

Method		**Precision** (%)	**Recall** (%)	**F1** (%)
BioCreative benchmarks*	Co-occurrence*	16.43	76.45	27.05
	Average result*	47.09	42.61	43.37
	Rank no. 1 result*	55.67	58.44	57.03
UET-CAM CID relation extraction*		**53.41**	**49.91**	**51.60**
+ silverCID corpus		**57.63**	**60.23**	**58.90**
SVM		44.73	50.56	47.47
SVM+ silverCID corpus		51.42	52.81	52.11
SVM+ CR EMC		47.64	50.28	48.93

Results provided by the BioCreative 2015 organizer.

**UET-CAM system includes SVM+ CR + MPS; SVM: SVM intra-sentence relation extraction.*

CR: Coreference resolution; MPS: Multi-pass sieve; EMC: expectation maximization clustering.

2.4 Conclusion

In this Chapter, we have presented a systematic extended study of our approach to the BioCreative V Chemical-Disease relation task. Our system, namely UET-CAM, is a modular system that handles the DNER (named entity recognition) and CID (relation classification) task separately. DNER is a joint decoding model for NER and NEN based on several state-of-the-art machine learning methods. For CID relation classification, we build an SVM-based model with a rich feature set and then improve it by using distant learning with silverCID corpus and crucially, applying a multi-pass sieve coreference resolution module. Our best performance achieved an F1 of 81.93 for DNER while that of the DNorm, the state-of-the-art DNER system based on SSI was 80.64%. The best performance for CID of our improved system had F1 of 58.90%, comparable to that of the highest ranked system in the CID task with 57.03%.

Chapter 3

Applying Deep Learning Models To Biomedical Named Entity Recognition

This chapter improves NER problem by applying deep learning method instead of traditional feature-based machine learning methods.

3.1 Introduction to deep learning

Statistical machine learning methods that require feature engineering require us to carefully exploit the characteristics of the data to propose useful features. Recently, deep neural networks (DNNs) have been effectively used to learn robust syntactic and semantic representations behind complex structures and increasingly been used for various NLP related tasks. With a deep learning model, relevant features are automatically extracted from data. They perform the 'end-to-end learning' – where the neural networks are given raw data and a task to perform, such as classification, and it learns how to do this automatically. They are extremely good at automatically feature engineering from noisy data, thus, not requiring a handcrafted feature set but still yielding good performances.

3.2 Proposed model

D3NER comprises of four layers, namely TPAC embeddings, context representing bi-LSTM, project and NER layer, being structured in an architect as depicted in Figure 3.1. After TPAC embedding layer, the embedded input vectors are fed into the bi-LSTM layer for modeling context information of each word. A project layer then encodes the output of bi-LSTM layer into a sequence of d-dimensional vectors (with d is the number of labels defined in tagging scheme). The on top layer CRF is used to labels NER tag for the whole sentence.

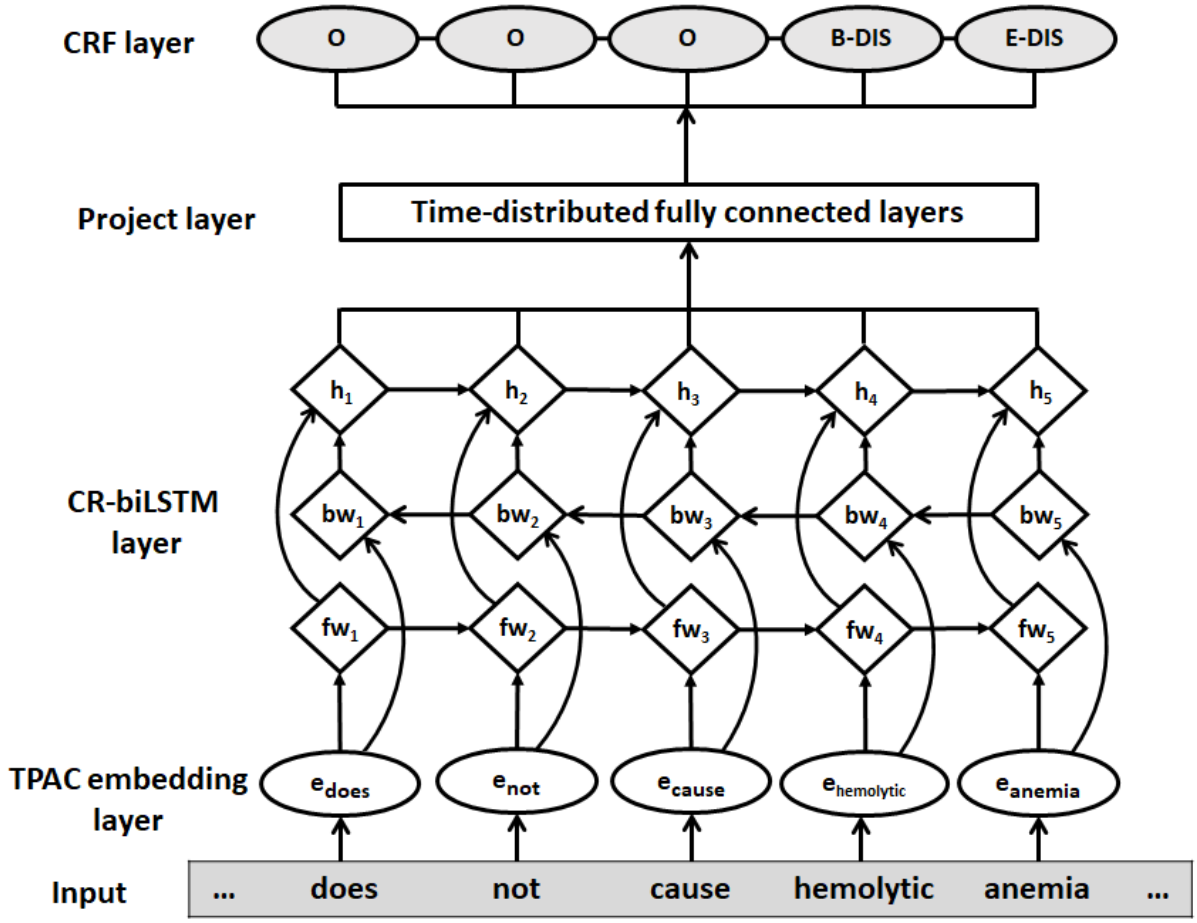

Figure 3.1: The D3NER architecture.

Example comes from the BioCreative V Chemical Disease Relation task corpus (PMC3425586).

To avoid overfitting, we apply dropout, with 0.5 and 0.15 respectively for the final hidden layer of CE-bi-LSTM and CR-bi-LSTM, and 0.5 for the first fully connected layer of the project layer. Early stopping is applied based on the D3NER performance on the validation sets (the model often stops at around 27, and 13 epochs on the BC5 CDR and on the NCBI Disease corpus, respectively).

3.3 Experimental results and discussion

We evaluate D3NER on three benchmark biomedical corpus the BioCreative V Chemical Disease Relation (BC5 CDR) corpus, the NCBI Disease corpus and FSU-PRGE (The FSU PRotein GEne) corpus.

The experimental results are shown on Table 3.1 and Table 3.2.

Table 3.1: Performance of D3NER and compared state-of-the-art models on two benchmark corpora for Disease and Chemical NER.

Model	**BC5 CDR Chemical**			**BC5 CDR Disease**			**NCBI Disease**		
	P	R	F1	P	R	F1	P	R	F1
Dnorm	-	-	-	82.00	79.50	80.70	82.20	77.50	79.80
tmChem	93.20	84.00	88.40	-	-	-	-	-	-
TaggerOne *	92.40	84.70	88.40	83.10	76.40	79.60	83.50	79.60	81.50
Habibi et al.,	92.18	89.94	91.05	84.19	82.79	83.49	86.43	82.92	**84.64**
Wei et al.,	-	-	-	85.28	83.30	84.28	-	-	-
Att-ChemdNER	93.49	91.68	92.57	-	-	-	-	-	-
Our model D3NER	93.73	**92.56**	**93.14**	83.98	**85.40**	84.68	85.03	**83.80**	84.41
TaggerOne **	**94.20**	88.80	91.40	85.20	80.20	82.60	85.10	80.80	82.90
Transition-based	-	-	-	**89.61**	83.09	**86.23**	**90.72**	74.89	82.05

Results are reported in %.

The highest values for each metric of each entity type are highlighted in bold.

**TaggerOne NER only, **TaggerOne joint model.*

Table 3.2: Performance of D3NER and compared state-of-the-art model on FSU-PRGE corpus for Gene/protein NER.

Model	**Precision**	**Recall**	**F1**
Habibi et al.,	**87.26**%	87.24%	87.25%
Our model D3NER	87.09%	**88.17**%	**87.62**%

3.4 Conclusion

This chapter presents D3NER, a novel biomedical named entity recognition using conditional random fields and bidirectional long short-term memory improved with jointly fine-tuned embeddings of various linguistic information. We evaluate D3NER on three benchmark datasets, i.e. the BC5 CDR corpus, the NCBI Disease corpus and the FSU-PRGE corpus which have also been used for performance evaluation in 7 very recent

state-of-the-art related models to which D3NER is compared. Experimental results demonstrate the power of D3NER in recognition of all disease, chemical and gene/protein named entities. D3NER could yield excellent performance for chemical NER and very good for both disease and gene/protein NER in terms of three popular performance scoring metrics.

Chapter 4

Applying Deep Learning Models to Biomedical Relation Classification

Following the success in applying deep learning to the NER problem in Chapter 3, this chapter continues with the use of these advanced neural networks for the relation classification problem.

4.1 A large-scale deep learning model for biomedical relation extraction

In this section, we present a large-scale deep neural model of state-of-the-art neural network architectures on four biomedical benchmark datasets, which represent a variety of language characteristics and semantic types.

4.1.1 Proposed model

Our 'Man for All SeasonS' (MASS) model comprises an embeddings layer, multi-channel bi-directional Long Short-Term Memory (BLSTM) layers, two parallel Convolutional Neural Network (CNN) layers and three $\mathrm{softmax}$ classifiers. The MASS model's architecture is depicted in Figure 4.1. MASS makes use of words and dependencies along the SDP going from the first entity to the second one using both forward and backward sequences. As is standard practice, an entity pair is classified as having a

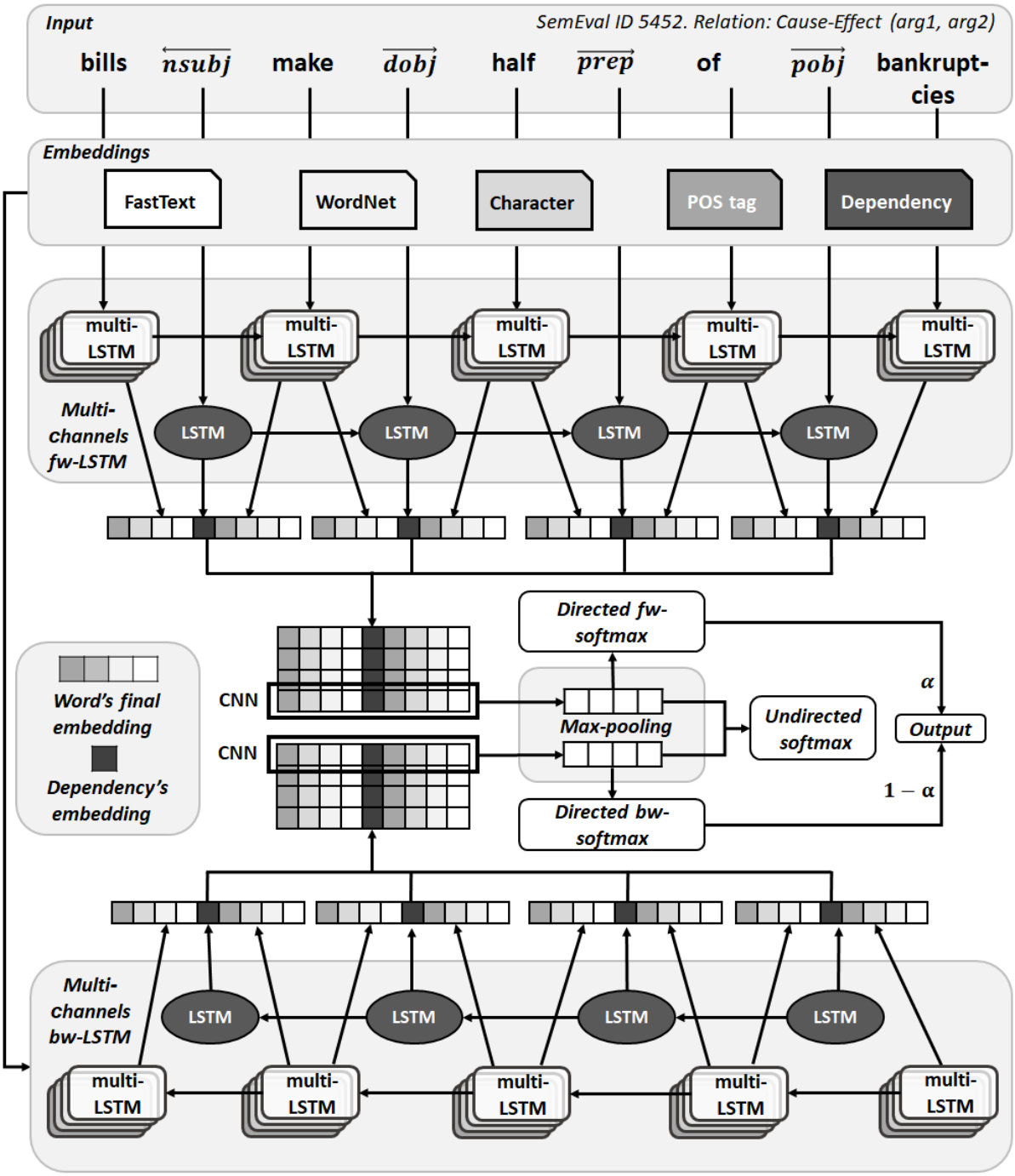

Figure 4.1: The architecture of MASS model for relation classification.
An embeddings layer is followed by multi-channel bi-directional LSTM layers, two parallel CNNs and three softmax classifiers. The model's input makes use of words and dependencies along the SDP going from the first entity to the second one using both forwards and backwards sequences.

relationship if and only if the SDP between them is classified as having that relation.

4.1.2 Experimental Results

Our experiments used four well-known benchmark corpora from different biomedical sub-domains, which have been used to evaluate various state-of-the-art relation classification systems: the ***DDI-2013*** corpus, the ***BC5 CDR*** corpus, the ***BB3*** corpus, and the ***Phenebank*** corpus.

In all corpora, the MASS model's results are always better than the baseline mode and very comparative to other methods.

4.2 An attentive augmented deep learning model for biomedical relation extraction

4.2.1 Richer-but-Smarter Shortest Dependency Path

The simple structure of the Shortest Dependency Path (SDP) is one of its weaknesses since there exists some useful information in the dependency tree that does not appear in the SDP. We notice that the child nodes attached to the shortest dependency paths and their dependency relation from their parents can provide supplemental information for relation classification. Depending on a specific set of relations, it turns out that not all children are useful to enhance the parent node; we select relevant children by applying several attention mechanisms with kernel filters. This new representation of relation is named Richer-but-Smarter SDP (RbSP). In this RbSP structure each token t is represented by itself and its attached children on the dependency tree.

4.2.2 Proposed Model

The overall architecture of our proposed model is shown in Figure 4.2. Given a sentence and its dependency tree, we build our model on the SDP between two nominals and its directed children on the tree. Here, we mainly focus on the SDP representation, which is composed of dependency embeddings, token embeddings, and token's augmented information. After SDP representation phase, each token and dependency relation is transformed into a vector. This sequence of vectors is then fed to a convolutional neural network to capture the convolved features that can be used to determine which relation two nominals are of.

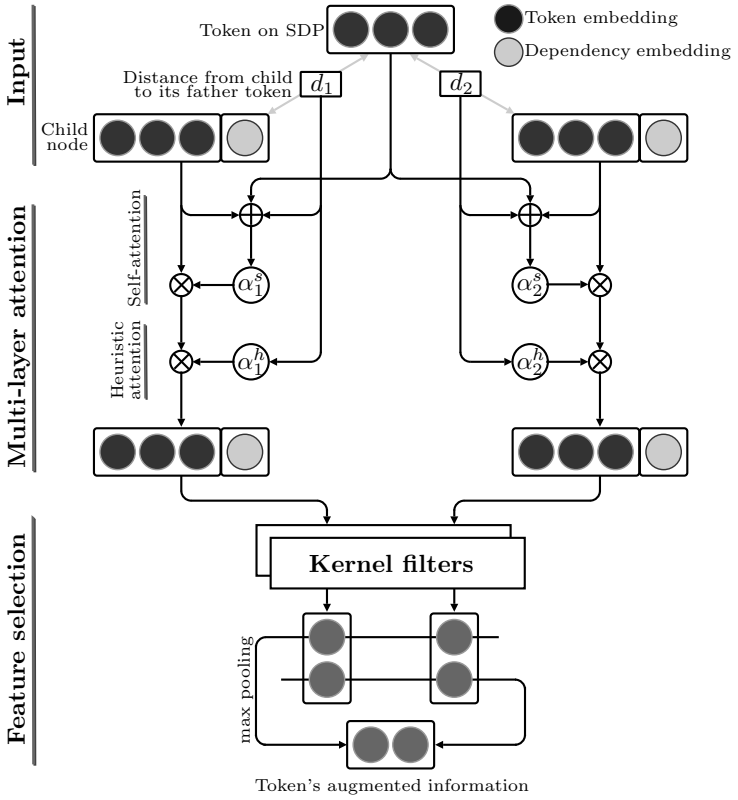

Figure 4.2: The architecture of RbSP model for relation classification.

A CNN model is applied to the output of the SDP representation. Our proposed model takes the Augmented SDP between two nominals that includes dependencies, tokens and their children as input.

4.2.3 Experimental results

The RbSP model's performance and comparisons

Table 4.1 summarizes the performance of our model and some comparative models.

4.3 A multi-fragment ensemble deep learning model for biomedical relation extraction

4.3.1 Bagging with bootstrap training data

Overfitting is one of the most remarkable problems of deep learning models. In the case of overfitting, the deep model often has low bias accurate predictions for the training data, i.e., it fits well with the training data, and training errors are low. The main hypothesis of ensemble methods is that if weak models are correctly combined, we can obtain more accurate and/or robust models. Ensemble methods is constructing multiple (same or different) models (often called 'weak learners' or 'base learners') and then classify

Table 4.1: The comparison of RbSP model with other comparative models on BC5 CDR corpus.

Model	Feature set	P	R	F1
BioCreative benchmarks*	Co-occurrence	16.43	76.45	27.05
	Average result	47.09	42.61	43.37
	Rank no.1 result	55.67	58.44	57.03
UET-CAM	SVM, rich feature set	53.41	49.91	51.60
	+ silverCID corpus	57.63	60.23	58.90
MASS	SDP, LSTM, CNN, WordNet	58.90	54.90	56.90
	+ Ensemble	56.80	57.90	57.30
	+ Post processing	52.80	71.10	60.60
ASM	Dependency graph	49.00	67.40	56.80
hybridDNN	Syntactic feature, word embeddings	62.15	47.28	53.70
	+ Context	62.39	47.47	53.92
	+ Position	62.86	47.47	54.09
ME+CNN	Contextual of whole sentence	59.70	57.50	57.20
	+ Cross-sentence	60.90	59.50	60.20
	+ Post processing	55.70	68.10	61.30
BRAN	Position, multi-head attention	55.60	70.80	62.10
	+ Data	63.30	67.10	**65.10**
	+ Ensemble	**64.00**	69.20	**66.20**
Baseline	Word embeddings	60.25	49.37	54.27
	+ Dependency Unit	60.33	50.36	54.90
cduCNN (our model)	Compositional Embedding, Dependency Unit	57.24	55.27	56.24
	+ Normalize conjunction	56.95	56.14	56.54
	+ Normalize object of a preposition	56.66	55.94	56.30
RbSP (our model)	**cduCNN** + Augmented Information	57.68	57.27	57.48
	+ Ensemble	58.78	57.20	57.98
	+ Post processing	52.38	**72.65**	**60.78**

* *Results are provided by the BioCreative V organizer.*

new data by taking a weighted or vote of their predictions. Ensemble models also often yield top ranking in many machine learning shared tasks.

4.3.2 Proposed models

The multi-fragment ensemble architecture is illustrated in Figure 4.3

In this work, we use the RbSP model proposed in Section 4.2 as the base model. Due to our hardware capabilities, we heuristically choose 100 base models for building the ensemble for BioCreative V dataset. In this work, we propose a multi-fragment

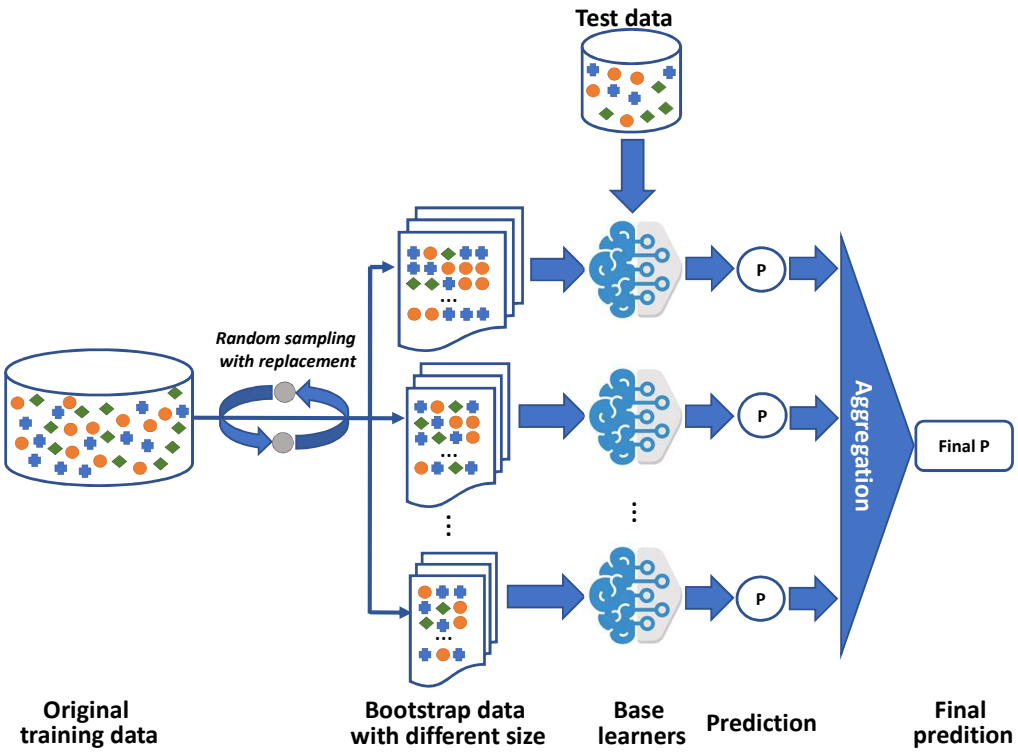

Figure 4.3: The multi-fragment ensemble architecture.

ensemble model, i.e., base models are trained on different sizes of bootstrap data, from 50% to 100%. For each instance, the result predicted by each model is considered as a vote. We apply the majority vote with threshold-moving technique to improve the performance.

4.3.3 Experimental results

Multi-fragment ensemble model results: Table 4.2 shows the experimental results on BioCreative V CDR dataset.

4.4 Conclusion

In this chapter, we proposed several novel deep neural architectures to overcome the limitation of traditional feature-based machine learning method as well as improve the performance.

In Section 4.1, we present ***MASS*** - a large-scale and well-balanced relation classification model that consists of several deep learning components applied on the Dependency Unit of Shortest Dependency Path. Experiments have proved the superiority and high adaptability of the model when applied to diverse data domains with a variety of lan-

Table 4.2: Multi-fragment ensemble results on BC5 CDR corpus.

		With replacement[†]			**Without replacement**[‡]		
		P	R	F1	P	R	F1
Averaged result		*57.79*	*56.42*	*57.10*	*57.68*	*57.77*	*57.73*
Size of bootstrap data[*]	10	57.38	55.42	56.39	58.28	54.30	56.22
	20	58.84	56.17	57.47	59.30	56.17	57.69
	30	58.33	56.92	57.62	58.69	56.73	57.70
	40	58.89	56.55	57.69	58.29	57.49	57.89
	50	58.63	57.77	58.20	59.00	57.58	**58.28**
	60	59.27	57.30	58.27	58.80	57.11	57.94
	70	59.68	57.86	**58.76**	58.75	57.39	58.06
	80	59.51	57.58	58.53	58.85	57.49	58.16
	90	59.81	57.49	58.62	59.21	57.02	58.09
	100	59.25	56.45	57.82	58.78	57.20	57.98
Multi-fragment bootstrap[+]	mf-50	61.25	56.83	58.96	60.56	56.64	58.54
	mf-60	60.73	57.30	**58.96**	60.49	57.02	**58.70**
	mf-70	60.57	57.20	58.84	60.36	56.45	58.34
	mf-80	60.31	57.02	58.62	59.76	56.83	58.26
	mf-90	60.31	57.02	58.62	59.68	56.64	58.12
+ *Post processing*	*mf-60*	*53.89*	*73.81*	*62.30*			

Results are reported in %.

100 *base models are used.*

[†], [‡]*Bootstrap data sets were built with or without replacement.*

[*]*Size of bootstrap data comparing to original size of training data, run from* 10 *to* 100.

[+]*Multi-fragment bootstrap 'mf-n' means using several bootstrap sizes from n to* 100.

guage characteristics and semantic types.

In Section 4.2, we propose a novel attentive augmented deep learning model ***RbSP*** that overcomes the disadvantages of traditional shortest dependency path and improves the attention mechanism with kernel filters to capture the features from context vectors.

In Section 4.3, we propose an effective ensemble mechanism, namely ***multi-fragment ensemble***, to avoid overfitting and improve the performance and stability for deep learning models.

Chapter 5

Inter-sentence Relation Classification in Biomedical Text

This chapter proposes a novel graph-based representation and a deep learning architecture to extract inter-sentence relations.

5.1 Background

Most previous relation extraction (RE) studies focused on intra-sentence relations and ignored inter-sentence relations, which explores entities at the document level rather than that at the specific mentions. An inter-sentence relation often explores entities at the document level rather than that at the specific mentions.

5.2 Materials and Methods

Figure 5.1 illustrates our proposed model for extracting the semantic relation at the abstract level, which contain four main phrases: (i) Firstly, we construct a document sub-graph to represent the relationship between entity pairs. We use the dependency information from the dependency tree and some non-local dependency information, includes: adjacent sentence information, title information, coreference information and knowledge-base information. (ii) In order to represent an instance by a set of paths, we apply several advanced techniques for finding, merging and choosing the relevant paths between entity pairs. (iii) In the next step, the advanced attention mechanism and several linguistic information are applied to explorer the information from the document sub-graphs more effectively. (iv) Lastly, to exploit these enriched representations effectively,

we develop a shared weight Convolutional Neural Network model (CNN).

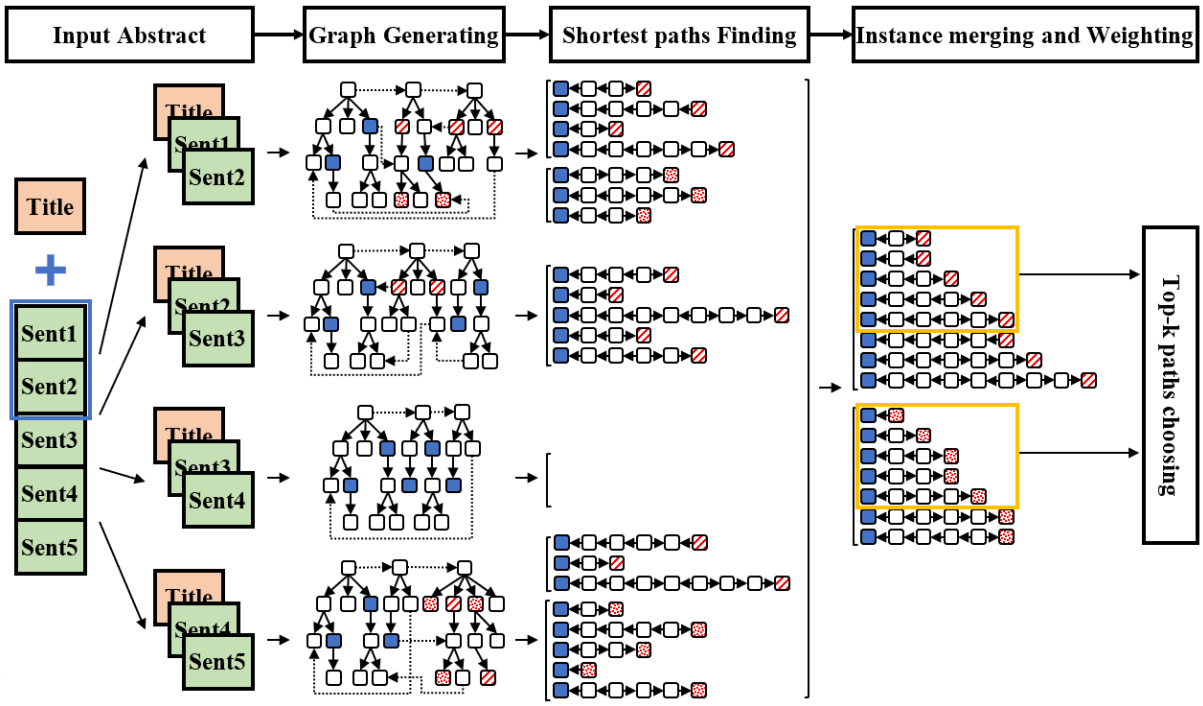

Figure 5.1: Proposed model for inter-sentence relation classification.

Figure 5.2 illustrates the overall architecture of our swCNN model, which is comprised of two main components: multi-path representation and classification. Given a set of multiple k paths as input, each path is converted into a separated embedding matrix. A shared-weight convolution with relu activation layer is followed to capture convolved features from a from these embedding matrices simultaneously. The essential features are gathered using a filter-wise pooling layer before classified by a fully connected layer with softmax classification.

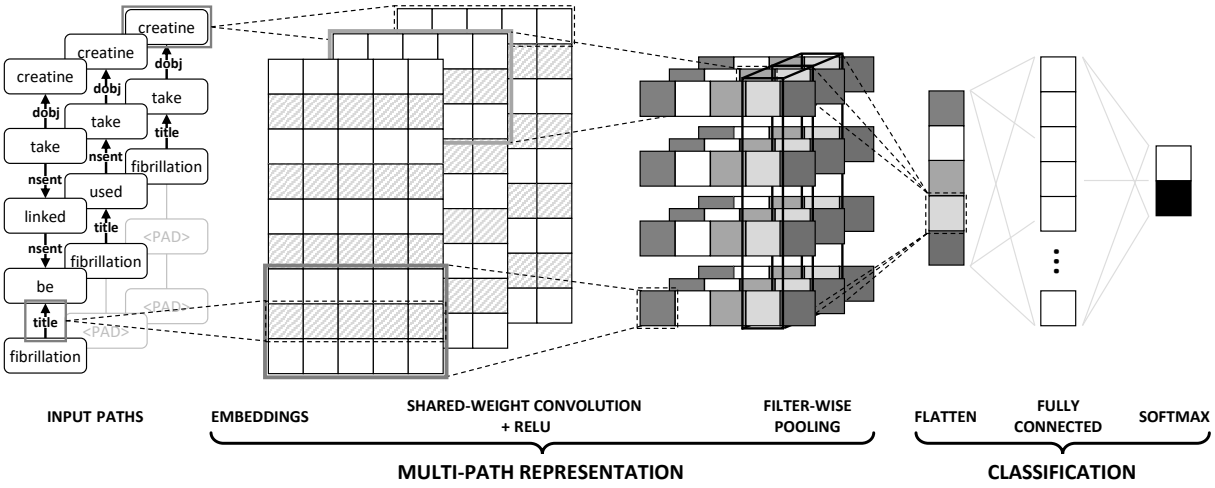

Figure 5.2: Diagram illustrating of a swCNN architecture.

5.3 Results

Table 5.1 summarizes the performance of our model and some comparative models. In which, the results of comparative models are reported both with and without using any additional enhancements.

Table 5.1 summarizes the performance of our model and some comparative models. In which, the results of comparative models are reported both with and without using any additional enhancements.

Table 5.1: The performance of document sub-graph-based model and some comparative models.

Method/model		Precision	Recall	F1
UET-CAM	Orginal	53.41	49.41	51.60
SVM + multi pass sieve coreference)	+ Distant supervision learning	57.63	60.23	58.90
SVM + rich feature set	Original	64.24	52.06	57.51
	+ Distant supervision learning	**65.59**	56.94	61.01
CNN+ME	Orginal	60.90	59.50	60.20
	+ Post-processing	55.70	68.10	61.30
LSTM-CNN on sequence of sentences	Orginal	24.00	52.00	32.80
	+ Entity replacing	54.30	65.90	59.50
BRAN (CNN + Attention on whole abstract)	Original	55.60	70.80	62.10
	+ Data	64.00	69.20	66.20
	+ Ensemble	65.40	**71.80**	**68.40**
Document-level Graph CNN	Original	52.80	66.00	58.60
Graph-based results	Original	57.93	68.17	62.59
	+ Distant supervision learning	62.48	67.17	64.74
	+ Ensemble	62.59	68.35	65.34

Results are reported in $\%$.

Highest result in each column is highlighted in bold.

Our model yields very competitive results when compared to other state-of-the-art models that have taken into account the inter-sentence relationships. Compare the original model without any additional enhancements, our model gives the best results with 62.59%.

5.4 Conclusions

In this chapter, we present a novel representation for a sequence of adjacent sentences in a document (namely document sub-graph). The graph is constructed using various types of information to capture local and non-local features. We also propose an instance merging mechanism and using a set of multiple paths for representing the relationship between entity pair. To explore the information in the document sub-graph, we construct a deep neural architecture based on a shared-weight convolutional neural network.

In experiments on BioCreative V CDR corpus, without using any external knowledge resources and additional enhancements, our proposed model outperforms all comparative models. Comparing the full model performance, our model still achieves comparable results, only lower than BRAN.

Conclusion

The dissertation presented a systematic study on biomedical relation extraction, a fundamental problem in the field of bioNLP. Relation extraction consists of two sub-problems: named entity recognition and relation classification, each of them was resolved as a separate problem. The contribution of the dissertation can be concluded as follow:

–The dissertation presented a detailed survey on the biomedical relation extraction problem.

– We proposed a novel representation for inter-sentence relation called document subgraph together with several additional techniques such as instance merging, using multiple paths to represent a single data instance. In addition, the dissertation contributed to the research community by creating a silver-standard dataset called *'silverCID'* for distant learning.

– The dissertation proposed several novel machine learning architectures that consist of both traditional feature-based and deep learning techniques. All proposed model had the potential and comparable results to the very state-of-the-art researches.

From the results achieved in the dissertation as well as the remaining limitations, there is some research direction for future works: (i) Try to apply advanced techniques more effectively, especially attention mechanism and ensemble manner. (ii) Continue research work with inter-sentence relation classification. (iii) To be able to apply to realistic systems, we should pay more attention to the processing speed and the use of hardware resources. Some components that have not been effective can be removed. (iv) The $n-$ary relations is an interesting research problem, which may be focused after we complete our work on inter-sentence relation extraction.

List of Publications

[LHQ1] **Hoang-Quynh Le**, Mai-Vu Tran, Thanh Hai Dang, Quang-Thuy Ha and Nigel Collier (2016). "Sieve-based coreference resolution enhances semi-supervised learning model for chemical-induced disease relation extraction". Database (2016), Vol. 2016: article ID baw102.

[LHQ2] **Hoang-Quynh Le**, Duy-Cat Can, Thanh Hai Dang, Mai-Vu Tran, Quang-Thuy Ha and Nigel Collier (2017). "Improving chemical-induced disease relation extraction with learned features based on convolutional neural network". In proceedings of the 9th International Conference on Knowledge and Systems Engineering (KSE), pp. 292-297. IEEE.

[LHQ3] Thanh Hai Dang, **Hoang-Quynh Le**, Trang M. Nguyen, Sinh T. Vu (2018). "D3NER: Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information". Bioinformatics, 34(20), pp 3539-3546. *(The first two authors should be regarded as Joint First Authors).*

[LHQ4] **Hoang-Quynh Le**, Duy-Cat Can, Sinh T. Vu, Thanh Hai Dang, Mohammad Taher Pilehvar and Nigel Collier (2018). "Large-scale Exploration of Neural Relation Classification Architectures". In proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2266-2277.

[LHQ5] Duy-Cat Can, **Hoang-Quynh Le** and Quang-Thuy Ha (2019). "Improving Semantic Relation Extraction System with Compositional Dependency Unit on Diverse Shortest Dependency Path". In proceedings of the 11th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2019), pp. 140-152. Springer, Cham.

[LHQ6] Duy-Cat Can, **Hoang-Quynh Le***, Quang-Thuy Ha and Nigel Collier. "A Richer-but-Smarter Shortest Dependency Path with Attentive Augmentation for Relation Extraction". In proceedings of the 2019 Conference of the North American Chapter

of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2902-2912. *(*Corresponding author).*

This list contains 6 publications.