Large-scale Exploration of Neural Relation Classification Architectures

Hoang-Quynh Le[1], Duy-Cat Can[1†], Sinh T. Vu[1†], Thanh Hai Dang[1*], Mohammad Taher Pilehvar[2] and Nigel Collier[2]

[1]Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi, Vietnam
[2]Department of Theoretical and Applied Linguistics, University of Cambridge, UK
{lhquynh, catcd, sinhvt, hai.dang}@vnu.edu.vn
{mp792, nhc30}@cam.ac.uk

Abstract

Experimental performance on the task of relation classification has generally improved using deep neural network architectures. One major drawback of reported studies is that individual models have been evaluated on a very narrow range of datasets, raising questions about the adaptability of the architectures, while making comparisons between approaches difficult. In this work, we present a systematic large-scale analysis of neural relation classification architectures on six benchmark datasets with widely varying characteristics. We propose a novel multi-channel LSTM model combined with a CNN that takes advantage of all currently popular linguistic and architectural features. Our 'Man for All Seasons' approach achieves state-of-the-art performance on two datasets. More importantly, in our view, the model allowed us to obtain direct insights into the continued challenges faced by neural language models on this task. Example data and source code are available at: https://github.com/aidantee/MASS.

1 Introduction

Determining the semantic relation between pairs of named entity mentions, *i.e.* relation classification, is useful in many fact extraction applications, ranging from identifying adverse drug reactions (Gurulingappa et al., 2012; Dandala et al., 2017), extracting drug abuse events (Jenhani et al., 2016), improving the access to scientific literature (Gábor et al., 2018), question answering (Lukovnikov et al., 2017; Das et al., 2017) to major life events extraction (Li et al., 2014; Cavalin et al., 2016). With a multitude of possible relation types, it is critical to understand how systems will behave in a variety of settings (see Table 1 for an example).

(i)	*<e1>**Three-dimensional digital subtraction angiographic**</e1> (<e2>**3D-DSA**</e2>) images from diagnostic cerebral angiography were obtained ...*
(ii)	*The metal <e1>**ball**</e1> makes a ding ding ding <e2>**noise**</e2> when it swings back and hits the metal body of the lamp.*

Table 1: Examples for different relation types: sentence (i) shows a *Synonym-of* relation, represented by an abbreviation pattern, which is very different from the predicate relation *Cause-effect* in (ii).

To the best of our knowledge, almost all relation classification models introduced so far have been experimentally validated on only a few datasets - often only one. This is despite the availability of established benchmarks. The lack of transparency as well as the possibility of having selection bias raise a question about the true capability of state-of-the-art methods for relation classification. In addition, despite such a wealth of studies, it still remains unclear which approach is superior and which factors set the limits on performance. For example, heuristic post-processing rules have been seen to significantly boost relation classification performance on several benchmarks; yet, they cannot be relied upon to generalize across domains. The novel approach we present in this paper draws inspiration from neural hybrid models such as that of Cai et al. (2016). In this work, we present a large-scale analysis of state-of-the-art neural network architectures on six benchmark datasets which represent a variety of language domains and semantic types. As a means of comparison against reported system performance, we propose a novel multi-channel long short term memory (Hochreiter and Schmidhuber, 1997, LSTM) model combined with a Convolutional Neural Network (Kim, 2014, CNN) that takes advantage of all major linguistic and architectural features cur-

[†]Contributed equally & Names are in alphabetical order
[*]Corresponding author

rently employed. We designate this as a 'Man for All SeasonS' (MASS) model because it incorporates many popular elements reported by state of the art systems on individual datasets.

The main contributions of the paper are:

1. We presented a deep neural network model, in which each component is capable of taking advantage of a particular type of major linguistic or architectural feature. The model is robust and adaptable across different relation types in various domains without any architectural changes.

2. We investigated the impact of different components and features on the final performance, therefore, providing insights on which model components and features are useful for future research.

2 Related Works

We focus here on supervised approaches to relation classification. Alternatives include hand built patterns (Aone and Ramos-Santacruz, 2000), unsupervised approaches (Yan et al., 2009) and distantly supervised approaches (Mintz et al., 2009). Traditional supervised and kernel-based approaches have made use of a full range of linguistic features (Miwa et al., 2010) such as orthography, character n-grams, chunking as well as vertex and edge walks over the dependency graph. Hand crafting and modeling with such complex feature sets remains a challenge although performance tends to increase with the amount of syntactic information (Bunescu and Mooney, 2005).

Recent successes in deep learning have stimulated interest in applying neural architectures to the task. Convolutional Neural Networks (CNNs) (Nguyen and Grishman, 2015) were among early approaches to be applied. Following in this direction, (Lee et al., 2017) achieved state of the art performance on the ScienceIE task of SemEval 2017. Other recent variations of CNN architectures include a CNN with an attention mechanism in Shen and Huang (2016) and a CNN combined with maximum entropy in Gu et al. (2017). Various auxiliary information has been reported to improve the performance of CNNs, such as the document graph (Verga et al., 2018) and position embeddings (Shen and Huang, 2016; Lee et al., 2017; Verga et al., 2018). Recurrent Neural Networks (RNNs) are another approach to capturing

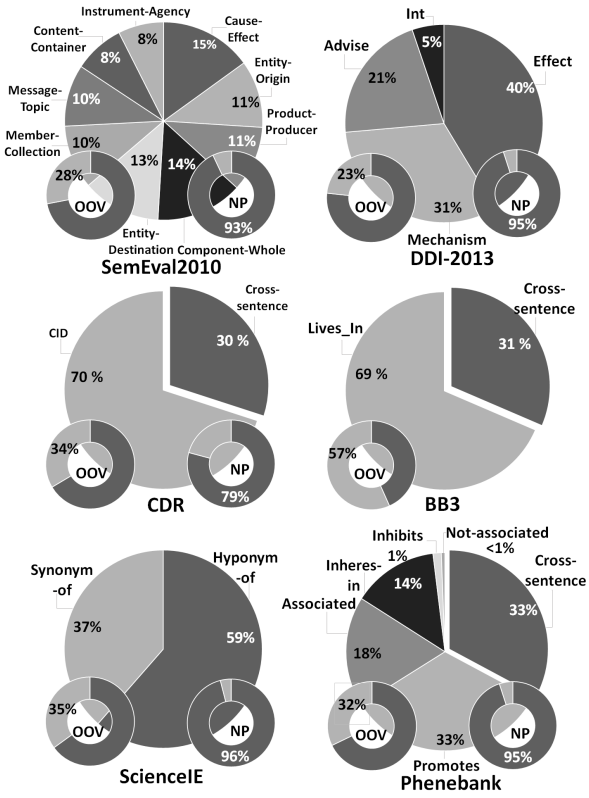

Figure 1: The statistics of corpora used in our experiments. Three aspects are considered: the distribution of relation types, the distribution of Out-Of-Vocabulary (OOV) in the test set and the distribution of new entity pairs (NP) that appeared in the test set but never appeared in the training data.

relations and naturally good at modeling long distance relations within sequential language data. Approaches include Mehryary et al. (2016) with the original RNN and Li et al. (2017); Ammar et al. (2017); Zhou et al. (2018) with RNNs having LSTM units which are used to extend the range of context. Apart from sentences themselves, RNN-based models often take as input information extracted from dependency trees, such as shortest dependency paths (SDP) (Mehryary et al., 2016; Ammar et al., 2017), or even whole trees (Li et al., 2017). Since RNNs and CNNs each have their own distinct advantages, a few models have combined both in a single neural architecture (Cai et al., 2016; Zhang et al., 2018).

3 Materials and Methods

3.1 Gold Standard Corpora

As noted above, our experiments used six well-known benchmark corpora from different domains, which have been used to evaluate vari-

#	Corpus	Domain	IAA	Size	Entity	Relation	% of negatives	Cross-sentence	Directed	Undirected	SDP length
1	SemEval *(SemEval 2010 - Task 8)*	Generic	0.74	8000 (2717)	–	9	17.4 %	–	✓	–	3.8 (13)
2	DDI-2013 *(SemEval DDIExtraction 2013)*	Biomedical	D: 0.84 M: 0.62	730 (175)	4	4	85.3 %	–	–	✓	9.0 (66)
3	CDR *(BioCreative5 CDR 2015)*	Biomedical	-	1000 (500)	2	1	61.4 %	✓	✓	–	6.8 (24)
4	BB3 *(BioNLPST BB-Event 2016)*	Biomedical	0.47	95 (51)	3	1	61.4 %	✓	✓	–	7.5 (25)
5	ScienceIE *(SemEval ScienceIE 2017)*	Scientific	0.45-0.85	400 (100)	3	2	88.5 %	–	✓	✓	6.5 (22)
6	Phenebank	Biomedical	0.56	1000 (500)	9	5	77.0 %	✓	✓	✓	6.2 (26)

Table 2: Characteristics of the six corpora used in this study. Domain: the domain of the corpus; IAA: the Inter-annotator Agreement score; Size: training set size (test set size in the brackets) in terms of the number of sentences (SemEval) or documents (all other corpora); Entity: the number of entity types; Relation: the number of relation types; % of negative: the distribution of positive and negative instances; Cross-sentence: if there are cross-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.

ous state-of-the-art relation classification systems. ***SemEval*** is a generic domain benchmark dataset (Hendrickx et al., 2009). The next four chosen corpora are from various biomedical domains: the ***DDI-2013*** corpus (Herrero-Zazo et al., 2013; Segura-Bedmar et al., 2014), the ***CDR*** corpus (Li et al., 2016), the ***BB3*** corpus (Delėger et al., 2016), and the ***Phenebank*** corpus. Finally, ***ScienceIE*** corpus contains scientific journal articles from three sub-domains (Augenstein et al., 2017). Inter-annotator agreement (IAA) as measured with Cohens kappa on these corpora indicates high variability in the range of $[0.45, 0.74]$, *i.e. moderate* to *substantial* agreement (McHugh, 2012).

As shown in Table 2, each of these corpora is distinct in many respects. CDR and BB3 were only annotated with one relation type, whilst other corpora have several relation types. In all corpora except SemEval, negative instances must be automatically generated by pairing all the entities appearing in the same sentences that have not been annotated as positives. As there are a large number of such entities, the number of possible negatives accounts for a large percentage of set of instances, *i.e.* up to 80% of the total in DDI-2013, ScienceIE and Phenebank. Further, the small percentage of positive examples includes several types, causing a severe imbalance in the data (He and Garcia, 2009) (see Figure 1 for further details).

Another challenge for relation classification is in modeling the order of entities in a directed relation type (Lee et al., 2017). In the six corpora, several relations are directed and order-sensitive, such as the *Cause-Effect* relation in SemEval and *Hyponym-of* in ScienceIE. Such relations require the model to predict both relation types and the entity order correctly. In contrast, for undirected relations, such as *Synonym-of* in ScienceIE and *Associated* in Phenebank, both directions can be accepted.

An interesting factor is that the length of the SDP in SemEval is considerably shorter than in the other corpora. The mean and maximum length SDP values for CDR, BB3, ScienceIE and Phenebank are quite similar, *i.e.* ~ 7 and $22 - 26$ tokens. DDI-2013 contains very complex sentences, with an averaged SDP length of 9 and the longest SDP of 66 token.

Figure 1 shows the Out-Of-Vocabulary (OOV) ratios in six corpora, which are quite large, ranging from 23% to 57%. More interesting is the percentage of entities (or nominal) pairs in the test set that have never appeared in the training set (NP: 79% on CDR and more than 93% on SemEval, DDI-2013, ScienceIE and Phenebank). These two characteristics indicate the importance of understanding the mechanisms by which neural networks can generalize, *i.e.* make accurate predictions on novel instances.

3.2 Model Architecture

Our 'Man for All SeasonS' (MASS) model comprises an embeddings layer, multi-channel bi-directional Long Short-Term Memory (BLSTM)

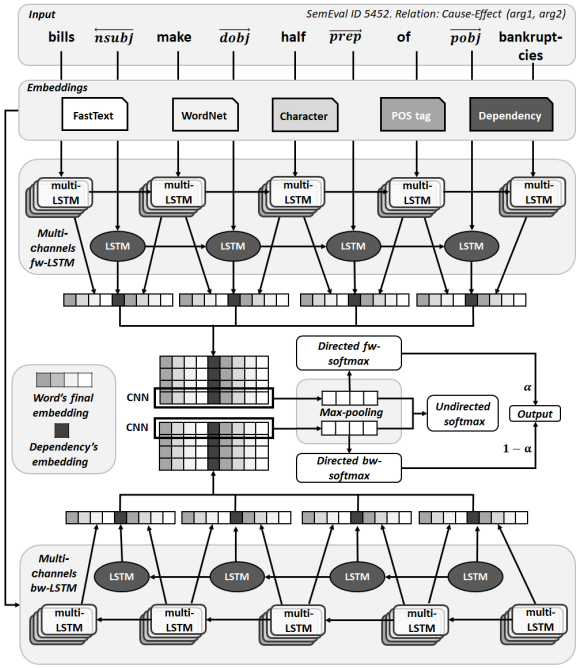

Figure 2: The architecture of MASS model for relation classification. An embeddings layer is followed by multi-channel bi-directional LSTM layers, two parallel CNNs and three *softmax* classifiers. The model's input makes use of words and dependencies along the SDP going from the first entity to the second one using both forwards and backwards sequences.

layers, two parallel Convolutional Neural Network (CNN) layers and three $softmax$ classifiers. The MASS model's architecture is depicted in Figure 2. MASS makes use of words and dependencies along the SDP going from the first entity to the second one using both forwards and backwards sequences. As is standard practice (Xu et al., 2015; Cai et al., 2016; Mehryary et al., 2016; Panyam et al., 2018) an entity pair is classified as having a relation if and only if the SDP between them is classified as having that relation.

3.2.1 Embeddings layer

Despite the presence of inter-sentential relations in the six corpora we make the simplifying assumption that relations occur only between entities (or nominals) in the same sentence. We model each such sentence using a dependency path. In order to classify novel dependency paths we represent a dependency relation d_i as a vector D_i that is the concatenation of two vectors as follow:

$$D_i = Dtyp_i \oplus Ddir_i$$

where $Dtyp$ is the undirected dependency vector, expressing the dependency type among 63 labels and, $Ddir$ is the orientation of the dependency vector *i.e.* from left-to-right or vice versa in the order of the SDP. Both are initialized randomly.

For **word representation**, we take advantage of four types of information, including:

- ***FastText pre-trained embeddings*** (Bojanowski et al., 2017) are the 300-dimensional vectors that represent words as the sum of the skip-gram vector and character n-gram vectors to incorporate sub-word information.
- ***WordNet embeddings*** are in the form of one-hot vectors that determine which sets in the 45 standard WordNet super-senses the tokens belong to.
- ***Character embeddings*** are denoted by $\mathbb{C}$, containing 76 entries for 26 letters in uppercase and lowercase forms, punctuation, and numbers. Each character $c_j \in \mathbb{C}$ is randomly initialized. They will be used to generate the token's character-based embeddings.
- ***POS tag embeddings*** capture (dis)similarities between grammatical properties of words and their lexical-syntactic roles within a sentence. We randomly initialized these vectors values for the 56 POS tags in OntoNotes v5.0.

Note that all initializations are generated by looking up the corresponding lookup table. The character and POS tag embeddings lookup tables were randomly constructed according to the $Glorot$ uniform initializer (Glorot and Bengio, 2010) and then treated as the model's parameters to be learned in the training phase.

3.2.2 Multi-channel Bi-LSTM

For a given linguistic feature type, LSTM networks (Hochreiter and Schmidhuber, 1997) are employed to capture long-distance dependencies along two directions, namely the forward and backward Bi-directional LSTM (BLSTM).

For the **dependencies**, BLSTMs take as input a sequence of dependency embeddings D_i, then gives output are the hidden states for dependencies between adjacent tokens w_i and w_{i+1} as $fwDEP_{ii+1}$ and $bwDEP_{ii+1}$.

Apart from the dependencies between tokens in SDPs, our model exploits four linguistic embeddings relating to words for **representing the**

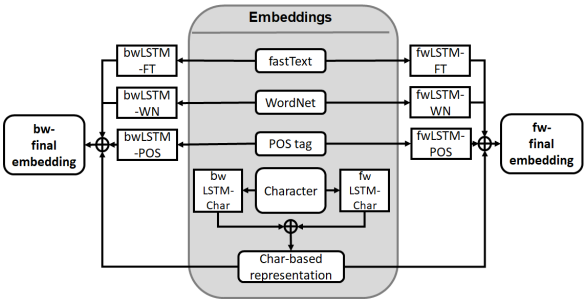

Figure 3: The multi-channel LSTM for word representation. Each token in the SDP is represented by using four word-related embeddings, including FastText word embedding, WordNet embedding, POS tag embedding and the character embedding. These four types of word-related information are fed into eight separate LSTMs, independently from each other during recurrent propagation.

words. These four types of word-related information are fed into eight separate LSTMs (four for each direction) independently from each other during recurrent propagation. These four BLSTM channels are illustrated in Figure 3. The morphological surface information is represented with character-based embedding using a BLSTM, in which the forward and backward LSTM hidden states are jointly concatenated (Ling et al., 2015; Dang et al., 2018). For other layers, the LSTM hidden states are concatenated separately as the forward and the backward vector to form two final embeddings for each token as follows:

$$fwW_i = fwFT_i \oplus fwWN_i \oplus Char_i \oplus fwPOS_i$$

$$bwW_i = bwFT_i \oplus bwWN_i \oplus Char_i \oplus bwPOS_i$$

3.2.3 CNN with dependency unit

Similar to Cai et al. (2016), the Convolutional Neural Networks (CNNs) in our model utilize **Dependency Units (DU)** to model the SDP. DU has the form of $[w_i - d_{ii+1} - w_{i+1}]$, in which w_i, w_{i+1} are two adjacent tokens and d_{ii+1} is the dependency between them. As a result, the low-dimensional forward and backward representation vectors of DU_j are created by concatenating the corresponding final embeddings of tokens w_j, w_{j+1} and the LSTM hidden state of the dependency d_{ii+1}. Formally, we have:

$$fwDU_j = fwW_j \oplus fwDEP_{jj+1} \oplus fwW_{j+1}$$

$$bwDU_j = bwW_j \oplus bwDEP_{jj+1} \oplus bwW_{j+1}$$

The forward and backward SDP representation matrices fwS and bwS are created by stacking the $fwDU$ and $bwDU$ vectors. We then apply two parallel CNNs to fwS and bwS to capture the context features (CF_j) around each dependency unit DU_j in the SDP as follows. These CNNs are designed similarly to the original CNN for sentence classification (Kim, 2014).

$$fwCF_j = f(We_{CNN} \cdot fwDU_j + b_{CNN})$$

$$bwCF_j = f^{'}(We^{'}_{CNN} \cdot bwDU_j + b^{'}_{CNN})$$

where We_{CNN} and $We^{'}_{CNN}$ are the weight matrices for the CNNs, b_{CNN} and $b^{'}_{CNN}$ are the bias terms for the hidden state vectors and f and $f^{'}$ are the non-linear activation functions.

The $n-$max pooling (Boureau et al., 2010) layer gathers the most useful global information G over the whole SDP (Collobert et al., 2011) from the context features of dependency units, which is defined as follows (in this work, we use $1-$max pooling).

$$fwG = \max_{d=1}^{k} fwCF_j$$

$$bwG = \max_{d=1}^{k} bwCF_j$$

where max is an element-wise function, and k is the number of dependency units in the SDP.

3.2.4 Softmax classifiers

Following (Cai et al., 2016), relation classification based on fwS and bwS simultaneously can strengthen the model's ability to judge the direction of relations. We, therefore, use two directed $softmax$ classifiers, one for each direction of the relation, with linear transformation to estimate the probability that each of fwS and bwS belongs to a directed relation (the direction taken into account). Formally we have:

$$p(fw) = softmax(W_f \cdot fwG + b_f)$$

$$p(bw) = softmax(W^{'}_f \cdot bwG + b^{'}_f)$$

where W_f and $W^{'}_f$ are the transformation matrices and b_f and $b^{'}_f$ are the bias vectors.

These two distributions are then combined to get the final distribution with a priority weight α:

$$p = \alpha \cdot p(fw) + (1 - \alpha) \cdot p(bw)$$

We also use the undirected $softmax$ to predict undirected distribution $p(ud)$. This $softmax$ is only used in the training objective function, which is the penalized cross-entropy of three $softmax$

classifiers. Our undirected softmax is quite similar to the idea of coarse-grain softmax used in Cai et al. (2016); Zhou et al. (2018).

$$p(ud) = softmax(W_f^{''} \cdot [fwG \oplus bwG] + b_f^{''})$$

where $W_f^{''}$ is the transformation matrix and $b_f^{'}$ is the bias vector.

3.3 Additional Techniques

Mehryary et al. (2016) demonstrated that random initialization can, to some extent, have an impact on the model's performance on unseen data, i.e, individual trained models may perform substantially better (or worse) than the averaged results.

Further, an ensemble mechanism, was found to reduce variability whilst yielding better performance than the averaging mechanism. Two simple but effective ensemble methods include strict majority vote (Mehryary et al., 2016) and weighted sum over results (Ammar et al., 2017; Lim et al., 2018; Verga et al., 2018). Since the former brings better results in our experiments, our ensemble system runs the model for 20 times and uses the strict majority vote to obtain the final results.

For dealing with the imbalanced data problem, we apply an under-sampling technique (Yen and Lee, 2006) during pre-processing for the DDI-2013 and Phenebank corpora. For a fair comparison we also apply some simple rules that was used by comparison models as the pre/post-processing step for DDI-2013 (following Zhou et al. (2018)), BCR (following Gu et al. (2017)) and ScienceIE (following Lee et al. (2017)) (for further details, see Appendix A).

Finally, we use several techniques to overcome over-fitting, including: *max-norm regularization* for Gradient descent (Qin et al., 2016); adding *Gaussian noise* (Quan et al., 2016) with mean 0.001 to the input embeddings; applying *dropout* (Srivastava et al., 2014) at 0.5 after all embedding layers, LSTM layers and CNN layers; and using *early stopping* technique (Caruana et al., 2000).

4 Results and Discussion

For each benchmark dataset we adopt the official task evaluations for system with $F1$ score, precision P and recall R. All official evaluations only considered the actual relations (excluding the *Other* relation and negatives) and worked on the abstract level (excepted SemEval). For a clearer comparison, we also report both averaged and ensemble results, in which, the averaged results are calculated over 20 different runs. Both results of the MASS model with and without applying pre/post-processing rules are also reported.

Model	Source of information	F1
SVM (Rink and Harabagiu, 2010)	Rich features	82.2
CNN + Attention Shen and Huang (2016)	Position, WordNet, words around nominals	85.9
BLSTM + CNN (Cai et al., 2016)	NER, WordNet w/o inversed SDP* w/ inversed SDP	 83.8 86.3
BLSTM + CNN + attention (Zhang et al., 2018)	Position embedding	83.7
Baseline model	WordNet, Character embeds	85.0
MASS model	WordNet, Character embeds (+ Inversed SDP) + Ensemble	85.9 85.4 **86.3**

Table 3: Comparison of our system with top performing systems on the SemEval 2010 corpus. The official evaluation is based on the macro-averaged F1. Since most of the comparative models did not report their P and R, we only report our F1 for comparison. All deep learning models use word embedding and POS tag information. *We report results for our implementation of Cai et al.'s system, without using the inversed SDP.

We compare the performance of the MASS model against three types of competitors: (i) A baseline model is used to verify the effectiveness of the multi-channel LSTM, in which we concatenate all embedding vectors used in MASS directly. (ii) The first ranked in the original challenges. (iii) Recent models with state-of-the-art results. The comparative results are shown in Tables 3 - 8.

In all corpora, the MASS model's results are always better than the baseline model. This is because directly concatenating many vectors with various value ranges seems to be causing information interference, and we cannot take advantages of each sequence of information separately anymore.

In **SemEval2010 corpus** (see Table 3), the macro-averaged $F1$ of the original model is 85.9% with the standard deviation of 20 runs is 0.33. This result outperforms all comparative models but Cai et al. (2016) which fed the inversed SDP to enrich the training data (we also tried feeding inversed SDP to the model, but the result became worse since this technique may be unsuitable for our model). Applying ensemble procedure boosts $F1$ for 0.45%, outperforming all comparative models.

Model	Source of information	P	R	F1
2-phase classification Hybrid kernel SVM[1]	Heterogeneous set of feature, rule-based negative filtering	64.6	65.6	65.1
2-phase classification SVM[2]	Rich features	73.6	70.1	71.8
BLSTM + Attention (Zhou et al., 2018)	Position-aware attention + Pre-processing	75.8	70.3	73.0
Baseline model	WordNet, Character embeds	51.6	52.9	52.2
MASS model	WordNet, Character embeds	54.0	56.3	55.1
	+ Ensemble	56.5	57.3	56.0
	+ Pre-processing	57.0	56.5	56.7

Table 4: Results on the DDI-2013 corpus. The official evaluation is the micro-averaged P, R and F1 at abstract-level. Note that all deep learning models use word embedding and POS tag information. [1]Chowdhury and Lavelli (2013). [2]Raihani and Laachfoubi (2017).

For dealing with **DDI-2013** (see Table 4)- an imbalanced data, comparative models often consider it as two sub-tasks, *i.e.* detection and classification. Chowdhury and Lavelli (2013); Raihani and Laachfoubi (2017) applied a two-phrase classification, in which one classifier detects positive instance and the other then classifies them. Zhou et al. (2018) used a binary softmax together with a multi-class softmax. Obviously, our model encounters a serious problem with imbalanced data. Since we treat the RE problem as a multi-class classification, in which, negative is also considered as a class, our results are much lower than comparative models. We applied negative undersampling technique and the pre-processing rules from Zhou et al. (2018) to remove some negatives, however the rules improved performance only slightly (0.3%).

Since our system just extracts the relations within a sentence, for **CDR** (see Table 5)- a corpus where 30% instances are cross-sentence relations, it is reasonable to explain why our recall is much lower than the comparative systems that can extract cross-sentences relations (Gu et al., 2017; Verga et al., 2018). Our results are still extremely encouraging since the $F1$ is better than other models which do not extract cross-sentences relations (Gu et al., 2017; Panyam et al., 2018). For a clearer comparison, we also try applying post-processing rules used by Gu et al. (2017), and they help to increase the $F1$ by 3.3%. Our $F1$ is just a little lower than the combined model of CNN and ME which extracts cross-sentence relations (Gu et al., 2017). The results for BRAN (Verga et al., 2018) however are much better than our MASS model. It is a a strong competitor on this benchmark that is designed to focus on cross-sentence relation classification by creating the document-level graph and is also trained using auxiliary data.

Model	Source of information	P	R	F1
CNN + ME[1] (Gu et al., 2017)	Contextual of whole sentence	59.7	57.5	57.2
	+ Cross-sentence	60.9	59.5	60.2
	+ Post processing	55.7	68.1	**61.3**
ASM[2] (Panyam et al., 2018)	Dependency graph	49.0	67.4	56.8
BRAN[3] (Verga et al., 2018)	Position, multi-head att	55.6	70.8	62.1
	+ Data	64.0	69.2	**66.2**
	+ Ensemble	63.3	67.1	**65.1**
Baseline model	WordNet, character embeds	56.6	54.1	55.3
MASS model	WordNet, character embeds	58.9	54.9	56.9
	+ Ensemble	56.8	57.9	57.3
	+ Post-processing	52.8	71.1	60.6

Table 5: Results on the CDR corpus. The official evaluation is reported at abstract-leve. All deep learning models use word embedding and POS tag information. [1]CNN + Maximum Entropy. [2]Approximate Subgraph Matching. [3]CNN + attention at abstract-level graph.

Model	Source of information	P	R	F1	IntraF
VERSE (SVM)[1]	Rich features	51.0	61.5	55.8	63.4
TurkuNLP (RNN)[2]		62.3	44.8	52.1	62.0
DET-BLSTM (Li et al., 2017)	Dynamic ext dep tree, distance embeddings	56.3	58.0	**57.1**	–
Baseline model	WordNet, Char embds	60.8	47.2	53.1	62.5
MASS model	WordNet, Char embds	59.8	51.3	55.2	64.6
	+ Ensemble	59.2	52.2	55.5	**64.8**

Table 6: Results on the BB3 corpus. The official evaluation is reported at both abstract- and intra sentence levels. All deep learning models use word embedding and POS tag information. [1]Lever and Jones (2016). [2]Mehryary et al. (2016)

In the **BB3** corpus (see Table 6), the original system outperforms all previously reported results at intra-sentence F. Using ensemble procedure, our results increase, but not much and still lower than the DT-BLSTM model, which is based on Dynamic Extended Tree (Li et al., 2017).

In the **ScienceIE** corpus (see Table 7), our results are only outperformed by one competitor. The reason may come from the characteristic of Hyponym-of and Synonym-of relations. Neither of these relations is expressed frequently by the linguistic information of tokens appearing in the SDP. In many cases, they are represented by different patterns with the same SDP. Therefore, our conclusion is that maybe the use of SDP does not match the ScienceIE corpus. The system from

Model	Source of information	F1
NTNU-2 (SVM) (Barik and Marsi, 2017)	Rich features	50.0
MIT (CNN) (Lee et al., 2017)	Relative position, NER + Post-processing	**64.5**
S2_rel (BLSTM) (Ammar et al., 2017)	Semisupervised, language model	54.1
	+ Ensemble	55.2
Baseline model	WordNet, character embeds	48.7
MASS model	WordNet, character embeds	54.6
	+ Ensemble	56.4
	+ Post- processing (Lee et al., 2017)	60.3
	+ Post- processing (rules ++)	**73.0**

Table 7: Results on the ScienceIE corpus. The official evaluation is based on the micro-averaged F1 at abstract-level. Since most of comparative models did not report their P and R, we only report our F1 for comparison. All deep learning models use word embedding and POS tag information.

			Baseline	Averaged	Ensemble
Sentence level	Macro-averaged	P	45.8	43.6	44.2
		R	39.2	42.6	41.1
		F	42.2	**43.1**	42.6
	Micro-averaged	P	56.5	53.2	55.4
		R	56.2	62.3	62.3
		F	56.4	57.3	**58.7**
Abstract level	Macro-averaged	P	45.8	43.6	44.2
		R	27.3	29.7	28.4
		F	34.3	**35.3**	34.6
	Micro-averaged	P	56.5	53.2	55.5
		R	37.5	41.6	41.6
		F	45.1	46.7	**47.5**

Table 8: Experimental results on the Phenebank corpus for the MASS model.

MIT (Lee et al., 2017) fed the whole sentence with the relative position as input, therefore it may catch many useful patterns which did not appear in the SDP. To test this hypothesis, we apply the post-processing rules used in Lee et al. (2017) and boosted $F1$ by 3.8%. In addition, when we applied some more simple linguistic rules to identify synonyms and hyponyms, the results improved beyond expectations by 16.6%, totally outperformed all other models.

For **Phenebank** (see Table 8), since this new corpus did not have an official evaluation, we report all possible MASS results. The micro-averaged results are much better than the macro-averaged. It is reasonable since Phenebank is an extremely imbalanced corpus, in which we can expect poor accuracy for rare classes, which together account for about 1% of positive data (and positive data only account for 23% of the whole corpus). The micro-averaged and macro-averaged results of the proposed model are always better than the baseline model, in both abstract and sentence-level. Interestingly, the ensemble model boosts the micro-averaged results (1.33% of $F1$ at sentence-level and 0.88% of $F1$ at abstract-level), but brings lower macro-averaged $F1$ (decreased 0.51% and 0.77% of $F1$ at sentence- and abstract-level respectively).

4.1 Components and Information resources

We study the contribution of each model's component and information sources to the system performance by ablating each of them in turn from the model and afterwards evaluating the model on all corpora. We compare these experimental results with the full system's results and then illustrate the changes of F1 in Figure 4. The changes of $F1$ show that all model's components and information sources help the system to boost its performance (in terms of the increments in $F1$) in all corpora. The contribution, however, varies among components, information types and among corpora.

Among information sources, FastText embedding (FT) often has the most important contribution, while using WordNet (WN) brings quite small improvements. Some examples clearly demonstrate that the impact of information sources varies greatly between benchmarks. The dependency embedding (DEP) and type embedding ($Dtyp$) have a very strong influence over the results in DDI-2013 and ScienceIE corpora but not much in other corpora. Furthermore, POS tag information (POS) plays a very important role in the BB3 corpus, surpassing FT, while its contribution in other corpora is not significant.

Also, the impact of model components shows relatively inconsistent across corpora. The baseline models always have lower $F1$ than MASS. This demonstrates the advantage of using a multi-channel LSTM to represent various linguistic information. Furthermore, the contributions of multi-channel LSTM and CNN are quite balanced. Interestingly, the undirected softmax always benefits the result although it was only used to calculate the penalty in the training step.

These experiments prove the effectiveness of using various information as well as architectural components. More importantly, these results show that our proposed MASS model can automatically adjust to each corpus, highlighting the flexibility of the MASS model which is able to adapt to various datasets with many different characteristics.

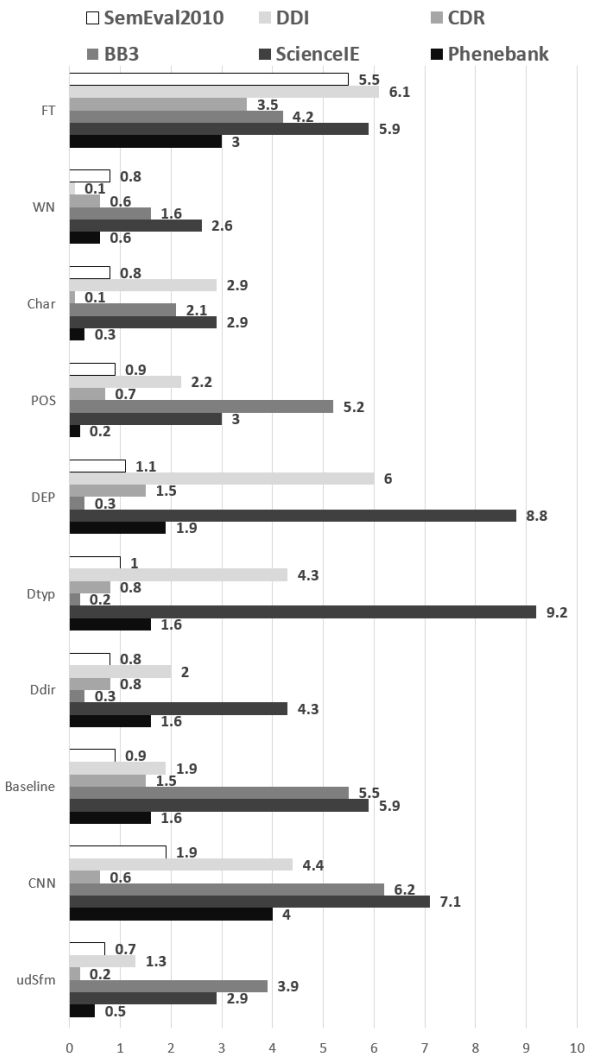

Figure 4: Ablation test results for various components and information sources: FastText (FT), WordNet (WN), Character-based (Char), POS tag, Dependency (DEP), dependency type (Dtyp) and dependency direction embedding (Ddir). Results are calculated based on the averaged F1 over 20 different runs. Baseline: Concatenating all embedding vectors to represent the words instead of using multi-channel LSTM. CNN: Using the final LSTM hidden states instead of CNN. udSfm: Removing the undirected softmax

4.2 Error Analysis

We studied model outputs to analyze system errors that defined the limitations of the model as well as to prioritize future directions. Many errors seem attributable to the parser. In some cases, we cannot generate the SDP, and in some cases where we have the SDP, information on the SDP is still insufficient or redundant to make the correct prediction. The directionality of relations is also challenging; in some cases the relation is predicted correctly but in the wrong direction. Other errors can be attributed to the limitations of our model, including (a) the inability to extract cross-sentence relations (accounting for 30% in CDR, BB3 and Phenebank), (b) the over-fitting problem (leading to wrong prediction - FP) and (c) limited generalisation power in predicting new relations (FN). Finally, we found some errors caused by the imperfect annotation. This problem may come from the different annotations assigned independently by two annotators (see IAA column in Table 2). We illustrate the above issues using realistic examples in Appendix C.

5 Conclusions

In this paper, we have presented a novel well-balanced relation classification model that consists of several deep learning components applied to the Dependency Unit of Shortest Dependency Path. We evaluated our model on six benchmark datasets, comparing the results with 15 recent state-of-the-art models. Experiments were also carried out to verify the rationality and impact of various model components and information sources. Experimental results demonstrated the robustness and adaptability of our system to classify different relation types in various domains without any architectural changes.

One existing issue with our model lies in its sensitiveness to class imbalance. This limitation resulted in significantly low performance on the DDI-2013 corpus (compared to state-of-the-art results). Our experiments also highlighted the existing challenges for neural relation classification models, including cross-sentence relations and imbalanced data. We aim to address these problems in future work.

Acknowledgments

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.052016.14. We also gratefully acknowledge the funding support of the EPSRC (N.Collier - Grant No. EP/M005089/1) and MRC (M. T. Pilehvar - Grant No. MR/M025160/1) for PheneBank. We also thank the anonymous reviewers for their comments and suggestions.

References

Waleed Ammar, Matthew Peters, Chandra Bhagavatula, and Russell Power. 2017. The ai2 system at semeval-2017 task 10 (scienceie): semisupervised end-to-end entity and relation extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 592–596.

Chinatsu Aone and Mila Ramos-Santacruz. 2000. Rees: a large-scale relation and event extraction system. In *Proceedings of the sixth conference on Applied natural language processing*, pages 76–83.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, and Lakshmi Vikraman andAndrew McCallum. 2017. Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 546–555. Association for Computational Linguistics.

Biswanath Barik and Erwin Marsi. 2017. Ntnu-2 at scienceie: Identifying synonym and hyponym relations among keyphrases in scientific documents. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 965–968.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118.

Razvan C Bunescu and Raymond J Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765.

Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS)*, pages 402–408.

Paulo R Cavalin, Fillipe Dornelas, and Sérgio MS da Cruz. 2016. Classification of life events on social media. In *29th SIBGRAPI (Conference on Graphics, Patterns and Images)*.

Md. Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. Fbk-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*, pages 351–355. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Bharath Dandala, Diwakar Mahajan, and Murthy V Devarakonda. 2017. Ibm research system at tac 2017: Adverse drug reactions extraction from drug labels. In *TAC*.

Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. 2018. D3ner: Biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 358–365.

Louise Delèger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferrè, Philippe Bessières, and Claire Nèdellec. 2016. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, pages 12–22. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10). Society for Artificial Intelligence and Statistics*.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database (Oxford)*, 2017:bax024.

Harsha Gurulingappa, Abdul Mateen-Rajpu, and Luca Toldo. 2012. Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1):15.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drugdrug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ferdaous Jenhani, Mohamed Salah Gouider, and Lamjed Ben Said. 2016. A hybrid approach for drug abuse events extraction from twitter. *Procedia computer science*, 96:1032–1040.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1746–1751.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017. Mit at semeval-2017 task 10: Relation extraction with convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 978–984.

Jake Lever and Steven JM. Jones. 2016. Verse: Event and relation extraction in the bionlp 2016 shared task. In *Proceedings of the the 4th BioNLP Shared Task Workshop*, pages 42–49.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database Oxford*, 2016:baw068.

Jiwei Li, Alan Ritter, Claire Cardie, and Eduard Hovy. 2014. Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1997–2007.

Lishuang Li, Jieqiong Zheng, and Jia Wan. 2017. Dynamic extended tree conditioned lstm-based biomedical event extraction. *International Journal of Data Mining and Bioinformatics*, 17(3):266–278.

Sangrak Lim, Kyubum Lee, and Jaewoo Kang. 2018. Drug drug interaction extraction from the literature using a recursive neural network. *PloS One*, 13(1).

Wang Ling, Chris Dyer, Alan W. Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530. Association for Computational Linguistics.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*, pages 1211–1220. International World Wide Web Conferences Steering Committee.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

Farrokh Mehryary, Jari Björne, Sampo Pyysalo, Tapio Salakoski, and Filip Ginter. 2016. Deep learning with minimal training data: Turkunlp entry in the bionlp shared task 2016. In *Proceedings of the the 4th BioNLP Shared Task Workshop*, pages 73–81. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction with complex event classification using rich features. *Journal of bioinformatics and computational biology*, 8(01):131–146.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48.

Nagesh C. Panyam, Karin Verspoor, Trevor Cohn, and Kotagiri Ramamohanarao. 2018. Exploiting graph kernels for high performance biomedical relation extraction. *Journal of biomedical semantics*, 9(1):7.

Pengda Qin, Weiran Xu, and Jun Guo. 2016. An empirical convolutional neural network approach for semantic relation classification. *Neurocomputing*, 190.

Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. 2016. Multichannel convolutional neural network for biological relation extraction. *BioMed research international*.

Anass Raihani and Nabil Laachfoubi. 2017. A rich feature-based kernel approach for drug-drug interaction extraction. *International journal of advanced computer science and applications*, 8(4):324–3360.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2014. Lessons learnt from the ddiextraction-2013 shared task. *Journal of Biomedical Informatics*, 51:152–164.

Yatian Shen and Xuanjing Huang. 2016. Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 536–540.

Yulan Yan, Naoaki Okazaki, Yutaka Matsuo, Zhenglu Yang, and Mitsuru Ishizuka. 2009. Unsupervised relation extraction by mining wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1021–1029.

Show-Jane Yen and Yue-Shi Lee. 2006. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Proceedings of Intelligent Control and Automation*, pages 731–740.

Xiaobin Zhang, Fucai Chen, and Ruiyang Huang. 2018. A combination of rnn and cnn for attention-based relation classification. *Procedia Computer Science*, 131:911917.

Deyu Zhou, Lei Miao, and Yulan He. 2018. Position-aware deep multi-task learning for drugdrug interaction extraction. *Artificial intelligence in medicine*, In Press.