# Leveraging Unannotated Texts for Scientific Relation Extraction

| | |
|---|---|
| 著者 | Qin DAI, Naoya INOUE, Paul REISERT, Kentaro INUI |
| journal or publication title | IEICE Transactions on Information and Systems |
| volume | E101.D |
| number | 12 |
| page range | 3209-3217 |
| year | 2018-12-01 |
| URL | http://hdl.handle.net/10097/00125635 |

doi: 10.1587/transinf.2018EDP7180

PAPER

# Leveraging Unannotated Texts for Scientific Relation Extraction

**Qin DAI**[†a)], **Naoya INOUE**[†,††b)], **Paul REISERT**[††c)], ***Nonmembers***, *and* **Kentaro INUI**[†,††d)], ***Member***

**SUMMARY** A tremendous amount of knowledge is present in the ever-growing scientific literature. In order to efficiently grasp such knowledge, various computational tasks are proposed that train machines to read and analyze scientific documents. One of these tasks, Scientific Relation Extraction, aims at automatically capturing scientific semantic relationships among entities in scientific documents. Conventionally, only a limited number of commonly used knowledge bases, such as Wikipedia, are used as a source of background knowledge for relation extraction. In this work, we hypothesize that unannotated scientific papers could also be utilized as a source of external background information for relation extraction. Based on our hypothesis, we propose a model that is capable of extracting background information from unannotated scientific papers. Our experiments on the RANIS corpus [1] prove the effectiveness of the proposed model on relation extraction from scientific articles.

***key words:*** *relation extraction, scientific document, word embedding, semantically related word*

## 1. Introduction

In recent years, with an increase in the number of scientific papers, it is prohibitively time-consuming for researchers to review and fully-comprehend all papers. To effectively and quickly access a large amount of scientific papers and acquire useful knowledge, a wide variety of computational studies for structuralizing scientific papers has been conducted, such as Argumentative Zoning [2], BioNLP Shared Task [3], and ScienceIE Shared Task [4]. One fundamental study is Relation Extraction (RE). In this paper, we explore the task of RE as an approach for effectively and quickly accessing a large amount of scientific papers and acquiring relevant knowledge.

RE is the task of capturing predefined semantic relations between entities from text. Thus, our task consists of the following: given a sentence that has been annotated with entity[*] mentions, we aim towards extracting relations among entities. Suppose the following sentence[**]:

(1) $\overset{\textit{entity}}{\underline{\textbf{\textit{RTMs}}}}$ $\overset{\textit{entity}}{\underline{\textit{achieve}}}$ $\overset{\textit{entity}}{\underline{\textit{top}}}$ $\overset{\textit{entity}}{\underline{\textit{performance}}}$ *in* $\overset{\textit{entity}}{\underline{\textit{automatic,}}}$ $\overset{\textit{entity}}{\underline{\textit{accurate,}}}$ *and* $\overset{\textit{entity}}{\underline{\textit{language independent}}}$ $\overset{\textit{entity}}{\underline{\textbf{\textit{prediction}}}}$ *of* $\overset{\textit{entity}}{\underline{\textit{sentence-level}}}$ *and* $\overset{\textit{entity}}{\underline{\textit{word-level}}}$ $\overset{\textit{entity}}{\underline{\textit{statistical machine translation}}}$ $\overset{\textit{entity}}{\underline{\textit{(SMT)}}}$ $\overset{\textit{entity}}{\underline{\textit{quality}}}$*.*

In Example 1, one of the scientific relations we aim to extract is the relation APPLY_TO(*RTMs*, *prediction*), which means that *RTMs* is the method that is used for the action of *prediction*. For notational convenience, we refer to a sentence where a relation is extracted from as a *target sentence*, and we refer to the related entity pair as a *target entity* pair.

The task of RE for entity pairs can be seen as a classification task. Specifically, given all possible entity pair combinations from a target sentence, the task is to categorize each pair into relation types including predefined relations and non-relation. For example, in Example 1, given the pair (*RTMs*, *prediction*), the output would be APPLY_TO(*RTMs*, *prediction*), and given the entity pair (*RTMs*, *top*), it would be non-relation(*RTMs*, *top*), which means that they do not belong to a predefined relation. With this level of fine-grained analysis, many applications, such as scientific question answering (QA) and scientific paper summarization, can benefit.

Many previous works on RE exist in the general domain [5], [6]. The earlier approaches depend on complex feature engineering such as manually prepared lexical-syntactic patterns [7]–[9], [etc.]. Recently, Neural Network (NN)-based approaches achieve close or even better performance to earlier approaches without complicated manually prepared features [10]–[12]. In the context of scientific RE, Ammar et al. [13] enhanced Miwa and Bansal [14]'s end-to-end general relation extraction model by incorporating external knowledge such as gazetteer-like information extracted from Wikipedia. However, no previous work leverages raw scientific documents as a source of background knowledge for RE.

In this work, we hypothesize that unannotated scientific papers can be utilized as a source of background knowledge for scientific RE. We attribute this to the fact that firstly the annotation scheme of scientific relations is based on scientific concepts such as Computer Science (CS) related concepts [1] like "Input" and "Computational_model", and biochemistry related concepts [15] like "Phosphorylate" and

Manuscript received May 18, 2018.
Manuscript publicized September 14, 2018.

[†]The authors are with Tohoku University, Sendai-shi, 980–8579 Japan.
[††]The authors are with RIKEN Center for Advanced Intelligence Project, Tokyo, 103–0027 Japan.
a) E-mail: daiqin@ecei.tohoku.ac.jp
b) E-mail: naoya-i@ecei.tohoku.ac.jp
c) E-mail: preisert@ecei.tohoku.ac.jp
d) E-mail: inui@ecei.tohoku.ac.jp

DOI: 10.1587/transinf.2018EDP7180

[*] In this work, *entity* refers not merely to concepts denoted by noun or noun phrase, it could be actions denoted by verb or verb phrase, and evaluation denoted by adjective or adverb etc.

[**] This example is taken from W13-2242, ACL anthology (http://aclanthology.info).

"Myristoylated_by". This implies that the corpus annotator is required to have external background knowledge about these scientific concepts such as "which entity is a computational_model/corpus/featrue". Secondly, the background information about these concepts are detailed in scientific paper. For instance, CS papers describe the background knowledge [1], which is like *"... proposed Database Semantics as a computational model for natural language semantics ... "*. Therefore, we hypothesize that if a RE system performs similar to the human annotator, the RE system will need to share with the human annotator similar background information about these scientific concepts, which could be extracted from scientific papers. In other words, we hypothesize that the background information about these CS related concepts can be automatically extracted from unannotated CS papers, and the extracted background information can facilitate RE in CS related dataset such as Tateishi et al. [1]'s RANIS corpus, which will be detailed in Sect. 3. Suppose the following sentence taken from the RANIS corpus:

(2) *$\underline{RTMs_A}$ achieve top performance in automatic, accurate, and language independent $\overline{prediction_B}$ of sentence-level and word-level statistical machine translation (SMT) quality.*

In Example 2, without any support of background information regarding the concept *RTMs*, such as "what is a *RTM*" (e.g., "computational model", "research team members", or "dataset"), its relation to the entity *prediction* can seem ambiguous. Specifically, if *RTMs* refers to a "computational model", a RE system might extract APPLY_TO(*RTMs*, *prediction*) relation, because the target sentence in Example 2 means that *RTMs* is the method or computational model that is **applied to** the action *prediction*. However, if *RTMs* refers to "research team members", the relation would be extracted as PERFORM(*RTMs*, *prediction*). Finally, if *RTMs* refers to a "corpus", the relation tends to be INPUT(*RTMs*, *prediction*).

Although the target sentence in Example 2 lacks enough background information about the target entity for disambiguating relation extraction, we could find the following sentences about the target entity *RTMs* from other sections of the same paper (Examples 3 and 4):

(3) *Referential translation* ***machines*** *(RTMs) provide a* ***computational model*** *for quality and semantic similarity judgments using retrieval of relevant training data ...*

(4) *... we* ***use*** *RTMs to automatically assess the correctness of student answers to obtain better result than the sate-of-the-art.*

Example 3 describes that the concept *RTMs* refers to a machine that could act as a *computational model*, and Example 4 mentions that *RTMs* could be **used** for some process. As discussed before, this information could be leveraged as background knowledge for disambiguating the relation as APPLY_TO(*RTMs*, *prediction*) rather than PERFORM(*RTMs*, *prediction*) or INPUT(*RTMs*, *prediction*), because *RTMs* is semantically closer to *computational model* rather than *research team members* or *corpus* in Examples 3 and 4.

For utilizing background knowledge, one possibility is to manually annotate useful background information about CS related concepts, such as "*RTMs* are a Computational Model" and "Using *WordNet* as a knowledge base", in scientific papers and apply the annotated scientific papers to RE. However, manual annotation is time consuming [16] and expensive [17].

To address this issue, in this work, we investigate the effectiveness of leveraging unannotated text for RE. Specifically, we propose two methods, term sentence (TS) and semantically related word (SRW), for automatically extracting background knowledge from unannotated scientific papers and utilizing the extracted background information for extending a state-of-the-art neural RE model. Our evaluation empirically demonstrates that incorporating the extracted TS and SRW from unannotated scientific papers improves the performance of RE.

## 2. Related Work

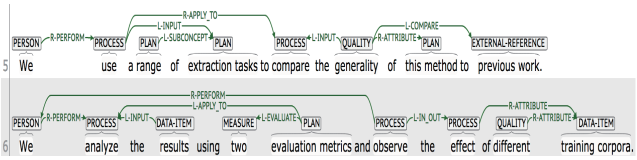
Conventional approaches for RE rely on human-designed, complex lexical-syntactic patterns [7], statistical co-occurrences [8] and structuralized knowledge bases such as WordNet [9], [18]. In recent years, exploring Neural Network (NN)-based models has been the dominant approach in the field. Zeng et al. [10] proposed a deep Convolutional Neural Network (CNN)-based framework, which depends on sentence-level features collected from an entire target sentence and lexical-level features from lexical resources such as WordNet [19]. Santos et al. [12] proposed a ranking CNN model, which is trained by a pairwise ranking loss function. To improve the ability of sequential modeling, Zhang et al. [11] proposed a recurrent neural network (RNN)-based model for RE. Other variants of RNN-based models have been proposed, such as Miwa et al. [14], who proposed a bidirectional tree-structured LSTM model. Additionally, similar NN-based approaches are used in scientific relation extraction. For instance, Gu et al. [20] utilized a CNN-based model for identifying *chemical-disease* relations from the abstracts of MEDLINE papers. Hahn-Powell et al. [21] proposed an LSTM-based RNN model for identifying *causal precedence* relationship between two event mentions in biomedical papers. Ammar et al. [13] enhanced Miwa and Bansal [14]'s relation extraction model via extensions such as gazetteer-like information extracted from Wikipedia. However, none of these approaches leverage unannotated scientific papers for RE.

## 3. Data

We evaluate the performance of RE using the RANIS cor-

**Table 1** Frequently appeared relation tags

| Type | Definition | Example |
|---|---|---|
| ATTRIBUTE(A, B) | B is an attribute or a characteristic of A | *accuracy*$_{\mathrm{A}}$ *of the tagger*$_{\mathrm{B}}$ |
| OUTPUT(A, B) | B is the output of a system or a process A; B is generated by A | *an image*$_{\mathrm{B}}$ *displayed*$_{\mathrm{A}}$ *on a palm* |
| APPLY_TO(A, B) | a method A is applied to achieve the purpose B | *CRF*$_{\mathrm{A}}$*-based tagger*$_{\mathrm{B}}$ |
| INPUT(A, B) | B is the input of a system or a process A; B is consumed by A | *corpus*$_{\mathrm{A}}$ *for training*$_{\mathrm{B}}$ |
| EVALUATE(A, B) | A is evaluated as B | *experiment shows an increase*$_{\mathrm{B}}$ *in F-score*$_{\mathrm{A}}$ *compared to the baseline* |
| SUBCONCETP(A, B) | A is-a, or is a part-of B | *a corpus*$_{\mathrm{B}}$ *such as PTB*$_{\mathrm{A}}$ |
| CONDITION(A, B) | The condition A holds in situation B, e.g, time, location, experimental condition | *a survey*$_{\mathrm{B}}$ *conducted in India*$_{\mathrm{A}}$ |
| EQUIVALENCE(A, B) | terms A and B refer to the same entity: definition, abbreviation, or coreference | *DoS*$_{\mathrm{B}}$ *(denial-of-service*$_{\mathrm{A}}$*) attack* |
| PERFORM(A, B) | A is the agent of an intentional action B | *a frustrated player*$_{\mathrm{A}}$ *of a game*$_{\mathrm{B}}$ |
| IN_OUT(A, B) | B is simultaneously INPUT and OUTPUT and is changed by a system or a process A | *a modified*$_{\mathrm{A}}$ *annotation schema*$_{\mathrm{B}}$ |


5 We use a range of extraction tasks to compare the generality of this method to previous work.

6 We analyze the results using two evaluation metrics and observe the effect of different training corpora.

**Fig. 1** Annotation example shown in brat rapid annotation tool. To more clearly illustrate the direction of relation, we add directional tag "L-" (means left hand side is the argument B) and "R-" (means right hand side is the argument B) before each relation tag.

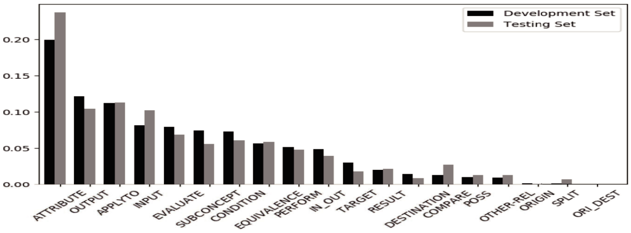


**Fig. 2** Distribution of relation types.

pus [1], a collection of computer science paper abstracts. The type of entity (referred to as Entity Type (ET) hereafter) and domain specific relation in the RANIS corpus has already been annotated with the annotation scheme proposed by [1], as shown Fig. 1. The corpus consists of ETs such as QUALITY, PROCESS and DATA-ITEM and domain specific scientific relations, such as INPUT, OUTPUT and APPLY_TO. Table 1 summarizes frequently appearing domain specific relations and provides both definitions and examples.

In total, the RANIS corpus contains 250 abstracts collected from ACL Anthology (230 abstracts in the development set and 20 abstracts in the test set) and 150 abstracts collected from ACM Digital Library. For training and testing our proposed model, we only use the 250 abstracts from ACL Anthology. From the ACL Anthology abstracts, we extract 11,520 relations from the development set of ACL Anthology and 1,142 relations from the test set of ACL Anthology. The distribution of relation types for both sets is shown in Fig. 2. For each ACL anthology abstract in the RANIS corpus, we collect its corresponding unannotated paper body from ACL Anthology Reference Corpus [22] as the source of background information for RE.

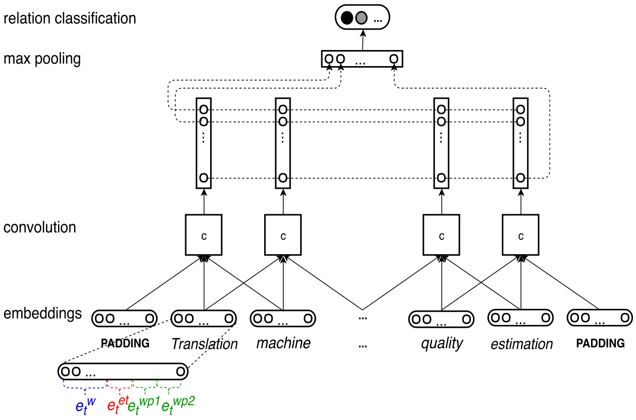


**Fig. 3** Baseline model architecture

## 4. Baseline Models

The baseline model is proposed by Santos et al. [12]. As shown in Fig. 3, it is composed of three layers. The first layer is an embedding layer, which maps each word of the target sentence into a low-dimensional word vector representation. The embedding layer is calculated via Eqs. (1)–(4), where $W_{emb}^{w}$ is a word embedding projection matrix, $W_{emb}^{et}$ is an entity type (ET) projection matrix, $x_t^w$ is a one-hot word representation, and $x_t^{et}$ is a one-hot entity type representation. The position vector $e_t^{wp}$ encodes the relative distance between the current word and the head of target entity pair. For instance, in Example 5, the relative distance of the word *"for"* is $[-1, 2]$.

(5) $\overset{entity}{\underline{We}}$ $\overset{entity}{\underline{introduce}}$ $\overset{entity}{\underline{referential\ translation\ machines}}$ $(\overset{entity}{\underline{RTM}}_A)$ *for* $\overset{entity}{\underline{quality\ estimation}}_B$ ...

This relative distance will be encoded into position vectors $e_t^{wp1}$ and $e_t^{wp2}$, respectively, via Eq. (3), where $W_{emb}^{wp}$ is a word position embedding projection matrix and $x_t^{wp}$ is a one-hot representation of the relative distance. Word embedding $e_t^w$, entity type embedding $e_t^{et}$ and word position embedding $e_t^{wp1}$ and $e_t^{wp1}$ are concatenated to create the final word representation $e_t$.

$$e_t^w = W_{emb}^w x_t^w \tag{1}$$

$$e_t^{et} = W_{emb}^{et} x_t^{et} \tag{2}$$

$$e_t^{wp} = W_{emb}^{wp} x_t^{wp} \tag{3}$$

$$e_t = concat(e_t^{w}, e_t^{et}, e_t^{wp1}, e_t^{wp2}) \tag{4}$$

$$z_t = concat(e_{t-(k-1)/2}, \ldots, e_{t+(k-1)/2}) \tag{5}$$

$$h_t = tanh(Wz_t + b) \tag{6}$$

The next layer is a convolutional layer, which generates a distributed convolutional window level vector $h_t$. $h_t$ is calculated by Eqs. (5) and (6), where $z_t$ is the concatenated embedding of $k$ words in the convolutional window, $k$ is convolutional window size, and $W$ is the weight matrix of the convolutional layer. In order to address the issue of referencing words with indices outside the sentence boundaries, the target sentence is padded with a special **PADDING** token $(k-1)/2$ times at the beginning and the end.

The third layer is a max pooling layer, which chooses the maximum value from each dimension of the convolutional window level feature and merges them as the sentence level feature $r$ via Eq. (7), where $i$ indexes feature dimensions, $M$ is the number of feature dimensions.

$$r_i = \max_t\{(h_t)_i\}, \ \forall i = 1, \ldots, M \tag{7}$$

Finally, the model predicts the semantic relationship between a target entity pair in a target sentence $x$, by computing the score for a class label $c \in C$ via dot product:

$$S_\theta(x)_c = r^T [W^{class}]_c, \tag{8}$$

where $C$ is a set of predefined semantic relationships, $r$ is the sentence level feature vector, and $W^{class}$ is the class embedding matrix. The column of $W^{class}$ represents the distributed vector representation of different class labels. It is worth mentioning that the model uses a logistic loss function, as shown in Eq. (9):

$$\begin{aligned} L = log(1 + exp(\gamma(m^+ - s_\theta(x)_{y^+})) \\ +log(1 + exp(\gamma(m^- + s_\theta(x)_{c^-})) \end{aligned} \tag{9}$$

where $s_\theta(x)_{y^+}$ is the score of correct class label, $s_\theta(x)_{c^-}$ is the score of the most competitive incorrect class label, $m^+$ and $m^-$ are margins, and $\gamma$ is a scaling factor. In our experiment, we use $m^+ = 2.5$, $m^- = 0.5$ and $\gamma = 2$.

## 5. Proposed Model

In this paper, we hypothesize that unannotated scientific papers can be utilized as a source of background information for RE. Therefore, we create a problem setting where we consider an annotated sentence in a paper abstract as a target sentence, and the corresponding unannotated paper body of the abstract (henceforth, *paper body*) as the source of background information. We hypothesize that the background information extracted from the paper body could facilitate relation extraction in paper abstracts. We believe that this setting can be easily adapted to a more general task setting, e.g. analyzing semantic relation in a whole document (not just in an abstract) via considering a collection of unannotated scientific papers as a source of background information.

Based on this hypothesis, we propose a new relation classification model that categorizes relations not only based on the target sentence, but also on the background information acquired from unannotated scientific papers, as illustrated in Sect. 1. To create such a model, we need to address the following questions:

1. From the perspective of knowledge acquisition, how do we extract the background information from unannotated scientific papers?
2. From the perspective of NN, how do we encode the extracted information into a vector representation for relation classification?

### 5.1 Retrieving Background Information from Unannotated Scientific Papers

For acquiring background knowledge from unannotated scientific papers, we propose two methods.

**Method 1:** extract all of the sentences containing the target entity of interest in the unannotated paper body as a representation of background information (henceforth, referred to as *Term Sentence(TS)*)[†]. Formally, $TS_A = w_{A1}, \ldots ent_A, \ldots, w_{Ai}, \ldots w_{An}$ and $TS_B = w_{B1}, \ldots ent_B, \ldots, w_{Bi}, \ldots w_{Bn}$, where $ent_A$ and $ent_B$ are target entities, $w_{Ai}$ ($w_{Bi}$) is the word of the sentence in which the target entity $ent_A$ ($ent_B$) exists. For example, given a target entity *RTM*, we could find the following TSs in its corresponding paper body:

(6) ***RTM** is a computational model for identifying the acts of translation for translating between any given two data sets with respect to a reference corpus selected in the same domain.*

(7) ***RTM** can be used for predicting the quality of translation outputs.*

Given multiple TSs for a target entity, this method simply concatenates all of the individual TSs (e.g., Examples 6 and 7) into an overall representation of TS and feeds it to the proposed model.

The intuition behind the method is that a TS could contain domain-specific background information about target entity for relationship analysis. For instance, Example 7 clearly mentions that "***RTM** can be used for predicting the quality* ..." and this is effective evidence for the existence of the scientific relationship APPLY_TO($\underline{RTM}_A$, $\underline{quality\ estimation}_B$) relationship in the target sentence (Example 8).

(8) *We introduce referential translation machines ($\underline{RTM}_A$) for $\underline{quality\ estimation}_B$ of translation outputs of*

[†] In this work, we only choose the noun phrase target entity to extract TS.

*sentence-level and word-level statistical machine translation (SMT) quality.*

**Method 2:** extract Semantically Related Word as a representation of background information for RE. In this work, we define SRW as the set of content words (e.g., nouns, verbs and adjectives) from a paper body that are semantically close to a given target entity.

The process of extracting SRW in this work is similar to the approach proposed by [23]. Specifically, based on word embeddings, we calculate cosine similarity between a given target entity (from a paper abstract) and each content word from its corresponding paper body, and then use a predefined criteria to select the member for its SRW. We manually set the SRW criteria (SRW_c) as 0.35, and only collect the word whose cosine similarity with the target entity is larger than the SRW_c as the member of SRW. The effect of SRW_c on RE performance will be discussed in Sect. 6.2. Formally, $SRW_A = \{w_{A1}, \ldots, w_{Ai}, \ldots w_{An} | cos(e_{ent_A}, e_{w_{Ai}}) > \text{SRW\_c}\}$ and $SRW_B = \{w_{B1}, \ldots, w_{Bi}, \ldots w_{Bn} | cos(e_{ent_B}, e_{w_{Bi}}) > \text{SRW\_c}\}$, where $SRW_A$ ($SRW_B$) is the SRW for entity A (B), $w_{Ai}$ ($w_{Bi}$) is the content words from the paper body, $e_{w_{Ai}}$ ($e_{w_{Bi}}$) is its word embedding and $e_{ent_A}$ ($e_{ent_B}$) is the word embedding of the target entity A (B).

The following example is a practical case of SRW extraction applied in this work. Given a target sentence (e.g. Example 9) with a marked target entity pair[†], the method automatically extracts SRW$_A$ and SRW$_B$, from its corresponding paper body for target entity pair, "***extraction***" and "***collections***"[††], respectively.

(9) *We are interested in the problem of word extraction$_A$ from Chinese text collections$_B$.*

SRW$_A$: *extraction, extracting, identification, retrieval, filtering*
SRW$_B$: *collections, corpora, sets, texts, corpus, data*

The intuition behind applying SRW for RE is inspired by its usage in word sense disambiguation [24]. Specifically, given an entity, its entity type might differ in distinct texts. For instance, the specific entity type for *"collections"* in Text1[†††] is different with the one in Text2[††††]. In Text1, *"collections"* belongs to the type of *corpus*, but in Text2, it refers to *parameters*. This difference could be illustrated by extracting SRW of *"collections"* from each Text, which is denoted in parenthesis. Since entity type information closely interacts with relation classification [14], [25], we hypothesize that SRW could illustrate the entity type information about target entity, thereby facilitating RE.

Text1: *Typically, a parallel training corpus is comprised of collections$_A$ of varying quality and relevance to the translation problem of interest.*
(*SRW$_A$: collections, corpus*)

Text2: *The model is defined by two collections$_A$ of parameters: the transition probabilities, which express the probability that a tag follows the preceding one (or two for a second order model); and the lexical probabilities,*
(*SRW$_A$: collections, parameters*)

For instance, suppose we intend to classify the relation between *"collections$_A$"* and *"model$_B$"* in the target sentence, *"We apply these collections$_A$ to train the model$_B$"*. In the context of Text1, the relation would be INPUT, because the SRW in Text1 indicates that *"collections"* is semantically similar to the entity *corpus*, and *corpus* is usually used as the input data for training a NLP model. In contrast, in the context of Text2, they have a low tendency to hold INPUT relation, when in fact, have high tendency to hold ATTRIBUTE relation, because in Text2, *"collections"* belongs to the type of *parameters*, and *parameters* is not the input data, but the attribute of the *"model"*. Similarly in Example 9, SRW$_B$ contains *"corpus"*, therefore the target entity, *"collections"*, has high tendency to participate in INPUT relation, which is the gold standard relation in RANIS corpus [1].

### 5.2 Architecture

The proposed NN model, in general, contains two main parts: Baseline model and Background Information Encoding model (BIE model, for short) as shown in Fig. 4. The former converts the target sentence into a vector representation, and the latter is responsible for converting the acquired TS pair and SRW pair into a vector representation.

The Baseline model is the CNN-based baseline model that has been described in Sect. 4. The BIE model, as shown in Fig. 4, is used for encoding SRW (or TS) of entity A and SRW (or TS) of entity B, thus having a parallel structure. The parallel CNN-model for each SRW (or TS) has independent convolutional weight matrix $W_1$ and $W_2$ but shares word embedding projection matrix $W^w_{emb}$. As shown in Fig. 4, BIE model consists of 3 layers: the first layer is the word embedding layer that maps each word from SRW or from TS into word vector via Eq. (10), where $X^{w_A}_t$ ($X^{w_B}_t$) is the one-hot of the word from $SRW_A$ ($SRW_B$) or from $TS_A$ ($TS_B$). The second layer is the convolutional layer, which generate the convolutional filter level vector $z^A_t$ and $z^B_t$ via Eqs. (11)–(13), where $k$ is the convolutional window size. The third layer is max pooling layer, which chooses a maximum value from each SRW (or TS) via Eq. (14), where $i$ indexes feature dimensions, $m$ is the number of feature dimensions. The final output of BIE model is calculated via Eq. (15).

$$e^{w_{A(orB)}}_t = W^w_{emb} x^{w_{A(orB)}}_t \tag{10}$$

$$z^{A(orB)}_t = concat(e^{w_{A(orB)}}_{t-(k-1)/2}, \ldots, e^{w_{A(orB)}}_{t+(k-1)/2}) \tag{11}$$

[†] This example is taken from J04-1004, ACL anthology (http://aclanthology.info).

[††] In this work, we only select the noun (phrase), verb (phrase) and adjective target entity and simply use its head word to extract SRW.

[†††] This example is taken from D09-1074, ACL anthology (http://aclanthology.info).

[††††] This example is taken from A94-1009, ACL anthology (http://aclanthology.info).

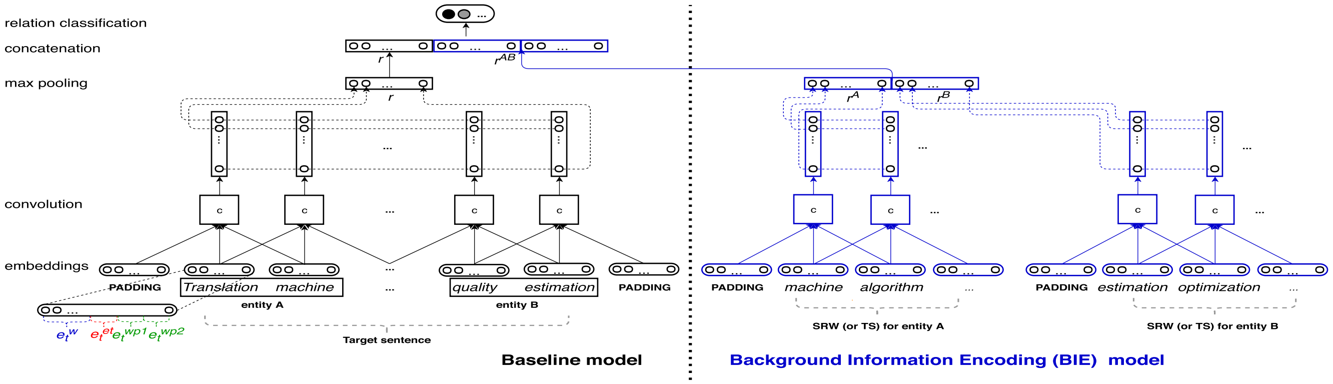


**Fig. 4** The architecture of the proposed model enhanced by LC (or TS) encoding.

**Table 2** Distribution of RELATED entity pairs.

| Data type | Percentage (RELATED/all) |
|---|---|
| training data | 17.0% (10,391/61,137) |
| validation data | 16.6% (1,129/6,792) |
| testing data | 17.1% (1,142/6,674) |

**Table 3** Hyperparameters for Relation Classification

| Parameter Name | Value |
|---|---|
| Word Emb. size | 200 |
| Word Entity Type Emb. size | 50 |
| Word Position Emb. szie | 100 |
| Convolutional Units (Baseline model) | 1000 |
| Context Window size (Baseline model) | 3 |
| Convolutional Units (BIE model) | 100 |
| Context Window size (BIE model) | 3 |
| The Number of Epoch | 25 |
| Learning Rate | 0.003 |

**Table 4** Performance of RE (mean ± standard deviation)

| Model | Precision | Recall | F-score |
|---|---|---|---|
| Baseline | 62.79±1.22 | 50.58±0.46 | 54.5±0.45 |
| Baseline + TS | 62.88±0.42 | 50.75±0.48 | 54.96±0.34 |
| Baseline + SRW | 63.02±0.7 | 51.67±0.52 | 55.56±0.46 |
| Baseline + TS + SRW | **65.14**±0.63 | **52.08**±0.58 | **56.47**±0.44 |

$$h_t^A = tanh(W_1 z_t^A + b_1) \tag{12}$$

$$h_t^B = tanh(W_2 z_t^B + b_2) \tag{13}$$

$$r_i^{A(orB)} = \max_t\{(h_t^{A(orB)})_i\}, \ \forall i = 1, \ldots, m \tag{14}$$

$$r^{AB} = concat(r^A, r^B) \tag{15}$$

Finally, the final vector representation of a SRW pair (or TS pair), $r^{AB}$, and the final output vector of the Baseline model, $r$, are concatenated and fed to a semantic relation classifier.

We use the back-propagation algorithm for training the model and choose the logistic loss function in Eq. (9) as the objective function.

## 6. Experiments

### 6.1 Setup

From the RANIS corpus, we extract 67,929 possible intra-sentence entity pairs from the ACL development set and 6,674 intra-sentence entity pairs from the ACL testing set. From the development set, we randomly select 90% of samples as training data and the rest as validation data for tuning hyper parameters such as the number of hidden layer dimensions, the number of epochs, learning rate, etc. In Table 2, we show the distribution of the RELATED entity pairs, which means that the entity pair belongs to a predefined relation such as INPUT. In Table 3, we show the selected hyper parameter values.

Previous works have shown that pre-trained word embeddings can improve training for relation extraction models [10]–[12]. Therefore, in this work, we trained scientific paper specific word embeddings on the ACL Anthology Reference Corpus [22] (in total: about 3 million sentences) by the skip-gram NN architecture made available by the Gensim word2vec tool[†]. We initialized the word embedding layer with the pre-trained domain-specific word embedding for RE.

We implemented the baseline model, proposed NN model, and the back-propagation algorithm with Theano [26]. To minimize the influence of random initialization of model parameters on RE, we ran each evaluation 5 times and took their mean value for comparison.

### 6.2 Result

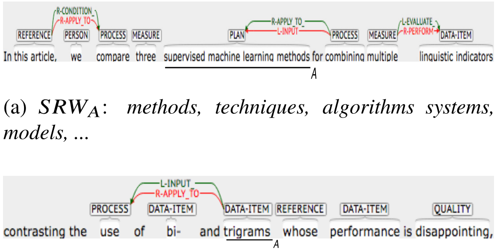
In this paper, we hypothesize that unannotated scientific papers could be used as a source of background information for scientific RE. We propose two methods for extracting background information: i) Term Sentence (TS), and ii) Semantically Related Word (SRW). For testing this hypothesis, we compare the performance of each method with the baseline approach, the CNN baseline model introduced in Sect. 4.

Tables 4 presents the overall performance of baseline model and each extension. It can be seen that all extension from our proposed method gets better performance than the baseline approach. Table 5 detects the influence of our proposed method on each individual relationship. It can be seen that the proposed methods perform better than the baseline approach over a majority of the relationships. The better performance indicates the following: unannotated scientific papers are useful resource of background information for

[†] https://radimrehurek.com/gensim

**Table 5** Performance (F-score) over selected relationship

| | **Baseline** | **Proposed Method** | | |
|---|---|---|---|---|
| **Relationship** | Baseline | Baseline+TS | Baseline+SRW | Baseline+TS+SRW |
| ATTRIBUTE | 75.09±0.5 | 73.73±0.74 | **75.35**±1.05 | 74.65±0.7 |
| APPLY_TO | 53.08±0.56 | 53.81±1.95 | **55.75**±1.56 | 55.53±1.2 |
| OUTPUT | 49.58±2.49 | 52.06±1.57 | 51.03±1.65 | **52.3**±1.48 |
| INPUT | 38.83±2.54 | 40.56±1.54 | 41.34±1.17 | **43.27**±2.44 |
| EVALUATE | 93.36±1.15 | 92.26±1.18 | 92.87±0.92 | **93.78**±0.54 |
| CONDITION | 38.47±3.92 | 37.54±3.97 | 36.41±3.71 | **39.27**±2.64 |
| EQUIVALENCE | 56.0±2.28 | 56.6±1.85 | 56.4±1.74 | **57.0**±1.1 |
| SUBCONCEPT | 22.47±5.64 | 22.95±2.74 | 24.81±3.39 | **32.4**±4.64 |
| PERFORM | 89.4±0.8 | 89.8±0.98 | 88.6±0.8 | **90.2**±0.75 |
| IN_OUT | 45.96±1.6 | **47.49**±2.0 | 46.82±4.15 | 46.93±1.32 |
| RESULT | 5.34±4.88 | 6.81±4.1 | 9.38±3.26 | **12.14**±4.74 |
| TARGET | 20.54±2.18 | 19.92±2.15 | **20.71**±3.28 | 20.21±1.49 |


R-CONDITION
R-APPLY_TO
REFERENCE
PERSON
PROCESS
MEASURE
PLAN
R-APPLY_TO
L-INPUT
PROCESS
MEASURE
L-EVALUATE
R-PERFORM
DATA-ITEM
In this article, we compare three supervised machine learning methods for combining multiple linguistic indicators
A
L-INPUT
R-APPLY_TO
PROCESS
DATA-ITEM
DATA-ITEM
REFERENCE
DATA-ITEM
QUALITY
contrasting the use of bi- and trigrams whose performance is disappointing,
A

(a) $SRW_A$: *methods, techniques, algorithms systems, models, ...*

(b) $SRW_A$: *bigram, trigram, unigram, tokens, words, ...*

**Fig. 5** Comparison between **Baseline + SRW** and **Baseline**, where red lines indicate the error from **Baseline**, while the green lines show the correctly identified relations from **Baseline + SRW**.

**Table 6** Performance of RE on the setting that **excludes** non-relation

| **Model** | **Precision** | **Recall** | **F-score** |
|---|---|---|---|
| Baseline | 68.88±1.3 | 66.88±0.77 | 66.34±1.23 |
| Baseline + TS | 68.75±0.74 | 67.18±0.79 | 66.78±0.61 |
| Baseline + SRW | 69.1±0.89 | 68.97±0.53 | 68.35±0.59 |
| Baseline + TS + SRW | **70.23**±0.44 | **70.19**±0.48 | **69.6**±0.43 |

RE, and for the two proposed methods, TS and SRW, especially the combination of TS and SRW, which achieved the highest scores, is effective method for extracting background information from unannotated scientific papers for scientific RE. Additionally, all of the proposed methods are unsupervised, and the results also confirm the feasibility of unsupervised method on tapping the potential of unannotated scientific papers for scientific RE.

Figure 5 compares some practical results between **Baseline + SRW** and **Baseline**. Take (b) as an example, although there is the target entity "use", which usually appears in relation APPLY_TO, the proposed system correctly identify the relation as INPUT, because SRW of "trigrams" contains such informative words like "tokens" and "words" that are frequently used as input data for some process.

In addition to comparing the performance over the relations that include non-relation, we also detect the influence of our proposed method when omitting the non-relation. Table 6 and Table 7 present the result on the setting that excludes non-relation. As shown in Table 6 and Table 7, the proposed methods outperform the baseline approach. Again, this comparison indicates the effectiveness of the proposed model for RE in scientific documents.

As mentioned in Sect. 5, we utilize a cosine similarity based criteria, SRW_c, to extract SRW from unannotated scientific papers. In Table 8, we compare the impact of using different SRW_c on the performance of scientific RE. It can be seen that, the best performance on RE is obtained with a moderate SRW_c like 0.35 and 0.45. This is understandable as the high CRW_c might limit the extraction of informative SRW and the low CRW_c might allow the extraction of noisy and irrelevant SRW from scientific papers, this could negatively affect the performance of RE.

**Table 7** Performance (F-score) over selected relationship on the setting that **excludes** non-relation

| | **Baseline** | **Proposed Method** | | |
|---|---|---|---|---|
| **Relationship** | Baseline | Baseline+TS | Baseline+SRW | Baseline+TS+SRW |
| ATTRIBUTE | 80.36±1.33 | 80.55±0.66 | 81.55±0.7 | **81.88**±0.82 |
| APPLY_TO | 74.79±1.42 | 73.95±1.85 | 76.92±1.58 | **77.91**±1.72 |
| OUTPUT | 60.83±3.06 | 62.35±1.31 | 63.46±0.94 | **65.7**±1.43 |
| INPUT | 52.39±1.26 | 54.66±2.56 | 56.5±2.12 | **58.7**±3.01 |
| EVALUATE | 97.67±0.59 | 97.35±0.56 | **98.33**±0.48 | 96.86±0.84 |
| CONDITION | 45.58±4.26 | 46.36±3.8 | 45.1±2.22 | **48.34**±1.51 |
| EQUIVALENCE | 77.8±5.53 | 82.2±0.75 | **85.0**±1.1 | 82.4±1.85 |
| SUBCONCEPT | 44.38±3.51 | 43.56±2.88 | 49.66±4.06 | **51.48**±3.0 |
| PERFORM | **92.4**±1.96 | 90.4±2.06 | 91.2±0.4 | 91.4±0.8 |
| IN_OUT | 47.8±1.51 | **49.69**±2.83 | 48.89±1.69 | 47.78±1.55 |
| RESULT | 42.32±7.97 | 40.09±8.7 | 47.26±4.94 | **58.88**±3.9 |
| TARGET | 23.31±3.98 | 25.06±3.37 | 25.67±2.78 | **27.82**±3.05 |

**Table 8** Impact of using different SRW_c on RE

| **SRW_c** | **Precision** | **Recall** | **F-score** |
|---|---|---|---|
| 0.15 | 65.0±1.53 | 50.26±0.51 | 54.44±0.65 |
| 0.25 | 65.84±1.82 | 49.84±0.61 | 54.67±0.76 |
| 0.35 | 63.02±0.7 | **51.67**±0.52 | **55.56**±0.46 |
| 0.45 | 65.56±0.8 | 50.81±0.5 | 55.37±0.54 |
| 0.55 | **66.3**±1.18 | 50.65±0.74 | 55.22±0.47 |

### 6.3 Error Analysis and Discussion

Towards understanding the disadvantage of our proposed method and improve the performance for future work, we randomly select 5 abstracts from the testing data and manually analyze the types of errors from the result of TS and SRW extension (**Baseline + TS + SRW**), which is visualized like Fig. 7. Based on the difference between the predicted relation and actual relation, we categorize the error into two types. The first type of error occurs between a relationship with high frequency and the one with low frequency, specifically, the model tends to confuse between EMTPY (means non-relation) and predefined relations such as INPUT and ATTRIBUTE, as shown in the third sentence in Fig. 7. This observation is also supported by the confusion matrix in Fig. 6, where this kind of error is marked by a blue rectangle. The second type of error is the error between definitionally similar relationships, which are frequently observed between INPUT and OUTPUT, INPUT and IN_OUT, APPLY_TO and INPUT, ATTRIBUTE and CONDITION etc. as shown in the first sentence of Fig. 7. This observation is also supported by the confusion matrix in Fig. 6, where this kind of error is marked by a red rectangle.

There are several optional solutions for addressing these errors. In order to deal with the non-relation bias, we assume that it would be effective to utilize syntactic information between target entities, because syntactically related entities might tend to be in some relation rather than in non-relation. Therefore by incorporating the syntactic path, the system might decrease the non-relation bias. For overcoming the definitionally similar relationships, we assume that it

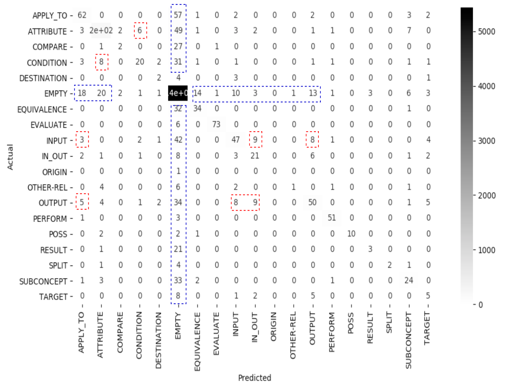
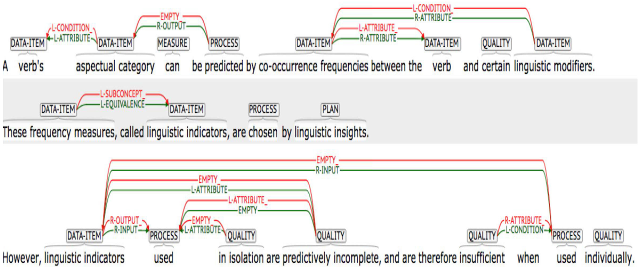


**Fig. 6** Confusion Matrix from **Baseline + TS + SRW**.



A verb's aspectual category can be predicted by co-occurrence frequencies between the verb and certain linguistic modifiers.

These frequency measures, called linguistic indicators, are chosen by linguistic insights.

However, linguistic indicators used in isolation are predictively incomplete, and are therefore insufficient when used individually.

**Fig. 7** Relationship identification error from **Baseline + TS + SRW**, where red lines indicate the error while the green line shows the gold standard relation.

would be effective to extract the information of selectional preference to distinguish these definitionally similar relationships. For instance, for distinguishing between INPUT and APPLY_TO, if one target entity involved in the relation is frequently observed as the OBJECT of the predicate "*apply*" and rarely observed as the OBJECT of "*generate*", the relation might have higher tendency to be in an APPLY_TO than INPUT. This is because the entity, such as "*method*", "*model*" and "*algorithm*", has such selectional preference and usually participates in APPLY_TO relation.

## 7. Conclusion

In this work, we address the task of relationship extraction in scientific documents by leveraging background information extracted from unannotated scientific papers. We design a novel neural network model that not only collects feature from target sentence, but also extracts background information from unannotated scientific papers. We proposed two unsupervised methods: Term Sentence (TS) and Semantically Related Word (SRW). Experimental results on the RANIS corpus demonstrated that unannotated scientific papers could be used as a source of background knowledge for scientific relationship extraction. The proposed unsupervised methods are also proven to be effective for acquiring background information from unannotated scientific papers for relation extraction. An error analysis showed that the proposed model had difficulty for identifying some relationships such as definitionally similar relationships. In our future work, we will improve our model by incorporating other background information, such as syntactic information and selectional preference information.

## Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR1513, Japan and KAKENHI Grant Number 16H06614.

## References

[1] Y. Tateisi, Y. Shidahara, Y. Miyao, and A. Aizawa, "Annotation of computer science papers for semantic relation extraction," LREC, pp.1423–1429, 2014.

[2] S. Teufel, Argumentative zoning: Information extraction from scientific text, Ph.D. thesis, University of Edinburgh, 2000.

[3] K.B. Cohen, D. Demner-Fushman, S. Ananiadou, and J. Tsujii, "Bionlp 2017," BioNLP 2017, 2017.

[4] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum, "Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications," arXiv preprint arXiv:1704.02853, pp.546–555, 2017.

[5] S. Kumar, "A survey of deep learning methods for relation extraction," arXiv preprint arXiv:1705.03645, 2017.

[6] D. Zhou, D. Zhong, and Y. He, "Biomedical relation extraction: from binary to complex," Computational and mathematical methods in medicine, vol.2014, pp.1–18, 2014.

[7] E. Boschee, R. Weischedel, and A. Zamanian, "Automatic information extraction," Proceedings of the International Conference on Intelligence Analysis, Citeseer, 2005.

[8] F.M. Suchanek, G. Ifrim, and G. Weikum, "Combining linguistic and statistical analysis to extract relations from web documents," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.712–717, ACM, 2006.

[9] Y.S. Chan and D. Roth, "Exploiting background knowledge for relation extraction," Proceedings of the 23rd International Conference on Computational Linguistics, pp.152–160, Association for Computational Linguistics, 2010.

[10] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, "Relation classification via convolutional deep neural network," COLING, pp.2335–2344, 2014.

[11] D. Zhang and D. Wang, "Relation classification via recurrent neural network," arXiv preprint arXiv:1508.01006, 2015.

[12] C.N.d. Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," arXiv preprint arXiv:1504.06580, 2015.

[13] W. Ammar, M. Peters, C. Bhagavatula, and R. Power, "The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction," Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp.592–596, 2017.

[14] M. Miwa and M. Bansal, "End-to-end relation extraction using lstms on sequences and tree structures," arXiv preprint arXiv:1601.00770, pp.1105–1116, 2016.

[15] B. Rosario and M.A. Hearst, "Multi-way relation classification: application to protein-protein interactions," Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.732–739, 2005.

[16] J.-D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," BMC bioinformatics, vol.9, no.1, p.10, 2008.

[17] G. Angeli, J. Tibshirani, J. Wu, and C.D. Manning, "Combining distant and partial supervision for relation extraction," EMNLP,

pp.1556–1567, 2014.

[18] Z. GuoDong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics, pp.427–434, 2005.

[19] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[20] J. Gu, F. Sun, L. Qian, and G. Zhou, "Chemical-induced disease relation extraction via convolutional neural network," Database, vol.2017, 2017.

[21] G. Hahn-Powell, D. Bell, M.A. Valenzuela-Escárcega, and M. Surdeanu, "This before that: Causal precedence in the biomedical domain," arXiv preprint arXiv:1606.08089, pp.146–155, 2016.

[22] S. Bird, R. Dale, B.J. Dorr, B.R. Gibson, M.T. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, Y.F. Tan, "The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics," LREC, 2008.

[23] L. Mascarell, "Lexical chains meet word embeddings in document-level statistical machine translation," Proceedings of the Third Workshop on Discourse in Machine Translation, pp.99–109, 2017.

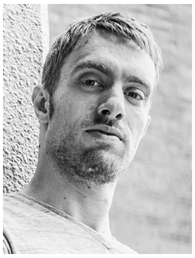
[24] O. Manabu and H. Takeo, "Word sense disambiguation and text segmentation based on lexical cohesion," Proceedings of the 15th conference on Computational linguistics-Volume 2, Association for Computational Linguistics, pp.755–761, 1994.

[25] T. Wang, Y. Li, K. Bontcheva, H. Cunningham, and J. Wang, "Automatic extraction of hierarchical relations from text," European Semantic Web Conference, Springer, vol.4011, pp.215–229, 2006.

[26] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A cpu and gpu math compiler in python," Proc. 9th Python in Science Conf, pp.1–7, 2010.

**Qin Dai** received his B.A. in Linguistics and Literature form Beijing Normal University, China in 2009, and his M.S. in in System Information Sciences from Tohoku University in Sendai, Japan in 2014. He is a graduate student in the Department of System Information Sciences, Tohoku University. His research interests include machine reading, knowledge acquisition and natural language processing.

**Naoya Inoue** received his M.S. degree of engineering from Nara Institute of Science and Technology in 2010 and his Ph.D. degree in Information Science from Tohoku University in 2013. He joined DENSO Corporation as a researcher in 2013. He has been an assistant professor at Tohoku University since 2015. He has also been a guest researcher at RIKEN Center for Advanced Intelligence Project since 2018. His research interests are in inference-based discourse processing and language grounding problems.



**Paul Reisert** received his B.S. in Computer Science from Purdue University in West Lafayette, Indiana in 2010 and his M.S. and Ph.D. degrees in System Information Sciences from Tohoku University in Sendai, Japan in 2017. He is currently a post-doctoral researcher at RIKEN, Japan. His research interests include argumentation mining, discourse analysis, and natural language processing.

**Kentaro Inui** is a professor of Graduate School of Information Sciences at Tohoku University and the the team leader of the Natural Language Understanding Team at AIP, RIKEN. He received his doctorate degree of engineering from Tokyo Institute of Technology in 1995 and has been working in the field of natural language processing. His research interests include semantic and discourse analysis of natural language text, knowledge acquisition and commonsense reasoning, and a broad variety of language technology applications. He served as an editorial board of Computational Linguistics for three years from 2009 and currently serves as an director of ANLP and the vice editor-in-chief for the Journal of Natural Language Processing.