Extraction of Gene-Environment Interaction from the Biomedical Literature

Jinseon You Jin-Woo Chung Wonsuk Yang Jong C. Park*
School of Computing
Korea Advanced Institute of Science and Technology
{jsyou, jwchung, derrick0511, park}@nlp.kaist.ac.kr

Abstract

Genetic information in the literature has been extensively looked into for the purpose of discovering the etiology of a disease. As the gene-disease relation is sensitive to external factors, their identification is important to study a disease. Environmental influences, which are usually called Gene-Environment interaction (GxE), have been considered as important factors and have extensively been researched in biology. Nevertheless, there is still a lack of systems for automatic GxE extraction from the biomedical literature due to new challenges: (1) there are no preprocessing tools and corpora for GxE, (2) expressions of GxE are often quite implicit, and (3) document-level comprehension is usually required. We propose to overcome these challenges with neural network models and show that a modified sequence-to-sequence model with a static RNN decoder produces a good performance in GxE recognition.[1]

1 Introduction

Identifying genetic information related to a disease is an effective method for discovering the etiology of the disease. Many researchers in biology have attempted to identify the relationship between different types of genetic information, such as genes, gene mutations or other biological events, and diseases.

One of the difficult aspects in the research is that it is necessary to consider various external factors, because they can affect whether such biological relationships hold or not. For example, it has been shown that there is no association of NAT2 gene and breast cancer (Zgheib et al., 2013), but after three years, other researchers made a conflicting claim that NAT2 gene is associated with breast cancer (Kasajova et al., 2016). There may be many factors causing this difference, but investigating the environmental factors has been one of the important research topics. Kasajova et al. (2016) found that NAT2 gene is associated with breast cancer when women with NAT2 gene polymorphisms have been exposed to long-period active smoking. As a result, active smoking has been considered as a crucial factor that determines the relation between NAT2 gene and breast cancer, which biologists called gene and environment interaction (GxE).

Since the importance of studying GxE is recognized, the amount of related work has steadily been increasing (Hunter, 2005). Nonetheless, there is still a lack of systems and databases that deal with this information (Simonds et al., 2016). For the purpose of addressing this situation, we present an automatic system that extracts environment terms indicating a change of gene-disease relations from the biomedical literature.

There are three major challenges that make it difficult to perform GxE recognition using existing systems in the biomedical domain. First, in contrast to general biomedical natural language processing (BioNLP) tasks, there are no preprocessing tools and corpora for GxE, though there are some resources for chemical-disease relations, such as named entity recognition systems specialized for chemical and disease names and corpora that annotate chemical and disease relations in abstracts (Wei et al., 2015b). Second, research on discovering biomedical relations usually specifies environmental information in the literature in various ways, using not only expressions that explicitly refer to certain biomedical concepts such as

*Corresponding author

[1]Our source code and gold standard corpus are available at http://biopathway.org/GxE

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 865–874,
Taipei, Taiwan, November 27 – December 1, 2017

pregnancy and smoking, but also statistical terms that refer to a comparison between two control groups, such as p-value and odds ratio, which are quite difficult to capture using conventional tools. Since the literature for GxE also tends to report experiment results in similar ways, the system needs to understand such implicit information to determine whether the result is meaningful or not. Third, information of this kind indicating GxE is usually not reported in a single sentence, requiring document-level comprehension of text.

To address this situation, we build an annotated corpus for GxE and develop an end-to-end system that recognizes environment information for gene-disease pairs given in single document. We exploit high-dimensional models based on neural networks to enable document-level understanding of text and to deal with the issues above. We also perform experiments with different neural network models to investigate which models are most suitable to GxE recognition.

2 Related work

One of the related areas that have been actively researched in BioNLP is biomedical event extraction. For example, a system was proposed in the recent shared tasks (Kim et al., 2013), attempting to extract ten biological events, with the best performance under a 0.6 F1-score. This score, however, was extremely skewed to particular event types. While all the systems showed good performance, with nearly a 0.8 F1-score, in extracting simple events such as gene expression and transcription, they showed quite poor performance for complex events such as binding and regulation. This is because, in contrast to simple events, complex events consist of more than two elements or another event. In particular, the task of extracting binding events is usually treated as finding ternary relations, where a system is supposed to recognize two biological entities together with a particular site where their binding takes place. This task is similar to our task as the relation between two entities can be changed according to the third entity. The best system for binding event extraction achieved a 0.49 F1-score.

Another recent work on dealing with complex relations in BioNLP is reading comprehension (RC), where the system is to find proper answers to given questions about a biological process within single document. For example, the system in (Berant et al., 2014) attempts to find answers through comparisons between two graphs constructed from given documents and questions. Although the system explicitly combines biological events extracted from sentences to construct a long biological process, possibly leading to the propagation of errors, they reported fairly good performance and meaningful results, considering that it is the first attempt for document-level biological information extraction.

On the other hand, there are quite a few systems and corpora for document-level comprehension from a similar perspective in other domains, such as news articles (Hermann et al., 2015) and children's books (Hill et al., 2016). One of the recent studies, (Hermann et al., 2015), addresses the reading comprehension task for which proposed models infer missing entities. From the perspective of evaluating how well such models understand documents for answering given questions, the task is similar to the present work. In this task, neural network models were shown to be effective for processing document-level information. More specifically, they demonstrated that a variant of neural network, RNN with attention mechanism, achieved the state-of-the-art performance (Chen et al., 2016).

WikiReading is most similar to our task as it deals with inference over entities based on a sequence of tokens (Hewlett et al., 2016). The authors were inspired from the QA task, treating given properties as questions and developing models to find proper entities that could be answers to the questions. There are two types of properties in WikiReading: (1) the categorical property that requires selection among a relatively small number of possible answers and (2) the relational property that requires extraction of rare or unique answers from the document. The authors compared various types of models from simple word embedding models to sequence-to-sequence models and showed that the sequence-to-sequence model gives rise to outstanding performance in both types of properties.

3 Task definition

3.1 GxE extraction

We formulate the GxE recognition task as extracting terms indicating a particular environment that is involved in a change of gene-disease relations. Figure 1 illustrates an example abstract that con-

Melatonin pathway genes and breast cancer risk among Chinese women. [PMID: 22138747]

Previous studies suggest that melatonin may act on cancer growth through a variety of mechanisms, most notably by direct anti-proliferative effects on breast cancer cells and via interactions with the estrogen pathway. Three genes are largely responsible for mediating the downstream effects of melatonin: melatonin receptors 1a and 1b (MTNR1a and MTNR1b), and arylalkylamine N-acetyltransferase (AANAT). It is hypothesized that genetic variation in these genes may lead to altered protein production or function. To address this question, we conducted a comprehensive evaluation of the association between common single nucleotide polymorphisms (SNPs) in the MTNR1a, MTNR1b, and AANAT genes and breast cancer risk among 2,073 cases and 2,083 controls, using a two-stage analysis of genome-wide association data among women of the Shanghai Breast Cancer Study. Results demonstrate two SNPs were consistently associated with breast cancer risk across both study stages. Compared with MTNR1b rs10765576 major allele carriers (GG or GA), a decreased risk of breast cancer was associated with the AA genotype (OR = 0.78, 95% CI = 0.62-0.97, P = 0.0281). Although no overall association was seen in the combined analysis, the effect of MTNR1a rs7665392 was found to vary by **menopausal status** (P-value for interaction = 0.001). **Premenopausal women** with the GG genotype were at increased risk for breast cancer compared with major allele carriers (TT or TG) (OR = 1.57, 95% CI = 1.07-2.31, P = 0.020), while **postmenopausal women** were at decreased risk (OR = 0.58, 95% 0.36-0.95, P = 0.030). No significant breast cancer associations were found for variants in the AANAT gene. These results suggest that common genetic variation in the MTNR1a and 1b genes may contribute to breast cancer susceptibility, and that associations may vary by **menopausal status**. Given that multiple variants in high linkage disequilibrium with MTNR1b rs76653292 have been associated with altered function or expression of insulin and glucose family members, further research may focus on clarifying this relationship.

Figure 1: An illustrated abstract describing GxE for lung cancer. In this figure, gene and disease are shown in blue and red, respectively. Expressions in bold-face are targeted environment terms. Sentences highlighted in gray are the evidence supporting the claim that an association between MTNR1a and breast cancer is changed by menopausal status.

Environment type	Example
Energy balance	dietary factors, physical activity
Lifestyle	smoking, alcohol, breastfeeding
Exogenous hormones	HRT, OC use
Endogenous hormones	menopausal status, age of menarche
Chemical environment	grilled foods/meats, heterocyclic amines
Drugs/treatment	statin, NSAIDS
Infection and inflammation	helicobacter pylori, autoimmune disease
Physical environment	x-rays, sun exposure

Table 1: The list of biological environment types (Simonds et al., 2016)

tains information about GxE for breast cancer, where gene and disease names are shown in blue and red, respectively (Wei et al., 2015a; Lee et al., 2016). There are four types of genes (MTNR1a, MTNR1b, AANAT, insulin) and two types of diseases (breast cancer, cancer). Therefore, we consider twelve gene-disease combinations for which our system attempts to find environment terms from the abstract. As an example of environment terms, it is claimed in the abstract that the association between MTNR1a and 1b genes and breast cancer may vary to *menopausal status*. Sentences highlighted in gray are the evidence supporting the claim. Expressions in bold-face are targeted environment terms to be extracted by our system.

Our model is given the abstract marked with genes and diseases, and considers each unique gene-disease combination, one at a time, to find its environment terms. For example, if we want to consider the combination of MTNR1a and breast cancer, the input is the abstract marked only with these two entities, without other entities marked such as MTNR1b or AANAT. We used two state-of-the-art named entity recognizers (Wei et al., 2015a; Lee et al., 2016) to identify gene and disease names from a given abstract. For each combination, we consider the following four cases: (1) the combination consists of an unassociated gene-disease pair, not affected by an environment; (2) although the combination consists of a pair that is basically unassociated, it becomes associated due to a particular environment; (3) the combination consists of an associated pair but it is not affected by an environment; and (4) the combination consists of an associated pair and the degree of its association is changed by an environment. Our system is trained to choose the most proper environment term for a given combination in cases (2) and (4), but not to choose any term in cases (1) and (3).

3.2 Corpus

For experimental data, we collected 253 raw abstracts that are taken from review papers on GxE for diverse diseases (Simonds et al., 2016; Dunn et al., 2011; DiGangi et al., 2013; Iyegbe et al., 2014; Hunter, 2005). To establish the gold stan-

dard data to train and test the system, we manually annotated the biological environments that should be extracted. For the clear definition of an environment, we only annotated the terms that can be categorized into one of the types in Table 1 and that are clearly reported as associated with gene-disease relations in the abstract.

Annotation was conducted by two experts in bioinformatics, who were given abstracts marked with gene and disease names. They did not annotate combinations consisting of entities that are incorrectly recognized by the named entity recognizer (i.e., entities that are neither gene nor disease). For each abstract, one annotator read the entire body of its text and annotated terms referring to an environment that is involved in a given combination of gene and disease, and then another annotator validated its correctness, in a way similar to other annotation tasks in BioNLP (e.g., Berant et al. 2014). The agreement on annotated environment terms between the two annotators is 0.81. If they did not agree on a certain annotation, they had a discussion on the disagreement and resolved it afterwards. The corpus contains a total of 1,429 combinations of unique genes and diseases. Among them, 341 combinations are annotated as being affected by environment, i.e., they are linked to some environment terms annotated in the same abstract.

4 Method

We use two types of models, a feature-based model and an neural-based model, that could be applied to document-level understanding of relations between entities in order to investigate which models are suitable to GxE recognition and whether or not there are important issues particularly in this new task. There are three variants based on the neural-based model: (1) an attentive reader (Hermann et al., 2015), (2) a sequence-to-sequence model (Sutskever et al., 2014), and (3) a static RNN decoder. We envision that different characteristics of these models would lead to different performance, according to the types of task.

In our experiment, the three models relied neither on any prior knowledge nor on external tools for collecting candidate environment terms. Even though such words as 'smoking' or 'alcohol' can be considered to have a higher probability to be a biological environment than other words, we did not use such information to prevent error propagation and to investigate the possibility of handling newly introduced terms.

4.1 A feature-based model

We combined two models proposed by Chen et al. (2016) and Xu et al. (2016): a model that adapts an entity-centric approach to the RC task, and a feature-based model that extracts chemical-disease relations on a document level.

Inspired by these two models, we use the following feature sets that we expect are suitable to our task. We describe each feature in detail below, where *g*, *d*, and *e* indicate gene terms, disease terms, and candidate environment terms, respectively: (1) shortest distance from *e* to *g* and *d* in the abstract, (2) whether *e* and *g* pair in the same sentence, (3) whether *e* and *d* pair in the same sentence, (4) whether *e*, *d*, and *g* pair in the same sentence, (5) whether *e* is included in MeSH (Medical Subject Headings) terms, (6) the frequency of *e* in an abstract, (7) the frequency of *e* in all abstracts, (8) whether *e* and *g* are connected by the dependency parser (De Marneffe et al., 2006), and (9) whether *e* and *d* are connected by the dependency parser (De Marneffe et al., 2006).

Using these features, the model tried to classify all terms that are present in the abstract. If the model assigns 1 to a term, we regard it as an environment. If the model classifies all terms for a particular gene-disease combination as 0, we assume that there is no environment for this combination in the abstract.

4.2 An neural-based model

We propose three neural-based models; 1) an attentive reader, 2) a sequence-to-sequence model, and 3) a static RNN decoder. The three models comprise two parts: converting text to vector representation, called encoding, and predicting the vector to answer, called decoding. The encoding is the same in all the three models. We look over the encoding and then compare each decoding part of the three models.

4.2.1 Encoding

Our encoding with attention is based on the model proposed by Chen et al. (2016), which shows better performance than any other encoders. The model runs in two steps, **text encoding** and **attention**, described in detail as follows.

Text encoding: All words are mapped to d-dimensional vectors using the PubMed/PMC word

embedding model (Pyysalo et al., 2013) with a limited dictionary size (V). We include special tokens, '<NOE>', that stands for no environment terms for the combination and '<UNK>', that stands for terms that are not included in the dictionary. The sequence of words in an abstract excluding stop words and special characters is encoded as $\mathbf{p}_1, ..., \mathbf{p}_m \in \mathbb{R}^d$ where m is the number of words in the abstract. Then, we pass the sequence $\mathbf{p}_1, ..., \mathbf{p}_m$ to bi-directional RNN:

$$\overrightarrow{\mathbf{h}_i} = RNN(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{p}_i) \in \mathbb{R}^h, i = 1, ..., m$$

$$\overleftarrow{\mathbf{h}_i} = RNN(\overleftarrow{\mathbf{h}}_{i+1}, \mathbf{p}_i) \in \mathbb{R}^h, i = m, ..., 1$$

$$\tilde{\mathbf{p}}_i = concat(\overrightarrow{\mathbf{h}_i}, \overleftarrow{\mathbf{h}_i}) = \begin{bmatrix} \overrightarrow{\mathbf{h}_i} \\ \overleftarrow{\mathbf{h}_i} \end{bmatrix} \in \mathbb{R}^{2h}, i = 1, ..., m$$

where h is the dimension of hidden units of RNN.

From $\mathbf{p}_1, ..., \mathbf{p}_m$, the model extracts marked gene and disease names. Let the set of gene names be $\{\mathbf{g}_1, ...\mathbf{g}_n\}$ where n is the number of gene names in the abstract. Let the set of disease names be $\{\mathbf{d}_1, ...\mathbf{d}_l\}$ where l is the number of disease names in the abstract. Then, we make the gene-disease combination vector $\mathbf{c}$ by element-wise summation of concatenated vectors:

$$\mathbf{c} = \mathbf{W}_c^T(\begin{bmatrix} \mathbf{g}_1 \\ \mathbf{d}_1 \end{bmatrix} \oplus \begin{bmatrix} \mathbf{g}_2 \\ \mathbf{d}_1 \end{bmatrix} ... \oplus \begin{bmatrix} \mathbf{g}_n \\ \mathbf{d}_l \end{bmatrix})$$

where $\mathbf{W}_c \in \mathbb{R}^{2d \times 2h}$ is the weight vector for gene and disease.

Attention: In order to enable the model to focus more on evidence for identifying environment terms in the abstract, we used the attention mechanism. In the QA task, the vector of questions is projected to a document for calculating the probability of relevance degree between a question and a document. Likewise, we project the gene-disease combination vector ($\mathbf{c}$) to the sequence of word vectors ($\tilde{\mathbf{p}}_1, ..., \tilde{\mathbf{p}}_m$). We applied a bilinear term, a variant of attention mechanism, to combine the combination vector and the sequence of vectors:

$$\mathbf{a} = softmax(\mathbf{c}^T \mathbf{W}_b \tilde{\mathbf{p}}_i), i = 1, ..., m$$

where $\mathbf{W}_b \in \mathbb{R}^{2h \times 2h}$.

And then, we generated an attention vector by summation of projecting the bilinear term to the sequence of vectors:

$$\tilde{\mathbf{a}} = \sum_i \mathbf{a}\tilde{\mathbf{p}}_\mathbf{i}, i = 1, ..., m$$

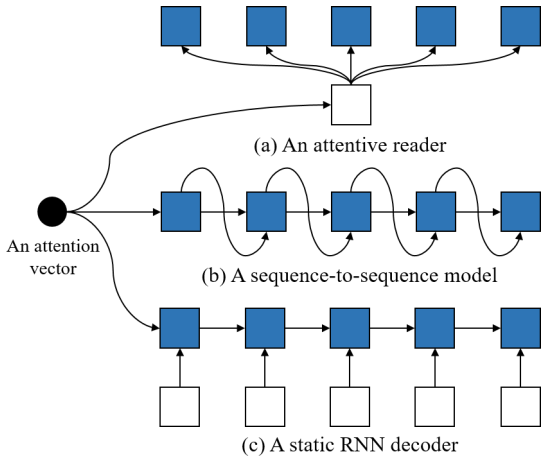

Figure 2: An overview of each decoding part in three models

4.2.2 Decoding

In this section, we describe each decoding of the three models. Figure 2 illustrates an overview of three models.

(a) An attentive reader

By mapping the attention vector ($\tilde{\mathbf{a}}$) to vocabulary, we compute output vector ($\mathbf{o}_a$) as follows,

$$\mathbf{o}_a = \mathbf{W}_a^T \tilde{\mathbf{a}}$$

where $\mathbf{W}_a \in \mathbb{R}^{2h \times V}$:

We choose terms that come from their conjunction showing the top values of the output vector and that are represented in the abstract, and consider them as environment. However, if the top value of the output vector indicates '<NOE>', we conclude that there is no environment.

(b) A sequence-to-sequence model

A decoder in the sequence-to-sequence model dynamically generates tokens from '<SOE>' (start of token) to '<EOE>' (end of token). The model is based on a previous hidden vector, a previous token vector and an encoding vector that is an output vector of the encoding. The previous token vector is computed by projecting a token generated in previous time step to an embedding layer. We try to set the attention vector ($\tilde{\mathbf{a}}$) to the encoding vector as we expect that the attention vector is more properly tuned to extract terms depending on the gene-disease combinations than the original encoding vector:

$$\mathbf{t}_{i-1} = \mathbf{W}_e^T o_{i-1}$$

$$\mathbf{y}_i = RNN(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{t}_{i-1}, \tilde{\mathbf{a}}) \in \mathbb{R}^{2h}, i = 1, ..., e$$

where e is the number of environment terms and $\mathbf{W}_e^T$ is an embedding layer.

$$o_i = argmax(\mathbf{W}_s^T \mathbf{y}_i), i = 1, ..., e$$

where $\mathbf{W}_s^T \in \mathbb{R}^{2h \times V}$. The o_i is the index of the vocabulary and a sequence of tokens, $(o_{1,...,e})$, is regarded as environment terms predicted by the model.

If the first decoding token indicates '<NOE>' in the output sequence, we assume that there is no environment.

(c) A static RNN decoder

As a modification to the sequence-to-sequence model, we suggest that the model uses a static RNN decoder, which does not use a previous token vector ($\mathbf{t}_{i-1}$). In particular, the model used randomly normalized token vectors. Because the model is needed to set the length of the decoder in advance, it seems to statically generate environment terms, which is an outstanding feature in comparison to the sequence-to-sequence model. Because our answer tokens are usually atomic and spread over the abstract, the previous output state, which is usually used when making a long sequence of tokens, is not useful for our task.

$$\mathbf{y}_i = RNN(\overrightarrow{\mathbf{h}}_{i-1}, \mathbf{t'}_i, \tilde{\mathbf{a}}) \in \mathbb{R}^{2h}, i = 1, ..., e'$$

where e' is the number of environment tokens that is set in advance.

$$o_i = argmax(\mathbf{W}_r^T \mathbf{y}_i), i = 1, ..., e'$$

where $\mathbf{W}_r^T \in \mathbb{R}^{2h \times V}$. The o_i is the index of the vocabulary and a sequence of tokens, $(o_{1,...,e'})$, is regarded as environment terms predicted by the model.

Similar to the sequence-to-sequence model, if the first decoding token indicates '<NOE>' in the output sequence, we assume that there is no environment.

5 Experiments

5.1 Corpus statistics

In the 253 abstracts that report the presence of GxE, 341 out of 1429 gene-disease combinations show a relationship and are considered affected by environment. Table 2 provides some statistics of the dataset. There are a total of 247 types of gene and 106 types of disease. In an abstract, there are about 304 tokens and 10 lines on average. Also, the average number of environment tokens is about 3 and the maximum number is 15. Assuming that the combination shows GxE in an abstract, the average number of unique environments per combination is 1.4. From the statistics, we see that the environment is made of just one or two words and that the combinations showing GxE appear rarely.

Category	**#**
Types of genes	267
Types of diseases	106
Avg. # of tokens	304.1
Avg. # of sentences	10.4
Avg. # of environment tokens	2.7
Min. # of environment tokens	1
Max. # of environment tokens	15
Avg. # of environments per combination	1.4
Min. # of environments per combination	1
Max. # of environments per combination	8

Table 2: Data statistics of the GxE dataset. All values are based on the statistics from the entire dataset.

In order to balance positive and negative values, we randomly sampled 146 combinations from almost 1000 redundant combinations that do not show GxE, and abstracts with a total of 487 combinations were given as input to the system. The input is already marked with gene and disease, and the annotated gold standard environment term was used as the target answer. We randomly selected 80% of the dataset (389) and used them for training, 10% for validation (49), and 10% for test (49).

5.2 Setup

For training the proposed model, we set common parameters empirically as follows. According to a given embedding model, the dimension of word vectors is 200. We built a dictionary using the most frequent $2.5K$ words. And we split sentences using the tool (Kazama and Tsujii, 2003) and tokenized the sentences using the supporting tool in BioNLP Shared Task 2011[2], where both tools are specialized to BioNLP.

[2] https://github.com/ninjin/bionlp_st_2011_supporting/blob/master/tools/GTB-tokenize.pl

Model	P	R	F1
Baseline model (DT)	0.157	0.172	0.152
Baseline model (SVM)	0.279	0.274	0.275
Baseline model (RF)	0.204	0.196	0.196
Baseline model (GB)	0.1	0.123	0.095
Baseline model (AB)	0.168	0.155	0.153
RNN reader (top-5)	0.321	0.359	0.338
Attentive reader (top-5)	0.362	0.373	0.366
Attentive reader (top-10)	0.290	0.542	0.378
Attentive reader (top-15)	0.283	0.639	0.390
Attentive reader (top-20)	0.214	**0.670**	0.324
Seq2seq model (basic encoding)	0.305	0.298	0.301
Seq2seq model (attention encoding)	0.322	0.319	0.320
Static RNN decoder (basic encoding)	**0.484**	0.389	**0.426**
Static RNN decoder (attention encoding)	0.450	0.380	0.409

Table 3: The performances of different models on GxE recognition. P and R stand for precision and recall, respectively. DT, SVM, RF, GB and AB stand for Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting, and AdaBoost, respectively. The RNN reader indicates attentive reader without attention encoding.

The baseline models followed the initial parameter setting of a machine learning framework, *sklearn*[3]. We tried to change the parameter setting, without any significant difference in performance.

In the attentive reader and the static RNN decoder, we used LSTM (Hochreiter and Schmidhuber, 1997), a variant of the RNN model, and set the hidden size and dropout rate of RNN to 64 and 0.5, respectively. On the other hand, the sequence-to-sequence model used GRU (Cho et al., 2014), another variant of the RNN model, and we set the hidden size of 64 and dropout rate of 0.5. In the case of the static RNN decoder, it is necessary to set the length of the model due to a static attribute. Therefore, we evaluated the performance of the model according to the length, and we found that the model with a length of 25 performs best.

All weights of three models are initialized from Gaussian distribution with 0 mean and 0.01 STD. At each update, we randomly sampled a mini-batch of 16, and the attentive reader, sequence-to-sequence model, and static RNN decoder used the Adam algorithm (Kingma and Ba, 2015) with 0.0001, 0.001, and 0.01 learning rates for optimization, respectively. Except for the attentive reader, we additionally used l2 regularizations. And we clipped the gradients when the norm of the gradients exceeds 10. We ran all neural network models up to 100 epochs.

We implemented the proposed neural network models using *TensorFlow*[4].

5.3 Results

The overall performance of our proposed models is shown in Table 3. We ran each model 10 times independently, and reported average scores in the table. Among others, it shows that it is hard to detect environment terms with feature-based models and that it is necessary to use high-dimensional models such as deep neural network. It also shows that the static RNN decoders outperform other models. It is an interesting result because, contrary to our result, the sequence-to-sequence model showed outstanding performance in WikiReading, which is the most similar task to ours. The fact that the static RNN decoder shows best performance is probably due to the characteristics of our corpus where environment tokens are more widespread and atomic.

If the reader model and the sequence-to-sequence model use attention as demonstrated in other studies, it shows higher performance than the model without attention. On the other hand, the static RNN decoders show a different case, in that our proposed attention seems to hamper the model in exactly extracting the environment.

In order to monitor how performance varies to the choice of top-k values, we evaluated the attention model with different top-k values ($k = 5, 10, 15, 20$). When we increase k values, recall increases and precision decreases. Overall, precision of the static RNN decoder is found better than that of the other models. On the other hand, the attentive reader model shows outstanding performance in extracting all relevant environment tokens.

5.4 Analysis

The baseline models mainly show F1-scores under 30, which are worse than we expected. Among them, SVM outputs skewed results, failing to find any environment terms, and classifying all test data to '<NOE>'. As a result, its performance depends on the number of input data showing no environment terms, which explains why the model shows even precision and recall. In contrast to SVM, the other two models can detect environ-

[3] http://scikit-learn.org

[4] https://www.tensorflow.org/

Polymorphisms in CRHR1 and the serotonin transporter loci: gene x gene x environment interactions on depressive symptoms. [PMID: 20029939] ... These data suggest that G x E interactions predictive of depressive symptoms may be differentially sensitive to **levels of childhood trauma**, ...		
attentive reader	**seq2seq**	**static RNN decoder**
high history low status <UNK> levels current hormone stress ...	<UNK>	<UNK> childhood levels high trauma
Peroxisome proliferator-activated receptor-alpha (PPARA) genetic polymorphisms and breast cancer risk: a Long Island ancillary study. [PMID: 18586686] ... but there was some evidence of interaction between PPARA variants and **aspirin use**, defined as **use at least once per week for 6 months or longer**. ...		
attentive reader	**seq2seq**	**static RNN decoder**
use aspirin <UNK> women levels body mass cancer index ...	hcas	aspirin

Table 4: Results of experiment with four proposed models. Words in red and blue indicate disease and gene names, respectively, and the words in bold-face indicate environment terms.

ment terms, but they also identify irrelevant tokens as environment terms in most cases.

In order to analyze how differently the proposed models output, we select three proposed models showing the best performance, the attentive reader, the sequence-to-sequence model (attention encoding), and the static RNN decoder (basic encoding) among them, and show the result of experiments for two abstracts. In Table 4, the first row represents a partial content of the abstract and the second row represents the results. We show a sequence of tokens in the attentive reader as much as possible, and the sequence is ordered by scores. On the other hand, a sequence of tokens in the sequence-to-sequence model and static RNN decoder is ordered in which they were made in the decoding part.

The first example in Table 4 provides GxE for two genes, where we ask our models to extract environment tokens for *serotonin transport* and *depressive symptoms*. Although the three models failed to identify all answer tokens, the static RNN decoder shows better performance than the others. In tokens extracted by the attentive reader, there are not only answer tokens but also error tokens, which results in decreasing the performance. These weaknesses sometimes work as an advantage for the cases with many environment terms as shown in the abstract in the second example.

In the second example, the number of tokens is bigger than that in the first example. Interestingly, the performance of the attentive reader and that of the static RNN decoder are reversed. Although there are many answer tokens, the static RNN decoder seems to extract minimal tokens. On the other hand, the majority of tokens extracted by the attentive reader are included in answer tokens.

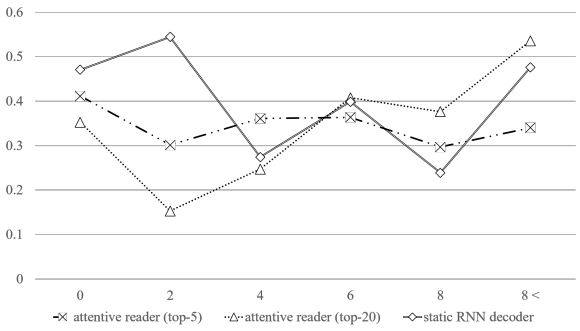

Figure 3: Performance of the two models, the attentive reader and the static RNN decoder model, according to the number of environment tokens. X-axis and Y-axis represent the number and F1 score, respectively.

As shown in Table 4, the static RNN decoder models work better in extracting a small number of tokens. On the other hand, the attentive reader is suitable to the data with a large number of tokens. And the choice of k seems an important factor to affect the performance of the attentive reader. In order to see that the observations are common, we compared F1-scores of three models, the attentive reader (top-5 and top-20) and the static RNN decoder, according to the number of environment tokens.

Figure 3 presents the change of F1-score when the number of environment tokens varies. When the number is smaller than 2, the static RNN decoder works best. While the static RNN decoder is sensitive to the number, the attentive reader (top-5) shows stable performance. Therefore, the attentive reader (top-5) shows better performance than the static RNN reader at both 4 and 8 points. The per-

Serious obstetric complications interact with hypoxia-regulated/vascular-expression genes to influence schizophrenia risk. [PMID: 18195713]	
basic encoding	**attention encoding**
<NOE>	obstetric serious complications

Table 5: An example showing the benefits of using attention model: the words in bold-face indicate environment terms.

formance of the attentive reader (top-20), however, increases steadily according to the number. As a result, the attentive reader (top-20) works best at 6 points afterwards.

At 6 points, the performance of the static RNN decoder and that of attentive reader (top-20) are reversed, so it seems that there is not much performance difference among them in the graph. But, the overall difference in performance is much bigger because the average of the numbers in the corpus is nearly 5. As a result, the static RNN decoder outperforms other models. From this observation, we anticipate that if we address the GxE task focusing on the number by combining the two models, the performance will exceed the current best score, 0.426.

In the static RNN decoder, the attention encoding did not seem to work well, which is in contrast to our assumption that it would be better to consider combinations for our task. However, there is a special case showing the benefits of attention encoding as shown in Table 5. Table 5 shows part of an abstract where there are 39 combinations (13 genes and 3 diseases) and four genes associated with schizophrenia among them have the same environment terms (*serious obstetric complications*). Unless considering the combination, it seems hard to identify environment terms due to many negative examples. As a result, given a combination, (*RGS4* and *fetal hypoxia*), without an environment term such as the example in Table 5, the static RNN decoder using a basic encoding seems to regard the majority of combinations including the combination with environment as a negative example, as shown in the first column of Table 5. However, interestingly, the static RNN decoder using attention encoding identified three tokens that are correct environment terms when the four combinations with environment are given. This incorrect result may be due to the lack of training examples like this case. So, if we are given many cases as shown in Table 5, we envision that attention encoding will help improve performance.

6 Conclusion

In this paper, we proposed various methods for GxE recognition and showed that our models achieved good performance, despite the inherent difficulty of the task. Unlike general approaches, such as CNN, RNN, and attention mechanism, in order to extract targeted relations or infer the correct answer, we used an RNN decoder as a sequence-to-sequence model with a static decoder, and demonstrated that it is suitable to the task in extracting terms from documents. It is necessary to identify conditional information that creates the contradiction of gene-disease relations to develop advanced systems for understanding the full etiology of a disease or the full genetic network. We anticipate that the model will help researchers not only to identify correct gene-disease relations but also to apply them to other tasks, such as extracting location information that indicates where events occur.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2017R1A2B4012788).

References

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D Manning. 2014. Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1510.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, volume 6, pages 449–454.

Julia DiGangi, Guia Guffanti, Katie A McLaughlin, and Karestan C Koenen. 2013. Considering trauma exposure in the context of genetics studies of posttraumatic stress disorder: a systematic review. *Biology of mood & anxiety disorders*, 3(1):2.

Erin C Dunn, Monica Uddin, SV Subramanian, Jordan W Smoller, Sandro Galea, and Karestan C Koenen. 2011. Research Review: Gene-environment interaction research in youth depression-a systematic review with recommendations for future research. *Journal of Child Psychology and Psychiatry*, 52(12):1223–1238.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems 28*, pages 1693–1701.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A Novel Large-scale Language Understanding Task over wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1545.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *International Conference on Learning Representations 2016*.

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

David J Hunter. 2005. Gene-environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298.

Conrad Iyegbe, Desmond Campbell, Amy Butler, Olesya Ajnakina, and Pak Sham. 2014. The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for gxe research. *Social psychiatry and psychiatric epidemiology*, 49(2):169–182.

Petra Kasajova, Veronika Holubekova, Andrea Mendelova, Zora Lasabova, Pavol Zubor, Erik Kudela, Kristina Biskupska-Bodova, and Jan Danko. 2016. Active cigarette smoking and the risk of breast cancer at the level of N-acetyltransferase 2 (NAT2) gene polymorphisms. *Tumor Biology*, 37(6):7929.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, pages 137–144.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013. The Genia Event Extraction Shared Task, 2013 Edition-Overview. In *Proceedings of the 3rd BioNLP Shared Task Workshop*, pages 8–15.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations 2015*.

Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2016. AuDis: an automatic CRF-enhanced disease normalization in biomedical text. *Database*, 2016:baw091.

S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. 2013. Distributional Semantics Resources for Biomedical Text Processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, pages 39–44.

Naoko I Simonds, Armen A Ghazarian, Camilla B Pimentel, Sheri D Schully, Gary L Ellison, Elizabeth M Gillanders, and Leah E Mechanic. 2016. Review of the Gene-Environment Interaction Literature in Cancer: What Do We Know? *Genetic epidemiology*, 40(5):356–365.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015a. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed research international*, 2015.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015b. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 154–166. Sevilla Spain.

Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Hee-Jin Lee, and Hua Xu. 2016. CD-REST: a system for extracting chemical-induced disease relation in literature. *Database*, 2016:baw036.

Nathalie K Zgheib, Ashraf A Shamseddine, Eddy Geryess, Arafat Tfayli, Ali Bazarbachi, Ziad Salem, Ali Shamseddine, Ali Taher, and Nagi S El-Saghir. 2013. Genetic polymorphisms of CYP2E1, GST, and NAT2 enzymes are not associated with risk of breast cancer in a sample of Lebanese women. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 747:40–47.