An Insight Extraction System on BioMedical Literature with Deep Neural Networks

Hua He\textsuperscript{1}, Kris Ganjam\textsuperscript{2}, Navendu Jain\textsuperscript{2}, Jessica Lundin\textsuperscript{2}, Ryen White\textsuperscript{2} and Jimmy Lin\textsuperscript{3}

\textsuperscript{1} Department of Computer Science, University of Maryland College Park
huah@cs.umd.edu
\textsuperscript{2} Microsoft
\{krisgan,navendu,jelundin,ryenw\}@microsoft.com
\textsuperscript{3} Cheriton School of Computer Science, University of Waterloo
jimmylin@uwaterloo.ca

Abstract

Mining biomedical text offers an opportunity to automatically discover important facts and infer associations among them. As new scientific findings appear across a large collection of biomedical publications, our aim is to tap into this literature to automate biomedical knowledge extraction and identify important insights from them. Towards that goal, we develop a system with novel deep neural networks to extract insights on biomedical literature. Evaluation shows our system is able to provide insights with competitive accuracy of human acceptance and its relation extraction component outperforms previous work.

1 Introduction

Biomedical literature offers a rich set of knowledge sources to discover important facts and find associations among them. For instance, MEDLINE contains over 18 million references to articles published since 1946 and sourced from over 5500 journals worldwide (Simpson and Demner-Fushman, 2012). Two major processing tasks performed on the biomedical text are: (1) identify and classify biomedical entities (NER) into predefined categories such as proteins, genes, or diseases, and (2) infer pair-wise relationships among named entities e.g., protein-protein interaction (Poon et al., 2014), gene-protein, and medical problem-treatment.

This paper presents a system that processes biomedical text to extract two specific types of relationships among biomedical entities: (a) cause-effect and (b) correlation.

This system is motivated by the need to better automate biomedical knowledge extraction and identify important information from them, as new scientific findings appear across a large collection of publications. For instance, given user sleep patterns, existing biomedical research can be better utilized to provide insights: inform about potential effect (e.g., “diabetes”, “obesity”) due to the cause (e.g., “sleep disorder”) and suggest appropriate treatment.

Since biomedical articles usually have title and abstract summarizing the contents of the full-text article, we focus on extracting the two relationship types from them. Unfortunately, mining this summary data still poses several key challenges. Similar to full-text, this data comprises unstructured text with domain-specific vocabulary, issues of synonymy (e.g., “heart attack” vs. “myocardial infarction”), acronyms, abbreviations and rapidly evolving terminology due to new scientific discoveries. While the titles are short and informative, they do not contain the key information that would be contained in the abstract.

Many of these challenges are also applicable for biomedical relation extraction. Further, identifying particular relation types is challenging because relations are expressed as discontinuous spans of text, and the relation types are typically application-specific. Finally, there is often little consensus on how to best annotate relation types resulting in lack of high quality annotated corpora for training.

In this study, we develop neural networks with novel similarity modeling for better causality/correlation relation extraction, as we map the extraction task into a representational similarity measurement task in the vector space. Our approach innovates in that it explicitly measures both relational and contextual similarity among representations of named entities, entity relations and contexts. Our system also provides a novel combination of recognizing named entities, predicting...
relationships (insights) between extracted entities, and ranking the output. We conduct human evaluations of the system to show it is able to extract insights with high human acceptance accuracy, and on a SemEval task evaluation its causality/correlation relation extraction compares favorably against previous state-of-the-art work.

1.1 Contributions
1. We build an end-to-end system to extract insights from biomedical literature.
2. We innovate in similarity measurement modeling with deep neural networks for better causality/correlation relation extraction.
3. Our human evaluation shows our system can achieve competitive acceptance accuracy.

2 Related Work
Most previous work in BioNLP focused on extraction of biomedical concepts (Craven, 1999; Finkel et al., 2005; Poon and Vanderwende, 2010; Simpson and Demner-Fushman, 2012; Liu, 2016), such as drug or protein names. We also conduct relation extraction on general named entities, such as “smoking” or “sleep quality”. Kabiljo et al. (2009) compared pattern-matching techniques against a baseline regular expression approach for gene/protein entity extraction. But existing tools for relation extraction are not as comprehensive as entity recognition tools.

Medical dictionaries and resources are heavily utilized by previous work. For instance, Chen et al. (2008) extracted disease-drug relation pairs with MedLEE (Friedman et al., 2004) system for clinical information extraction of EHR records. Liu et al. (2015) developed a text-mining system to search for associations among human diseases, genes, drugs, metabolites and toxins against large collections of text-rich biological databases. Previous research efforts also lead to semantic representation program SemRep (Rindflesch and Fiszman, 2003), which exploits biomedical domain knowledge and linguistic analysis of biomedical text. Other unconventional resource such as web query logs are also utilized (Paparrizos et al., 2016) to provide early warnings about the presence of devastating diseases.

Feature engineering was the dominant approach in most biomedical relation extraction work with machine learning techniques (Dogan et al., 2011; Yala et al., 2016); different sparse features were explored. For example, word n-gram features, knowledge-based features from medical dictionaries and word position features. Our work instead propose neural network models that do not require sparse features as in most previous work.

Recent shift from feature engineering to model engineering with neural networks has significantly improved accuracy on many NLP tasks. Jagannatha and Yu (2016) adopted an LSTM model for medical entity detection given patient EHR records. There are recent work with the use of deep reinforcement learning on healthcare study (Li, 2017). Our approach is inspired by recent embedding learning work to jointly represent texts and knowledge base (Toutanova et al., 2015, 2016), previous work on embedding transfer learning (Bordes et al., 2013) and noise-contrastive estimation (Rao et al., 2016). Lastly our work models insight extraction as a similarity measurement problem, and is inspired by similarity measurement work (He et al., 2016; He and Lin, 2016) on pairwise word interaction modeling with deep neural networks.

3 System Overview
We provide a recipe to build a system for biomedical insight extraction and use it as a guide for the remainder of this paper (Algorithm 1).

To make our discussion concrete, we will use a sample biomedical article in Example 1. Given the text, at line 4 of Algorithm 1 we firstly look for all named entities using a shallow parser and public medical dictionaries (see details in Section 4). Many named entities could be found, for example, "clinical study", "sleep disturbances in middle-aged men" and "diabetes". Next given any pair of previously extracted entities within a sentence,
RESEARCH METHODS: A group of 6,599 initially healthy, nondiabetic middle-aged men took part in a prospective, population-based study. The incidence of diabetes during a mean follow-up of 14.8 years was examined in relation to self-reported difficulties in falling asleep.

RESULTS: A total of 615 subjects reported either difficulties in falling asleep or use of hypnotics (seen as makers of sleep disturbances). Among those, 281 of the men developed diabetes during the follow-up period. The clinical study suggests sleep disturbances in middle-aged men are likely associated with diabetes.

Example 1: Sample Text

at line 6 our neural network-based relation extractor checks if a valid causality/correlation relationship exists (Section 5). For example, our models can identify that the entity “sleep disturbances in middle-aged men” has a correlation relationship with “diabetes” but not with “clinical study”. Later each valid entity pair is scored via the ranking component at line 7 (Section 6). In the final step, the system returns top ranked insight(s) to users: “sleep disturbances in middle-aged men → diabetes” given this example.

Figure 1 presents the system which consists of three major neural network-based components: (1) a named entity extractor, (2) a causality/correlation relation extractor, and (3) an insight ranker. Our system reads in biomedical texts, then provides insights in the end. We primarily innovate in the relation extraction component. Next, we describe each of these components in detail.

4 Named Entity Extraction

Named entity extraction in biomedical domain is challenging due to the domain-specific and rapidly evolving terminology. For example, “Diabetes mellitus type 1”, “Type 1 diabetes”, “IDDM”, or “juvenile diabetes” all express the same concept. Given frequent evolution of entity naming for new drugs, diseases or abbreviations, this task becomes more complicated.

Most existing off-the-shelf biomedical entity recognizers narrowly focus on specific biomedical terms. Instead we aim to improve the system recall by extracting both specific biomedical concepts such as “gene tmem230” or “prostate cancer” as well as general noun phrases such as “sleep qual-
two named entities, then representations of the two entities should be semantically similar and close to the representation of the relation in a low-dimensional vector space. Therefore we map the causality/correlation relation extraction into a similarity measurement task in the vector space.

Our novel approach learns representations of named entities \( (\vec{A}, \vec{B}) \), context words and the relation vector \( \vec{R} \), then explicitly measures two aspects of the similarity: 1) relational similarity between entities and relation (Sec. 5.2); plus, 2) contextual similarity between entities and sentence context (Sec. 5.3).

The intent of our approach is to enforce such structure of the vector space: as the similarity among entities, relation and contexts gets stronger, a fit of all should be observed for better causality/correlation relation extraction. We develop two neural network models with such property; both are utilized in the relation extraction component of the system.

We define input sentence representation \( S \in \mathbb{R}^{t \times d} \) to be a sequence of \( t \) words, each with a \( d \)-dimensional word embedding vector. \( x_t \in \mathbb{R}^d \) denotes the embedding vector of the \( t \)-th word \( (t \in [1, l]) \) in \( S \). Model details are described in the following sections.

### 5.1 Context Modeling

Different words occurring in similar contexts should have a higher chance to contribute to similarity measurement and relation extraction. We use bidirectional LSTMs (BiLSTM) for context modeling as a basis for all following models.

LSTM (Hochreiter and Schmidhuber, 1997) is a special variant of Recurrent Neural Networks (Williams and Zipser, 1989). At time step \( t \), given an input word \( x_t \) and previous LSTM hidden state \( h_{t-1} \), \( LSTM(x_t, h_{t-1}) \) outputs current hidden state \( h_t \in \mathbb{R}^{dim} \). BiLSTM consists of two LSTMs that run in parallel in opposite directions. The BiLSTM hidden state \( h_t^b \in \mathbb{R}^{2dim} \) is a concatenation of forward LSTM’s \( h_t^f \) and backward LSTM’s \( h_t^b \), representing contexts of input word \( x_t \) in the sentence. We define \( \text{concat} \) operation and output sentence context representation \( H^S \in \mathbb{R}^{t \times 2dim} \) below:

\[
\begin{align*}
    h_t &= LSTM(x_t, h_{t-1}) \\
    h_t^b &= \text{concat}(h_t^f, h_t^b) \\
    H^S[t] &= h_t^b
\end{align*}
\]

Context modeling with BiLSTM allows our following model components to be built over contexts rather than over individual words. Given named entity positions of the sentence, we get \( \vec{A} \) and \( \vec{B} \) from context \( H^S \).

### 5.2 Relational Similarity Modeling

Relational similarity modeling focuses on interactions between named entities and relations in the vector space. When the named entity \( \vec{A} \) goes through a transformation process induced by the relation \( \vec{R} \), our intent of relational similarity modeling is to force the transformed entity to be translated to the other named entity \( \vec{B} \) in the same vector space so that the relation \( \vec{R} \) holds between the two named entities.

We show the following objective function of our relational similarity modeling:

\[
\vec{A} + \vec{B} - \vec{R} \simeq 0 \tag{4}
\]

To model the transformation process in Equation 4, we need to know how to measure the similarity of the triplet \( (\vec{A}, \vec{B}, \vec{R}) \). Therefore we develop a similarity measurement function \( \text{SimiScore}(\vec{A}, \vec{B}, \vec{R}) \) with learnable weights \((W^*)\), the similarity function takes an input named entity pair of \((\vec{A}, \vec{B})\) and a relation \( \vec{R} \), returns a similarity score \( \text{dist} \in \mathbb{R}^1 \) representing how semantically close \((\vec{A}, \vec{B}, \vec{R})\) are, as in Function 1.

We utilize a ranking approach during training to incorporate the constraint of Equation 4 into the relational similarity model. Our goal is to learn a function \( \text{SimiScore}() \) so that the positive triplet \( (\vec{A}, \vec{B}, \vec{R}^+) \) is assigned a larger score than that of the negative triplet \( (\vec{A}, \vec{B}, \vec{R}^-) \):

\[
\text{SimiScore}(\vec{A}, \vec{B}, \vec{R}^+) > \text{SimiScore}(\vec{A}, \vec{B}, \vec{R}^-) \tag{5}
\]

where \( \vec{R}^+ \) denotes the positive causality/correlation relation, \( \vec{R}^- \) denotes a non-causality/non-correlation relation. The ranking approach maximizes the similarity score between the entity pair \((\vec{A}, \vec{B})\) and a positive relation \( \vec{R}^+ \) while minimizing the score with the negative \( \vec{R}^- \).
Lung cancer, the context words ⃗R Non-Caus. Relation ⃗R is ⃗C : Lung cancer is most likely caused by smoking

Named Entities (⃗A, ⃗B)

Contextual Similarity

Causality Relation ⃗R+

Non-Caus. Relation ⃗R−

Relational Similarity

thus ensuring that the positive connection is larger than the negative one as in Figure 2.

Our relational similarity model and the ranking training approach facilitate the transformation process of (⃗A, ⃗B) and ⃗R in the vector space, which in the end leads to better constraint satisfaction of objective Equation 4.

The relational similarity model is placed on top of BiLSTM (Sec 5.1) as part of the system. We initialize named entities ⃗A/⃗B as \( h^A_t / h^B_t \) from the BiLSTM model, then initialize relation representations \( R^+ / R^- \) as random vectors. During training both \( R^+ / R^- \) are updated.

### 5.3 Contextual Similarity Modeling

Since not all words of a given title/abstract are created equal, important context words around named entities that can better contribute to the causality/correlation relation extraction deserve more model focus. We develop a contextual similarity model that can increase model weights onto important context words to better utilizing contextual information.

For example, given a sentence, *lung cancer is most likely caused by smoking*, the context words *caused by* are important clues to suggest there exists a causality/correlation relationship between the two named entities. Clue words that require model attentions usually include, e.g. *lead to, is associated with, because of*, while others are not obvious, such as *promote, reflect, reduce, make*.

Our system does not require a manually prepared list of clue words, but an attention mechanism (Bahdanau et al., 2014) is utilized to better identify them by conducting similarity measurement between context word representation \( h^t_w \) (not including entity words) and extracted named entities (⃗A, ⃗B) (from Sec. 4). Resulting similarity scores of words are accumulated in \( \text{atten} \in \mathbb{R}^d \).

\[
\text{mix} = W^a \cdot \text{concat}(\bar{A}, \bar{B}) \quad (6)
\]

\[
E[t] = \text{dotProd} (\text{mix}, h^t_w) , \forall t \in [1, l] \quad (7)
\]

\[
\text{atten} = \text{softmax} (E) \quad (8)
\]

where we concatenate both entity representations (⃗A, ⃗B), apply linear transformation with weights \( W^a \) to obtain a representation \( \text{mix} \) of both entities. We then use dot product \( \text{dotProd} \) to measure the similarity between \( \text{mix} \) and each context word, finally normalize the attention weights \( \text{atten} \) with \( \text{softmax} \). The weights of \( \text{atten} \) indicate the importance of each context word with respect to both named entities.

The attention weights should better guide the focus of the model onto important context words of the sentence. That is, context words that are closer to entity representation \( \text{mix} \) should have better chances to be clue words. We define the attention re-weighted sentence representation \( \text{attenSen} \in \mathbb{R}^{2 \cdot \dim} \):

\[
\text{attenSen} = \text{atten} \odot H^S \quad (9)
\]

where \( \odot \) represents element-wise multiplication.

Figure 2 illustrates an example where representation \( \text{mix} \) of named entities attends to context words one at a time. Important context clue words “caused by” should receive higher attention weights than irrelevant neighbor words.

The re-weighted sentence representation \( \text{attenSen} \) is used together with entity representations (⃗A, ⃗B) for final prediction.

In summary, both models described in this section focus on different aspects of similarity measurement in relation extraction: the contextual similarity model utilizes context information around named entities, while the relational similarity model focuses on enforcing a transformation constraint between entities and relation in the vector space. We adopt both models for better relation extraction, in the end only pairs of named entities that are recognized positively by either one of the models are passed to the next stage of the system.
6 Ranking of Extracted Insights

The last major component of our system is to rank extracted relations \((\vec{A}, \vec{B}, \vec{R})\) from the output of the relation extraction component, as there could be many extracted relations but not all of them are important enough as insights of the article. Importance scores of extracted relations are obtained by following a set of rules below:

1. We utilize the output classification probability \(\in [0, 1]\) of the relational similarity model as the base ranking score.
2. We use a multi-perspective convolutional neural network model (MPCNN) (He et al., 2015) to measure the similarity \(\in [0, 1]\) between the title of the article and extracted relation, since the MPCNN model has competitive performance on multiple benchmarks for textual similarity measurement. We compare title text with “\(\vec{A}\) leads to \(\vec{B}\)” of an extracted relation, if the similarity score is over a threshold of 0.75, we increase the extracted relation’s ranking score by 15%. If the extracted relation is from the title text, we also boost its ranking score by 15% because of its location importance.

Once all extracted relations are scored, our system only returns the top ranked insights to users.

7 Experiment Setup

Datasets. Experiments are conducted on two datasets: our own dataset of medical/health publications annotated on Universal Human Relevance System (UHRS), a crowdsourcing platform for end-to-end system evaluation; and SemEval-2010 task 8 dataset for training and evaluation of our relation extraction component:

1. The first dataset consists of 100 publications from recent biomedical/health journals, which are then annotated on UHRS to evaluate our system. In order to ensure high-quality human annotations, Figure 3 provides an annotation interface on UHRS, which displays instructions, title/abstract texts of publications and a list of top ranked extracted insights from the system output. For fair evaluation the order of extracted insights is randomized then we ask expert annotators with suitable background to verify the correctness of each.
2. SemEval-2010 Task 8 (Hendrickx et al., 2009) defines 9 relation types between named entities: Cause-Effect, Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, Entity-Destination, Component-Whole, Member-Collection and Message-Topic, and a tenth relation type Other when two named enti-
ties do not have the first 9 relations. SemEval-2010 dataset consists of 10,717 sentences, with 8,000 for training and 2,717 for test. The dataset is human annotated, and each instance provides one sentence which includes two named entities and a relation type between the two entities.

Since our system focuses on extracting insights, we only use Cause-Effect subset of SemEval-2010 dataset as the positive training/testing examples and treat the remaining 9 categories data such as Content-Container, Message-Topic as negatives. We use this dataset for training and evaluating our relation extraction component (Sec. 5) only.

**Training.** Two loss functions are adopted to train relation extraction neural network models.

For contextual similarity model (Sec. 5.3), a hinge loss is used. The training objective is to minimize the following loss, summed over examples $(x, y_{gold})$:

$$\text{loss}_{\text{contextSim}}(w, x, y_{gold}) = \sum_{y' \neq y_{gold}} \max(0, 1 + f_w(x, y') - f_w(x, y_{gold})) \quad (10)$$

where input $x$ represents an entity pair $(A, B)$ plus its sentence context, $y_{gold}$ is the ground truth label and $y'$ is the model predicted label. Both $y'$ and $y_{gold}$ indicate the relation type with directionality (e.g., directional causality). $w$ represents weights of contextual similarity model with BiLSTM, function $f_w(x, y')$ outputs the model predicted label value, function $f_w(x, y_{gold})$ outputs the model ground truth label value, and $n$ is the number of training examples.

For relational similarity model (Sec. 5.2), a Bayesian Personalized Ranking (BPR) loss (Rendle et al., 2009) is used. The label of the relational similarity model is binary because the BPR loss ranks positive inputs above negative inputs, thereby requiring the supervision signal to distinguish positives from negatives. Due to BPR loss’s ranking nature, each training instance of the relational similarity model include one positive input $(x, R^+)$ and one negative input $(x, R^-)$. Given a positive correlation/causality input $(R^+)$, we generate negative training examples by matching the input $x$ with each of the negative relation labels ($R^-$). BPR loss is shown to be better tailored for ranking tasks empirically (Verga et al., 2016):

$$\text{loss}_{\text{relSim}}(w, x, R^+, R^-) = \sum_{R^-} - \log(\sigma(f_w(x, R^+) - f_w(x, R^-))) \quad (11)$$

where $\sigma$ is the sigmoid function, function $f_w(x, R)$ represents the relational similarity model with BiLSTM, and outputs a similarity score for ranking purpose (Sec. 5.2).

In all experiments, we perform optimization using RMSProp (Tieleman and Hinton, 2012) with backpropagation (Bottou, 1998) and a learning rate fixed to $10^{-4}$ and a momentum parameter 0.9.

**Settings and Preprocessing.** We preprocess both datasets with Stanford CoreNLP toolkit (Manning et al., 2014). We tokenize, lowercase, sentence split and dependency parse all words of both datasets. We set LSTM hidden state $\text{dim} = 500$.

Two sets of $d = 300$-dimension word embeddings are utilized. The first one is 300-dimension GloVe word embeddings (Pennington et al., 2014) trained on 840 billion tokens; for better biomedical/health domain adaptation, we also train second word embeddings using the GloVe toolkit on biomedical research articles with over 1 billion tokens. We do not update word embeddings in all experiments.

During system deployment, we only initialize input words with the medical word embeddings if they do not exist in GloVe embeddings’ vocabulary. We also concatenate embeddings of both input words and their head words on dependency trees as input for relation extraction models. We follow the task settings and compute F1-score with the official evaluation script only on Cause-Effect subset of SemEval-2010 data, then the best model based on F1 is selected for final system deployment. We set a distance limit and do not extract relations between two named entities if the distance is larger than 15.

**8 Evaluation and Results**

**Human Evaluation of the Entire System.** We firstly provide a full end-to-end evaluation of the system on UHRS with human annotators.

For each biomedical publication, top 10 candidate insights from the system are listed for further inspection. The annotators are required to understand the texts, carefully inspect each insight, finally either accept it if it is one of the article insights or simply reject it. The annotation
Figure 4: Human evaluation results of the full system and a baseline system on UHRS. We show the acceptance accuracy for each of the top ten positions given both systems’ output lists. We primarily focus on the first 1 and 3 positions, namely Precision@1 and Precision@3.

<table>
<thead>
<tr>
<th>System Ablation Study</th>
<th>Precision@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full System</td>
<td>63%</td>
</tr>
<tr>
<td>- Remove ReRanker (Sec. 6)</td>
<td>-5%</td>
</tr>
<tr>
<td>- Replace with BiGRU (Sec. 5)</td>
<td>-42%</td>
</tr>
</tbody>
</table>

Table 1: Ablation studies on the full system.

We also provide a baseline system, of which its relation extraction component is a bidirectional gated RNN model (BiGRU) (Cho et al., 2014). BiGRU model and the ranking component are major differences between our full and baseline system. Since typically only a limited number of key findings is presented in one article, we evaluate the system with an averaged acceptance accuracy at top 1 (Precision@1) and top 3 (Precision@3) positions of the output rank list, which represent on average the number of extracted insights accepted by annotators among the first 1 and 3 output.

Figure 4 shows annotation results with acceptance accuracies for each of the ten output positions given biomedical titles and articles. The Precision@3 of our full system is 50.6%, which is significantly better than the baseline system’s 21.3%. For top 3 extracted insights on the list, our full system on average have 1.5 insights accepted by annotators. Furthermore, the acceptance accuracy Precision@1 of our system is 63% in comparison to that of the baseline system’s 21%.

Table 1 shows the ablation study on the removal of the ranking component (Sec. 6) and the replacement of BiGRU model for causality/correlation relation extraction. We observe significant performance difference.

Evaluation of Relation Extraction Component.

We also evaluate the relation extraction component (Sec. 5) on Cause-Effect subset of SemEval-2010 dataset. Note our causality/correlation relation extraction component is not supposed to be a general purpose one, since our system only focuses on insight extraction of biomedical/health literature. We compare our relation extraction models against previous work on the Cause-Effect subset of the data, Table 2 shows our relational similarity model, without the use of sparse features or external resources such as WordNet, outperforms recent state-of-the-art treeLSTM model (Miwa and Bansal, 2016). It also shows BiGRU model is reasonably competitive on this dataset, which is why we use it in our baseline system for comparison purpose.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 score+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tymoshenko and Giuliano (2010)</td>
<td>82.30%</td>
</tr>
<tr>
<td>Tratz and Hovy (2010)</td>
<td>87.63%</td>
</tr>
<tr>
<td>Rink and Harabagiu (2010)</td>
<td>89.63%</td>
</tr>
<tr>
<td>BiGRU</td>
<td>89.89%</td>
</tr>
<tr>
<td>Miwa and Bansal (2016)</td>
<td>91.57%</td>
</tr>
<tr>
<td>Contextual similarity modeling</td>
<td>90.77%</td>
</tr>
<tr>
<td>Relational similarity modeling</td>
<td>92.28%</td>
</tr>
</tbody>
</table>

Table 2: Test results (F1 score) on the Cause-Effect subset(*) of SemEval-2010 dataset. Results are grouped as 1) Top 3 participating teams in SemEval-2010 competition; 2) Baseline BiGRU model; 3) Recent state-of-the-art treeLSTM model (Miwa and Bansal, 2016); 4) Our work.

9 Result Analysis and Case Study

Visualization of Contextual Similarity Model.

We show values of attention weights, $atten$ of...
Excess oil, dirt, and bacteria cause acne.

The bombing resulted in the deaths of 1318 in Hanoi.

Ambient vanadium pentoxide dust produces irritation of the eyes.

Electron beam is generated by an explosive emission cathode.

Table 3: Visualization of model attention weights $\text{atten}$ given four SemEval-2010 test sentences.

---

Equation 8 and 9 from within the contextual similarity model (Sec. 5.3). Given four sentences in the test set of SemEval-2010 data, the model predicts that all provided entity pairs (in bold) have the causality/correlation relation. From Table 3 we observe the model is able to do its expected job: it can recognize important clues words, such as “result in”, “produce”, “generated by” and “cause”; the model produces attention weights (each $\in [0, 1]$) to tell the importance of clue words for causality/correlation relation extraction. We also observe the model tends to focus more on prepositions of clue words, such as “by” of “generated by” and “in” of “result in”, this is probably because we use head words as extra inputs (Sec. 7) to the model.

**Case Study.** We lastly provide case study of our system. We show two biomedical articles’ titles and abstracts as examples, with only necessary omissions to remove irrelevant texts due to the space limit.

**Case 1:** Scalp recorded direct current potential shifts associated with the transition to sleep in man. Abstract: Cortical direct current (DC) potentials are considered to reflect the state of cortical excitability which may change characteristically from wakefulness to sleep. The present experiments examined changes in the scalp recorded DC potential in 10 healthy humans ... It is reasonable to assume that the slow negative shift of the DC potential at the transition from wakefulness to sleep reflects increased cortical excitability.

**Case 2:** Alcohol consumption and self-reported sunburn: a cross-sectional, population-based survey. Abstract: Heavy drinking has been associated with several cancers, including melanoma and basal cell carcinoma. ... 299,658 adults reported their use of alcohol in the preceding month and a history of sunburn in the preceding year. Approximately 33.5% of respondents reported a sunburn within the past year. ... Excessive drinking is associated with higher rates of sunburn among American adults. The observed relationship typifies the high-risk behavior associated with excessive drinking and suggests one pathway linking alcohol use with skin cancer.

---

10 Conclusion

We build an end-to-end system for insight extraction on biomedical literature. We develop novel similarity measurement modeling with deep neural networks to extract causation/correlation relations. Our evaluation shows the system is able to extract insights with competitive human acceptance accuracy and its relation extraction component compares favorably against previous work.

Acknowledgments

We thank Yifeng Liu, Scott Wen-tau Yih for insightful discussions, and Max Ma for engineering support. We also thank three anonymous reviewers for highly helpful reviews.
References


Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky.2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd


