Self-training improves Recurrent Neural Networks performance for Temporal Relation Extraction

Chen Lin[a], **Timothy A. Miller**[a], **Dmitriy Dligach**[b], **Hadi Amiri**[a], **Steven Bethard**[c], **Guergana Savova**[a]
[a]Boston Children's Hospital Informatics Program, Harvard Medical School
{firstname.lastname}@childrens.harvard.edu
[b]Loyola University Chicago
ddligach@luc.edu
[c]University of Arizona
bethard@email.arizona.edu

Abstract

Neural network models are oftentimes restricted by limited labeled instances and resort to advanced architectures and features for cutting edge performance. We propose to build a recurrent neural network with multiple semantically heterogeneous embeddings within a self-training framework. Our framework makes use of labeled, unlabeled, and social media data, operates on basic features, and is scalable and generalizable. With this method, we establish the state-of-the-art result for both in- and cross-domain for a clinical temporal relation extraction task.

1 Introduction

Neural network methods have obtained spectacular successes in the fields of computer vision (He et al., 2016; Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012; Graves and Jaitly, 2014), and machine translation (Sutskever et al., 2014), where large datasets are available for training. For extracting information from text, however, performance gains have been minimal or non-existent, with published work emphasizing that such performance parity is not obtainable without extensive feature engineering. Unlike other settings that have seen performance gains, information extraction tasks related to text typically have much smaller supervised training sets, and the neural network algorithms presumably do not see enough instances to optimally tune the large parameter space.

In this paper, we examine the important information extraction task of temporal relation extraction from clinical text. The state-of-the-art for this task is a machine learner with a heavily-engineered set of features (Sun et al., 2013; Lin et al., 2016a). The identification of temporal relations from the clinical text in the electronic medical records has been drawing growing attention because of its potential to provide accurate fine-grained analyses of many medical phenomena (e.g., disease progression, longitudinal effects of medications), with many clinical applications such as question answering (Das and Musen, 1995; Kahn et al., 1990), clinical outcomes prediction (Schmidt et al., 2005), and recognition of temporal patterns and timelines (Zhou and Hripcsak, 2007; Lin et al., 2014). Obtaining large supervised datasets for clinical tasks is expensive and difficult, so it has been challenging to show meaningful improvements from the recent explosion of sophisticated neural network methods.

Our hypothesis is that the range of interesting phenomena found in clinical data is much broader than what is covered by available gold standard datasets for temporal information extraction. The results of Clinical TempEval 2017 (Bethard et al., 2017) strongly support this latter point, as the performance of submitted systems drops severely when trained on gold instances in one domain and tested on a new domain. We are thus inspired to make use of unlabeled data in addition to gold standard data with a simple semi-supervised learning method–self-training and combine it with varieties of pre-trained word embeddings to overcome gaps in training data coverage. In self-training (Yarowsky, 1995; Riloff et al., 2003; Maeireizo et al., 2004), a classifier is first trained on existing labeled data, and then applied to unlabeled data (typically a much larger amount). The predicted instances above a confidence threshold are added to the training set and the classifier is re-trained. Self-training is especially attractive in a neural network setting because the primitive feature types used by these networks (i.e., tokens) are computationally more efficient to obtain than the sophisticated features typically used by feature engineering methods.

For pre-training, we investigate the use of multiple external data sources to train word embeddings that form the input layer of the model. Since our

Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018), pages 165–176
Brussels, Belgium, October 31, 2018.

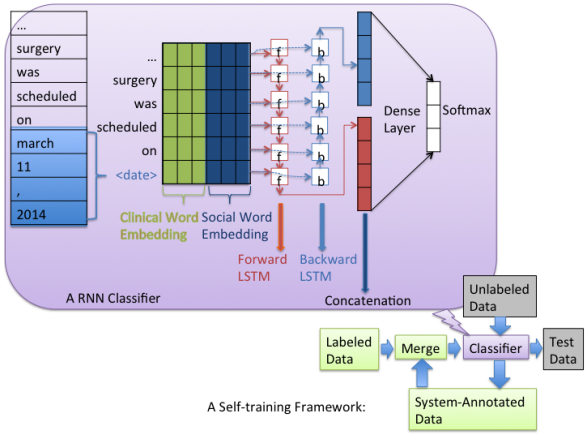


Figure 1: A RNN-based Self-training Framework

task is in the clinical setting, we use available clinical data sources, but also experiment with general domain sources trained on much larger datasets.

Besides showing that neural network approaches to information extraction can outperform feature-engineering approaches, we find that self-training works better in the neural network setting than with existing state-of-the-art feature-engineering approaches. Finally, we show that these methods generalize to new clinical domains better than the feature-engineering approaches we compare them to, obtaining state-of-the-art performance in an unsupervised domain adaptation setting.

2 Related Work

In recent years, several shared tasks on temporal relation extraction from clinical text have been organized. Among them, the i2b2 temporal challenge evaluates the i2b2 corpus (Sun et al., 2013), and Clinical TempEval series (Bethard et al., 2015, 2016, 2017) evaluate systems using the THYME corpus (Styler IV et al., 2014), which is annotated with time expressions (TIMEX3), events (EVENT), and temporal relations (TLINK) per an extension of the TimeML specifications (Pustejovsky et al., 2003; Pustejovsky and Stubbs, 2011). Challenge participants develop methods to extract EVENT and TIMEX3 entities, CONTAINS relations and document creation time relations. Herein, we focus on CONTAINS relation, which signals an EVENT occurs entirely within the temporal bounds of an *narrative container*. The *narrative container* is either another EVENT or TIMEX3.

Conventional learning methods, such as support vector machines (SVM) and conditional random fields (CRF) (Sun et al., 2013), have been developed for this task. Neural networks used in general relation extraction (Hashimoto et al., 2013; Socher et al., 2012), have also been adopted in clinical temporal relation extraction, such as structured perceptron (Leeuwenberg and Moens, 2017), convolutional neural networks (CNNs) (Dligach et al., 2017; Lin et al., 2017) and Long Short-Term memory (LSTM) networks (Tourille et al., 2017; Dligach et al., 2017). Classifiers are usually trained and tested in the same domain for the same medical condition, e.g. models are trained and tested on the colon cancer set of the THYME corpus for Clinical TempEval 2015 and 2016 (Bethard et al., 2015, 2016).

Clinical TempEval 2017 introduces the task of domain adaptation, as the most frequent use case would be the application of a model on a domain different from the domain it was trained on. The source domain of Clinical TempEval 2017 is colon cancer clinical text while the target domain is brain cancer clinical text. Few domain adaptation techniques are applied by the participants: 1) modeling unknown words to accommodate unseen vocabulary in the new domain; 2) using pretrained domain-independent word embeddings; 3) for supervised domain adaptation, assigning higher weights to samples from the new domain during model training. The performance on the domain adaptation task plummetted. Other domain adaptation methods used in general relation extraction include (Nguyen et al., 2014; Nguyen and Grishman, 2014; Plank and Moschitti, 2013).

Semi-supervised learning has been a popular approach for improving coverage and model generalizability for various information extraction tasks by exploring unlabeled data. Besides semi-supervised methods developed for feature-based learners (Le and Kim, 2015; Li and Zhou, 2010), there are such algorithms for deep neural network structures (DNN) (Laine and Aila, 2016; Kingma et al., 2014). Self-training or bootstrapping is a standard and straightforward semi-supervised learning method and widely used (Agichtein and Gravano, 2000; Pantel and Pennacchiotti, 2006; Greenwood and Stevenson, 2006; Rosenfeld and Feldman, 2007; Xu, 2008; Xu et al., 2007, 2010). To our best knowledge, we are the first to use self-training in a deep neural network setting for a clinical relation extraction task. Our motivation lies in two folds: 1) Self-training is computationally efficient as there is no other parallel learning goals such

as minimizing the reconstruction errors in Generative Adversarial Networks-based semi-supervised learning. With primitive features, DNN-based self-training can effectively and efficiently evaluate a large amount of instances; 2) We hypothesize that not all unlabeled data are useful. Our goal is to use a straightforward method like self-training to study the unlabeled space and help to select the most informative instances.

3 Data

We collect a variety of external data sources, described below, to supplement the THYME dataset (Styler IV et al., 2014).

3.1 Labeled Clinical Data

Our labeled data is the THYME corpus (Styler IV et al., 2014) used for the Clinical TempEval tasks. The corpus contains internal medicine, oncology, pathology, and radiology reports for 200 *colon* cancer patients and 200 *brain* cancer patients for a total of 1200 notes. Following the unsupervised domain adaptation setting of Clinical TempEval 2017, we use colon cancer notes for model development, and brain cancer notes for cross-domain validation.

3.2 Unlabeled Clinical Data

We augment the labeled data with additional clinical notes for colon cancer patients for a total of 27,157 notes (average length=135 words) from the same medical center as the THYME corpus from Section 3.1. On average, each patient has 125 notes of varied types – primary care, specialty care, pathology, radiology, etc. This set includes all electronic medical record notes at a single medical center for the 200 colon cancer THYME patients. We use it to automatically derive additional training instances, and refer to these generated instances as *silver* instances. We do not have access to additional unlabeled out-of-domain data (i.e. brain cancer clinical notes).

3.2.1 Clinical Word Embeddings

To train word embeddings with good vocabulary coverage and high representational power, we took advantage of the clinical notes from MIMIC-III (Medical Information Mart for Intensive Care) dataset (Johnson et al., 2016). The publicly available MIMIC III contains 879 million words from Beth Israel Deaconess Medical Center's Intensive Care Unit. We merged MIMIC-III data with the unlabeled colon cancer set above and trained 300-dimension embeddings with fastText (Joulin et al., 2016) and skip-gram (Guthrie et al., 2006) models.

3.2.2 Social Media Word Embeddings

While unlabeled clinical data provides a domain-matched source for training embeddings, additional data can be freely obtained from social media posts about colon cancer. To explore the benefits of extra coverage of such datasets versus the domain specificity of clinical embeddings, we obtain another set of embeddings using user-generated content about colon cancer from two social media platforms, namely Twitter and Reddit. For this purpose, we first generate a keyword list from two sources: a) the most frequent medical terms in the unlabeled colon cancer notes, these include any term that maps to the Unified Medical Language System concept unique identifiers (UMLS CUIs) (Bodenreider, 2004), b) the most frequent terms that map to ICD-9 billing codes related to colon cancer. These two lists results in a total number of 143 keywords. We use these keywords as a filter to collect 1.7 million publicly-available tweets about colon cancer. In addition, we collect 19K Reddit posts that contain at least one mention of colon cancer. We remove all occurrences of usernames, hash tags, URLs, and non-ASCII characters from the resulting data and employ fastText (Joulin et al., 2016) to obtain social media word embeddings.

In addition to the above embeddings, we utilize the Google News embeddings[1] trained by word2vec (Mikolov et al., 2013).

4 Methods

We develop a self-training framework to generate additional *(silver)* instances of CONTAINS relation (see Figure 1, lower-right). We focus on within-sentence CONTAINS relations and set aside all cross-sentence relations based on two motivations. First, the majority of the gold standard CONTAINS relations occur within a sentence.[2] Second, a sentence is a complete semantic and syntactic structure, which makes it an ideal unit for a sequence model, like RNN, to operate on. We therefore ignore cross-sentence CONTAINS links and focus on within-sentence CONTAINS relations. In addi-

[1] https://code.google.com/archive/p/word2vec/

[2] 4,3654 within-sentence vs. 743 cross-sentence CONTAINS relations in colon cancer test set. We note that it is impractical to link all cross-sentence events and/or time expressions pairs due to the large number of potential links.

tion, since we use the official Clinical TempEval 2017 scoring tool, our models are penalized for the missed cross-sentence relations.

4.1 Preprocessing

We process the labeled and unlabeled clinical data through the sentence detection and tokenization modules of Apache cTAKES[3]. For the labeled clinical data, we use gold standard event and time expression annotations and their time classes (Styler IV et al., 2014) for both model development and final validation. For the unlabeled clinical text data, we use the cTAKES event annotator (Lin et al., 2016a) and time expression annotator (Miller et al., 2015) to automatically annotate event and time expressions along with their time classes (e.g., TIME, DATE, SET). Both labeled and unlabeled corpora are transformed to lower case as shown in Figure 2.

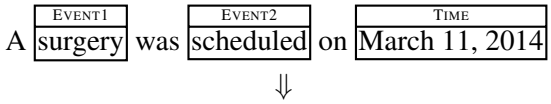
4.2 Instance Representation

We first create a dataset of within-sentence CONTAINS-relation candidates from the colon cancer text of the labeled clinical data. Given all gold standard events and time expressions within a sentence, we link every pair of events, and every event to a time expression (if present) to form CONTAINS candidates.

To mark the position of the relational arguments in a candidate pair, we adopt the same xml-tag marked-up token sequence representation as previous work (Dligach et al., 2017), and encode the time expression with its time class (Lin et al., 2017) for better generalizability. Figure 2 illustrates the marked-up token sequence representations for all three relational candidates, in which the event in an event-time relation pair is marked by ⟨e⟩ and ⟨/e⟩ and the time expression is marked by ⟨t⟩ and ⟨/t⟩. The time expression is further encoded by its time class, ⟨t⟩ ⟨date⟩ ⟨/t⟩, which is a gold standard attribute of a time expression annotation (Styler IV et al., 2014). Event-event instances are marked with additional indexes 1 and 2, e.g. *a ⟨e1⟩ surgery ⟨/e1⟩ is ⟨e2⟩ scheduled ⟨/e2⟩ on march* 11.

We also follow previous best practice in applying transitive closure to existing gold CONTAINS relations on the training data (Mani et al., 2006; Lin et al., 2016a). Depending on the order of the relational arguments, there are three types of gold standard relational labels, CONTAINS, CONTAINED-BY, and NONE.

[3]http://ctakes.apache.org



Candidate 1: a ⟨e⟩ surgery ⟨/e⟩ was scheduled on ⟨t⟩ ⟨date⟩ ⟨/t⟩;
Candidate 2: a surgery was ⟨e⟩ scheduled ⟨/e⟩ on ⟨t⟩ ⟨date⟩ ⟨/t⟩;
Candidate 3: a ⟨e1⟩ surgery ⟨/e1⟩ was ⟨e2⟩ scheduled ⟨/e2⟩ on march

Figure 2: Representations of event-event and event-time relational candidates in a sentence

4.3 Bidirectional RNN Classifier

We use a bi-directional recurrent neural network to model the relational context similar to the state-of-the-art model (Tourille et al., 2017). As shown in Figure 1 (upper-left), each token in the token sequence input is represented by one set of clinical embeddings and one set of additional embeddings (either cancer-related social media embeddings or Google news embeddings) to capture the semantics exhibited by clinical and non-clinical terms.

As described in section 3.2.1, the clinical embeddings are derived from combining the MIMIC III and unlabeled colon cancer datasets. For the unlabeled colon cancer data, we use the extracted relational candidates as shown in Figure 2 to train embeddings, so that all xml-tag marked-up tokens and time-class tokens, e.g. ⟨/e⟩, ⟨/e1⟩, ⟨/t⟩, ⟨/date⟩, are represented. For each set of embeddings, an `UNK` token represents out-of-vocabulary words to accommodate unseen words in a new domain. Table 1 shows the coverage of each embedding set and their combinations over the labeled colon cancer training set. We will show the effect of the different embedding combinations in the experiments.

The two sets of embeddings for a given token are concatenated and fed into the two sequences of hidden states of RNN: forward states and backward states. The output of the two states is concatenated and fed into a dense layer and through a softmax layer to predict three relational labels as described in section 4.2. We evaluate two RNN models, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated recurrent units (GRUs) (Chung et al., 2014).

We implement the network in Keras (Chollet, 2015) with Theano (Theano Development Team, 2016) backend. We train our models with a batch

corpora	word#	coverage
(1) Clinical	136K	94.66%
(2) Cancer-related social media	60K	76.67%
(3) Google News	3M	83.69%
(1) + (2)	171K	95.69%
(1) + (3)	3M	95.70%

Table 1: Embedding word coverage (percentage of words in the THYME corpus covered by the vocabulary in each corpora); Clinical embeddings derived from the combination of MIMIC and unlabeled colon cancer datasets, see section 3.3; Cancer-related embeddings derived from the combination of relevant Reddit posts and tweets, see section 3.4

size of 256, Stochastic Gradient Descent using Adam optimizer (Kingma and Ba, 2014), and a learning rate of 0.0001, on a GTX TitanX GPU. The hyper-parameters are optimized through a random search algorithm (Li et al., 2016) and the size of the hidden states of the forward and backward recurrent neural networks are set 512. We keep 10% of the training samples as a validation split, and applied a 0.5 dropout ratio and 0.0001 L2-regularized penalties to the embedding layers. For the high-precision model, we increased the weight of the L2-regularizer from 0.0001 to 0.001.

4.4 Self-Training

We apply the high-precision bi-directional RNN model trained on the labeled data to generate CONTAINS predictions on the unlabeled colon cancer data for *silver* annotations. We retain instances with a confidence score as generated by the `softmax` function of greater than 0.9 (a higher threshold will result in too few positive instances, a lower threshold will result in disproportionately many negative instances). We find that a lower threshold leads to low quality predictions and a higher threshold generates too few CONTAINS relations. The retained silver instances are merged with the gold ones and input into the bi-directional-RNN for a second-round of training.

As a comparison, we use self-training with the state-of-the-art SVM model (Lin et al., 2016a,b) to generate *silver* relations. The SVM-based THYME system is the latest release of Apache cTAKES v4 temporal module. For a comparison with the best setting of RNN-based self-training, we add all positive (CONTAINS, CONTAINED-BY) *silver* relations with the confidence threshold of greater than 0.9 to the gold training data of THYME corpus and then retrain the SVM model.

method	all silver	positive silver
joint bi-lstm	1.533M	19,441
SVM event-time	1.244M	57,462
SVM event-event	2.521M	36,960

Table 2: Number of generated silver training instances

Table 2 shows the number of silver instances generated by each learning algorithm. The high-precision bi-directional RNN model (joint-bi-lstm) is built upon LSTM networks with clinical and social media embeddings, and trained on the training split of the colon cancer set of THYME corpus.

5 Experiments

We experimented with several combinations of clinical and cancer-related social media and Google news embeddings. We tested three modes of merging silver instances with gold annotations (Figure 1, lower right): 1) Posi-Merge: merging the positive predictions (i.e. CONTAINS and CONTAINED-BY relations) with the gold relations; 2) sub-Merge: merging a subset of the silver data (a random sample of 45K silver samples including CONTAINS, CONTAINED-BY, and NONE relations) with the gold relations; and 3) all-Merge: merging all silver data with the gold relations. After merging, we shuffled gold and silver instances together to balance the batch-wise computation.

Models utilizing self-training were trained on the gold colon cancer training set of the THYME corpus and silver instances predicted from the unlabeled colon cancer data. Models were tested on the gold colon cancer and gold brain cancer development sets of the THYME corpus, comparing in-domain and cross-domain performance to select the best models for testing. The best models were tested on the gold colon cancer and brain cancer test sets (Clinical TempEval 2017 test sets).

All models were evaluated with the metrics precision (P), recall (R) and F1-score (F), using the standard Clinical TempEval evaluation script, where the P and R definitions are enhanced through temporal closure (UzZaman and Allen, 2011; UzZaman et al., 2012): when calculating precision, we run temporal closure on the gold relations but not on the system-generated ones; when calculating recall, we run temporal closure on the system-generated relations but not on the gold ones.

6 Results

Table 3 shows performance of the THYME system and various bi-directional RNN methods on the colon cancer and brain cancer development sets. For RNNs, we evaluated both LSTM and GRU models. For embedding combinations, we tested using the clinical embedding alone (C), using both clinical and cancer-related social media embeddings (CS), using both clinical and Google News embeddings (CG), and using Google News Embeddings alone (G). For ways to merge silver samples with gold instances we tested *no-self-training* in which no silver instances were used, *all-Merge* in which all silver instances were used, *sub-Merge* in which a subset of silver samples were used, and *Posi-Merge* in which only the positive silver instances were used. Among all settings, *bi-LSTM CG Posi-Merge* and *bi-LSTM CS Posi-Merge* achieved the best F1-score (F1b) on the brain development set; *bi-LSTM CS Posi-Merge* had the best F1-score (F1c) on the colon development set. These two best performing neural models along with the *THYME no-self-training* system were tested on the Clinical TempEval test splits.

Table 4 shows that the bi-LSTM models outperform the SVM-based THYME system and the Clinical TempEval 2017 top system, especially on the cross-domain experiments. The THYME system performance on the colon test set is 0.621 F1 which is an improvement over previously reported results (Lin et al., 2016b). The THYME system result on the brain cancer test is reported here for the first time. Note that the THYME system was trained on all gold colon cancer annotations (training, development and test), while the bi-LSTM models were trained on gold training colon cancer data and positive silver colon cancer samples. The best Clinical TempEval result on the gold colon cancer test set – 0.613 F1-score – is reported by the LIMSI-COT system which makes use of cTAKES-generated features (Tourille et al., 2017). The best Clinical TempEval result on the gold brain cancer test set – 0.34 F1-score – is achieved by the GUIR system (MacAvaney et al., 2017), while LIMSI-COT obtains 0.33 cross-domain F1-score.

7 Discussion

7.1 Comparison with SVM Self-Training

The top two rows of Table 3 show that the self-training technique did not improve the SVM-based THYME system. While recall reached its peak with the self-trained SVM, the precision trade-off was disastrous and F1 suffers dramatically. Our interpretation of this result is that the SVM is simply adjusting its class priors, labeling more instances as positive, but its fixed feature set and linear model constrain it from learning anything of interest from the silver data. The SVMs we use have extensively-engineered representations that were implicitly fit to the training and development sets of the colon cancer data. These feature sets may not have the representational power to find useful new patterns in the silver data. In contrast, the neural network models learn to extract features in their lower layers, and when given new data (e.g., silver data from self-training), the representation learning parts of the model are able to adapt and potentially find new patterns. This suggests that self-training for neural networks has higher potential than for SVMs, and that in the SVM setting, self-training should be accompanied by additional feature engineering.

Another difference between the models is that the SVM model relies on sophisticated linguistic features (parse trees, event and time expression attributes) that cannot be as reliably extracted from silver data. A token-sequence neural model, in contrast, makes use of very basic features and maintains a relatively accurate performance on the unlabeled data. It is possible that SVM performance is actually hurt by the lower quality features available from the silver training instances it encounters.

It is also worth noting that extracting additional silver instances for the SVM model is slower as it takes longer to generate the complex features that the SVM models use, while the token-based features of the neural model are extremely fast.

For all these reasons, we believe that neural networks are a more practical solution and better suited for a semi-supervised learning framework such as self-training.

7.2 Impact of Embeddings

Adding a broader range of embeddings as input to the bi-LSTM self-trained models improved the performance for the cross-domain task (rows 6-8 of table 3). It is possible that the clinical embeddings, even though trained on the mixture of MIMIC III and unlabeled colon cancer corpora, still do not provide semantic representation for the brain cancer notes. The diseases, symptoms, procedures, linguistic choices, etc. may vary substantially be-

Model	F1 drop ratio: (F1c-F1b)/F1c	colon cancer relations			brain cancer relations		
		P	R	F1c	P	R	F1b
1. THYME no-self-training	15.46%	0.661	0.587	0.621	0.533	0.518	0.525
2. THYME Posi-Merge	27.11%	0.185	**0.608**	0.284	0.123	**0.664**	0.207
3. bi-lstm CS no-self-training	16.59%	0.711	0.541	0.615	0.514	0.511	0.513
4. bi-lstm CS all-Merge	**8.87**%	**0.727**	0.431	0.541	**0.582**	0.428	0.493
5. bi-lstm CS sub-Merge	10.48%	0.712	0.549	0.620	0.567	0.543	0.555
6. bi-lstm C Posi-Merge	13.50%	0.712	0.551	0.622	0.528	0.549	0.538
7. bi-lstm CS Posi-Merge	9.63%	0.690	0.567	0.623	0.523	0.609	**0.563**
8. bi-lstm CG Posi-Merge	10.63%	0.684	0.584	**0.630**	0.513	0.624	**0.563**
9. bi-gru CS Posi-Merge	10.43%	0.702	0.559	0.623	0.522	0.600	0.558
10. bi-lstm G Posi-Merge	14.33%	0.673	0.530	0.593	0.475	0.545	0.508

Table 3: Model performance of *CONTAINS* relation on colon cancer and brain cancer development sets. C: clinical embeddings representation; CS: clinical and social media embeddings representation; CG: clinical and Google News embeddings representations; G: Google News embeddings. all-Merge: all silver instances added to gold training data; Posi-Merge: positive silver instances added to gold training data; sub-Merge: a subset of silver data added to gold training data.

Model	F1 drop ratio (F1c-F1b)/F1c	colon cancer relations			brain cancer relations		
		P	R	F1	P	R	F1
best Clinical TempEval	44.54%	0.657	0.575	0.613	0.52	0.25	0.34
THYME no-self-training	15.46%	0.661	**0.587**	0.621	**0.533**	0.518	0.525
bi-lstm CS Posi-Merge	13.14%	**0.700**	0.563	**0.624**	0.520	0.566	**0.542**
bi-lstm CG Posi-Merge	**13.04**%	0.692	0.576	**0.629**	0.514	**0.585**	**0.547**

Table 4: *CONTAINS* relations on colon cancer and brain cancer test set

tween these two cancer populations. Cancer-related social media and Google News embeddings come in with additional word coverage and more general semantic representations and thus help with the cross-domain performance. Word coverage increments are shown in Table 1. However, using non-clinical (Google News) embeddings on its own (row 10 of table 3) decreased both in-domain and cross-domain performance, even worse than the THYME system (row 1). It's possible that even though Google News embedding have good word coverage, general senses dominate clinical-specific senses, demonstrating the need for some clinical-specific data.

One interesting fact is that the cancer-related social media embedding has a much smaller vocabulary size than the Google News embeddings. Still, the CS option achieves the same F1-score as the CG option on the gold development brain set. Because of its better coverage and general semantic representation, CG option performs the best on the colon development set and the test sets of both colon and brain cancer data as shown in Table 4. We experimented with concatenating all three embeddings (clinical, cancer-related social media, and Google News), but did not observe any performance improvements.

7.3 Sampling of Silver Instances

Adding all high-confidence silver data to the gold training data clearly hurts performance (row 4). One possible explanation is the negative-to-positive instance ratio which is much higher in the silver data (80:1) than in the gold data (8:1). Adding the highly unbalanced silver samples may weight the system towards predicting the negative class, thus row 4 has higher precision but lower recall. Adding a random subset of silver samples to the gold samples provides additional information without skewing the class distribution, and we observe that in this setting the bi-LSTM model outperforms the THYME system, row 5 of Table 3. However, this setup may provide unpredictable performance due to the randomness of sampling the silver data.

The best merging option is the Posi-Merge. The models in rows 6-9 of Table 3 all outperform the THYME system, even for a single clinical embedding setting in row 6 of Table 3. Posi-Merge provides a stable sample of the silver data, strengthens the positive signals and achieves good cross-

domain performance.

7.4 Analysis of Improvements

We are interested in understanding the different contributions of self-training and pre-trained embeddings. Embeddings can provide a kind of adaptation for words in a new domain that are similar to words in the training data (e.g., *brain* in the brain cancer corpus may behave similarly to *colon* in a colon cancer corpus). However, self-training may still provide benefit if there are words in the test set that do not have correlates in the training data, but that can be found in the silver data. In these cases, confident silver instances provide information to the neural network about how these words should be integrated into the learned representations for predicting the relation category.

To investigate this possibility, we visualized the embeddings for gold training data, silver data, and development set data, all for colon cancer patients. We hope to find a sub-space in the embedding space where there is overlap between words in the silver data and development, but no nearby words from the training data. Figure 3 shows a visualized scatter-plot (Maaten and Hinton, 2008) of one such space, showing words from gold training set (blue), silver data (red), and the gold development set (yellow), given the clinical embedding. The upper-left cluster of silver words encloses several words occurring in the development set which are not represented or even close to the nearest words from the training set visualized in the lower right corner. Figure 3 shows that through self-training the vocabulary coverage is extended to less represented areas thus the model variance error is reduced which makes the model more generalizable.

7.5 LSTM vs. GRU

Given the same settings (rows 7 and 9 of Table 3), a GRU model performs similarly to a LSTM model for the in-domain task, but differently for the cross-domain task. GRUs are related to LSTMs, both utilize gating mechanisms to manage the vanishing gradient problem, though GRUs have fewer parameters. The performance difference may not be meaningful; we selected the LSTM for the test set evaluation due to its nominally better performance. However, given the small magnitude of these differences, future work may investigate whether GRUs may have advantages in reducing overfitting.

Figure 3: A part of T-SNE-visualized space

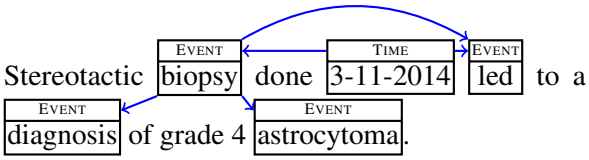


Figure 4: System annotations for a brain cancer sentence. Each arrow represents a CONTAINS relation.

7.6 Error Analysis

By comparing the error outputs of the THYME system and the best self-trained bi-LSTM system of Table 3 (rows 1 and 8) on the gold brain cancer development set, we find that the THYME system tends to pick up short-distance relation pairs, while the bi-LSTM model performs well on both short- and long- distance relations. One such example is shown in Figure 4. It represents a complex set of relations between four events and one time expression. All marked entities are participating in at least one CONTAINS relation, e.g. CONTAINS (3-11-2014, biopsy), CONTAINS (3-11-2014, led), CONTAINS (biopsy, led), CONTAINS (biopsy, diagnosis), CONTAINS (biopsy, astrocytoma). The link between two of the events, *biopsy* and *astrocytoma*, spans almost across the entire sentence. The bi-LSTM model predicts all relations correctly even without the assistance of transitive closure. We hypothesize that the benefit is due to the bidirectional setting of the LSTM model, which models the sentence structure very well. With the additional silver instances, two sets of embedding representations, and the memory capabilities, the self-trained bi-LSTM model adapts to a new domain to cover both short- and long-distance relations.

8 Conclusion

We show that neural models for temporal information extraction are able to take advantage of self-training. Compared with SVM models that leverage sophisticated features, our RNN-based self-training framework for temporal relation extraction operates on primitive features, models the sentence structure well, and is highly scalable and generalizable. Our RNN framework establishes a new state-of-the-art result for Clinical TempEval 2017 domain adaptation task. Experiments with externally-trained embeddings suggest that health-related social media or large scale general-domain text data can complement domain-specific text for a domain adaptation task. We will open source our learning framework in the near future.

Acknowledgments

The project is supported by 1U24CA184407-01 from the National Cancer Institute and R01LM010090 from the National Library Of Medicine at the US National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank James H. Martin and the anonymous reviewers for their valuable suggestions and constructive criticism. The Titan Xp GPU used for this research was donated by the NVIDIA Corporation.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062.

Steven Bethard, Guergana Savova, Martha Palmer, James Pustejovsky, and Marc Verhagen. 2017. Semeval-2017 task 12: Clinical tempeval. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 563–570.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

François Chollet. 2015. Keras. https://github.com/fchollet/keras.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Amar K Das and Mark A Musen. 1995. A comparison of the temporal expressiveness of three database query methods. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 331. American Medical Informatics Association.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. *EACL 2017*, page 746.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772.

Mark A Greenwood and Mark Stevenson. 2006. Improving semi-supervised acquisition of relation extraction patterns. In *Proceedings of the Workshop on Information Extraction beyond the Document*, pages 29–35. Association for Computational Linguistics.

David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1–4. sn.

Kazuma Hashimoto, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. Simple customization of recursive neural networks for semantic relation classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1372–1376.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

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

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Michael G Kahn, Larry M Fagan, and Samson Tu. 1990. Extensions to the time-oriented database model to support temporal reasoning in medical expert systems. *Methods of information in medicine*, 30(1):4–14.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Thanh-Binh Le and Sang-Woon Kim. 2015. Modified criterion to select useful unlabeled data for improving semi-supervised support vector machines. *Pattern Recognition Letters*, 60:48–56.

Tuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. 2016. Hyperband: A novel bandit-based approach to hyperparameter optimization. *arXiv preprint arXiv:1603.06560*.

Yu-Feng Li and Zhi-Hua Zhou. 2010. Improving semi-supervised support vector machines through unlabeled instances selection. *arXiv preprint arXiv:1005.1545*.

Chen Lin, Dmitriy Dligach, Timothy A Miller, Steven Bethard, and Guergana K Savova. 2016a. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, 23(2):387–395.

Chen Lin, Elizabeth W Karlson, Dmitriy Dligach, Monica P Ramirez, Timothy A Miller, Huan Mo, Natalie S Braggs, Andrew Cagan, Vivian Gainer, Joshua C Denny, and Guergana K Savova. 2014. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association*.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016b. Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 108–113. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. *BioNLP 2017*, pages 322–327.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. Guir at semeval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics.

Beatriz Maeireizo, Diane Litman, and Rebecca Hwa. 2004. Co-training for predicting emotions with spoken dialogue data. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 28. Association for Computational Linguistics.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 753–760. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Timothy A Miller, Steven Bethard, Dmitriy Dligach, Chen Lin, and Guergana K Savova. 2015. Extracting time expressions from clinical text. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015)*, pages 81–91. Association for Computational Linguistics.

Minh Luan Nguyen, Ivor W Tsang, Kian Ming A Chai, and Hai Leong Chieu. 2014. Robust domain adaptation for relation extraction via clustering consistency. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 807–817.

Thien Huu Nguyen and Ralph Grishman. 2014. Employing word representations and regularization for domain adaptation of relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 68–74.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics.

Barbara Plank and Alessandro Moschitti. 2013. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1498–1507.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Benjamin Rosenfeld and Ronen Feldman. 2007. Using corpus statistics on entities to improve semi-supervised relation extraction from the web. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 600–607.

Reinhold Schmidt, Stefan Ropele, Christian Enzinger, Katja Petrovic, Stephen Smith, Helena Schmidt, Paul M Matthews, and Franz Fazekas. 2005. White matter lesion progression, brain atrophy, and cognitive decline: the austrian stroke prevention study. *Annals of neurology*, 58(4):610–616.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211. Association for Computational Linguistics.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

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.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.

Julien Tourille, Olivier Ferret, Aurelie Neveol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 224–230.

Naushad UzZaman and James F Allen. 2011. Temporal evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 351–356. Association for Computational Linguistics.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Fei-Yu Xu. 2008. *Bootstrapping relation extraction from semantic seeds*. Saarland Univ., Department of Computational Linguistics and Phonetics.

Feiyu Xu, Hans Uszkoreit, Sebastian Krause, and Hong Li. 2010. Boosting relation extraction with limited closed-world knowledge. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1354–1362. Association for Computational Linguistics.

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*, pages 584–591.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical dataa review with emphasis on medical natural language processing. *Journal of biomedical informatics*, 40(2):183–202.