Attention-Based Bidirectional Long Short-Term Memory Networks for Extracting Temporal Relationships from Clinical Discharge Summaries

DOI:
10.1016/j.jbi.2021.103915

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Alfattni, G., Nenadic, G., & Peek, N. (2021). Attention-Based Bidirectional Long Short-Term Memory Networks for Extracting Temporal Relationships from Clinical Discharge Summaries. *Journal of Biomedical Informatics*, *123*, [103915]. https://doi.org/10.1016/j.jbi.2021.103915

Published in:
Journal of Biomedical Informatics

Attention-Based Bidirectional Long Short-Term Memory Networks for Extracting Temporal Relationships from Clinical Discharge Summaries

Ghada Alfattni[a,b], Niels Peek[c,d,e], Goran Nenadic[a,e]

[a] *Department of Computer Science, University of Manchester, Manchester, UK*
[b] *Department of Computer Science, Jamoum University College, Umm Al-Qura University, Makkah, Saudi Arabia*
[c] *Centre for Health Informatics, Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK*
[d] *National Institute of Health Research Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK*
[e] *The Alan Turing Institute, UK*

Abstract

Temporal relation extraction between health-related events is a widely studied task in clinical Natural Language Processing (NLP). The current state-of-the-art methods mostly rely on engineered features (i.e., rule-based modelling) and sequence modelling, which often encodes a source sentence into a single fixed-length context. An obvious disadvantage of this fixed-length context design is its incapability to model longer sentences, as important temporal information in the clinical text may appear at different positions. To address this issue, we propose an Attention-based Bidirectional Long Short-Term Memory (Att-BiLSTM) model to enable learning the important semantic information in long source text segments and to better determine which parts of the text are most important. We experimented with two embeddings

Email address: `gafattni@uqu.edu.sa` (Ghada Alfattni)

and compared the performances to traditional state-of-the-art methods that require elaborate linguistic pre-processing and hand-engineered features. The experimental results on the i2b2 2012 temporal relation test corpus show that the proposed method achieves a significant improvement with an F-score of 0.811, which is at least 10% better than state-of-the-art in the field. We show that the model can be remarkably effective at classifying temporal relations when provided with word embeddings trained on corpora in a general domain. Finally, we perform an error analysis to gain insight into the common errors made by the model.

Keywords:
TLINKs, NLP, Clinical text, Discharge summaries, Deep learning

1. Introduction

Free-text notes in electronic health records (including, for example, hospital discharge summaries, outpatient letters, handover notes, etc.) store key clinical information with the physician's explanation of patient's condition, their medical history, duration of symptoms, confirmed and rejected diagnostic hypotheses, patient preferences, treatment experience, etc. Automated extraction of information from such data sources has been used to unlock information on large scale to support clinical practice and epidemiological research [1]. One of the key tasks in clinical Natural Language Processing (NLP) is establishing temporal relations between clinical events, as this is essential for understanding the patient's trajectory and their health status, as well as improving quality of service, enhancing care and increasing healthcare utilisation [2]. However, temporal information is often represented by complex language expressions and requires advanced NLP to extract and categorise temporal relations.

Temporal links (TLINKs) are used to establish temporal relations between clinical events, and between clinical events and temporal expressions (TIMEXs). A recent systematic review of the work in temporal relation extraction from clinical free-text [3] has revealed that a small set of TLINKs (namely *before*, *after*, *overlap* and *contains*) has been more widely studied, as opposed to other types (e.g., *started by*, *finished by*, *precedes*) that remain challenging. The previous efforts were often placed in the context of shared tasks and benchmark datasets to assess and advance the state-of-the-art, such as the i2b2 2012 shared task [4] or the clinical TempEval series of shared challenges [5, 6, 7], with nearly all work evaluated on few publicly available corpora.

The vast majority of earlier methods used lexical resources and manually engineered features to leverage the linguistic knowledge [8, 9, 10, 11, 12, 13, 14, 15]. While results were encouraging, these efforts revealed several unresolved issues. For example, automatic extraction of high-level features (such as part of speech tags, entity identification (i.e., NER) and dependency paths) often used for TLINK mining typically resulted in error propagation, and thus overall performance degradation [16, 17, 18]. Similarly, feature engineering used in many approaches is time-consuming, and manually engineered features (i.e., regular expressions) often generalise poorly due to the varied nature of relations in text and the modest coverage of existing training datasets [19]. More recently, deep learning methods, which learn features automatically, have been used to tackle some of these issues [3]. Several deep learning methods based on various network architectures have been adopted for the task, including sequential modelling approaches such as Convolutional Neural Networks (CNNs) [18, 20, 21, 22] and Recurrent Neural Networks (RNNs) [23, 24]. However, these methods only consider the current input and what has been learnt from the inputs that are received previously.

Bidirectional RNNs introduce a mechanism to look at both prior and subsequent inputs before generating an output at a time step [25, 26, 27]. They still struggle to extract temporal relationships when the distance between related entities is long. The longer the input sequence length (i.e., the length between the relation entities), the more difficult it is to capture the context. Another issue is that some trigger phrases in the text (e.g., history of, continued, repeat, consecutive, subsequently) might act as a dominant feature and improve classification [14].

Attention mechanisms have been used to guide models to focus on parts of the text that are most influential with respect to target [28]. Attention-based neural network architectures have recently gained much attention and have been proven to be effective in several NLP tasks such as machine translation [26], question answering [27], recognizing textual entailments [28], and relation classification [29]. In the context of relation classification, several recent efforts (e.g., [16], [29], [30]) have successfully employed attention mechanisms to extract general relationships (e.g., Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, etc.) and have shown that they can perform as well as state-of-the-art relation classification systems based on features and neural networks. To the best of our knowledge, attention-based architectures for TLINKs have previously only been explored by Liu et al. [31], but they limited their experiments and evaluation to one type of temporal relation (i.e., intra-sentence temporal relations). In our study, we explored the attention mechanism that is integrated into a Bidirectional Long Short-Term Memory Network (BiLSTM) on a wider set of temporal relations (intra-sentence temporal relations, cross-sentence temporal relations, and references to document creation times).

Therefore, in this paper, we explore the attention mechanism that is integrated into a Bidirectional Long Short-Term Memory Network (BiLSTM) on a wider set of TLINKs (i.e., intra-sentences, cross-sentences and documents creation time relations). BiLSTM networks allow to make full use of context and capture the most important semantic information between relation entities. Thus, they minimise the performance dependency on features derived from lexical resources or NLP pre-processing. Afterwards, the attention mechanism is used to better determine which parts of the text are most influential for identifying temporal relations[1]. Using the i2b2 2012 temporal relation corpus, we demonstrate that BiLSTM networks with attention mechanism can be used for temporal relation classification between health-related events in clinical texts. We experiment with various embeddings and compare the performances to traditional state-of-the-art methods that require elaborate linguistic pre-processing and hand-engineered features. Finally, we perform an extensive error analysis to gain insight into the errors made by the models.

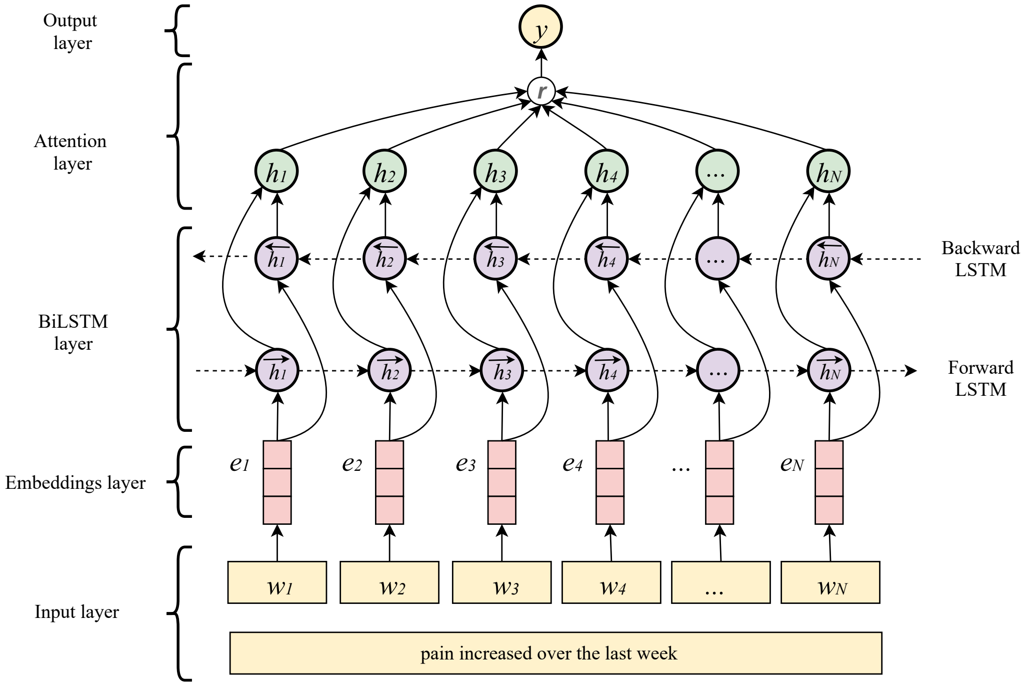

Figure 1: The Att-BiLSTM model architecture

2. Material and Methods

Figure 1 shows the architecture of the proposed Attention-based BiLSTM (Att-BiLSTM) model. It composes of five different layers: Input layer, Embeddings layer, Bidirectional LSTM layer, Attention layer and Output layer. The embeddings layer maps each word in an input into a vector representation. The sequence of vector representations corresponding to a sequence of words are input to the BiLSTM layer, which utilizes BiLSTM networks to capture important word-level features from the embedding layer. Then, the attention

[1] The framework is available on GitHub at `https://github.com/GhadaAlfattni/att-tlinks`

layer guides the networks to focus on specific information by producing a weight vector. After multiplying the weight vector, the word-level features from each timestep are converted into a sentence-level feature vector. Lastly, the output layer uses the sentence-level feature vector for relation classification and outputs the most likely relation type based on the sequence of probability vectors from the previous layer. These layers are described in more detailed below.

Input layer. Typically, this layer takes the position-marked relation entities with the surrounding tokens as inputs. For example, the sentence "pain increased over the last week" will be presented as "<E1>pain increased</E1> over the <T1>last week</T1>" where position markers are used to refer to the relation entities (E1 and T1).

Embeddings layer. Given the input context S composing of N words with markup of the relation entities $S = \{w_1, w_2, ..., w_N\}$, every word w_i is tokenized and mapped into a low-dimensional vector (i.e. embeddings) e_i to provide lexical and semantic features. This is done by using the matrix-vector product:

$$e_i = Wv^i$$

where W is the embeddings matrix, and v^i is a vector which has value of 1 at index e_i and 0 in all other positions. Then the sentence is transferred into the next layer as real-valued vectors.

Bidirectional LSTM layer. BiLSTM networks were used to calculate hidden states by processing sequence of token representations forwards and backwards (i.e. left-to-right and right-to-left). The forward LSTM network encodes the context of an input sentence and the backward LSTM network encodes the context of the reverse sentence. The output of the i^{th} word is calculated by the following equation:

$$h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}]$$

where $\oplus$ denotes the element-wise addition of outputs from forward and backward LSTM.

In this layer, we adopted a variant introduced by Graves [32] and then used for relation extraction by Zhou et al. [16]. It adds weighted peephole connections from the Constant Error Carousel (CEC) to the gates of the same memory block. By directly employing the current cell state to generate the gate degrees, the peephole connections allow all gates to inspect into the cell (i.e. the current cell state) even when the output gate is closed [32].

Attention layer. This layer takes a matrix H consisting of output vectors $h_1, h_2, ..., h_N$ that is produced by the the BiLSTM layer. Subsequently, the sentence vector r is computed as the weighted sum of α.

$$M = tanh(H)$$

$$\alpha = softmax(p^T M)$$

$$r = H\alpha^T$$

where p is a trained parameter vector and p^T is its transpose. The final sentence-pair representation that is used for classification in the output layer is performed by:

$$h^* = tanh(r)$$

For each relation instance, the embeddings layer, the Bidirectional LSTM layer and the attention layer together encode the relation instances into a multi-dimensional vector r. The encoded relation vector r is then fed into a fully connected layer (i.e., output layer). The output dimension of the output layer is set to the number of potential labels, which is 3 in this study.

Output layer. In this layer, we used a softmax classifier to predict the label (i.e., relation type) from a discrete set of classes for each sentence; the cost function is the negative log-likelihood of the true class labels. For regularization, we followed Zhou et al. [16], and combined dropout with L2 regularization to prevents neural networks from overfitting.

2.1. Dataset

We applied and evaluated the proposed method on the publicly available clinical corpus [4] that formed the basis for the 2012 i2b2 Temporal Relations challenge [33]. The corpus consists of 310 discharge summaries-190 summaries for training and 120 for testing. The gold standard annotations include time expressions (TIMEXs), events (EVENTs, both medical and general), and temporal relations (TLINKs). TLINKs can be assigned between:

1. EVENTs and document creation times (DCT), that is, the time stamp associated with the time when the clinical document was created;

2. EVENTs and TIMEX in one sentence, and

3. EVENTs in adjacent sentences.

TLINK type attributes can be BEFORE, AFTER or OVERLAP. Table 1 provides descriptive statistics of TLINKs types in the training and test sets of the 2012 i2b2 temporal relations corpus.

Relation type	**Training set**	**Test set**
AFTER	2981	2521
BEFORE	17348	14825
OVERLAP	11856	8948

Table 1: The number of annotated relations in the training and test sets in the 2012 i2b2 temporal relations corpus.

2.2. Training and Hyper-parameters

In this study we used the standard split established by the i2b2 organizers, using the training set for evaluating models and tuning model parameters, and evaluating our best models on the test set. Since there is no official development set, we randomly selected 10% of the training data for validation.

The hyper-parameters of our models were tuned to optimize the performance through the randomized parameter optimization algorithm, where each setting is sampled from a distribution over possible parameter values [34]. The choices generated by this process are as below: the model is trained using AdaDelta [35], with a learning rate of 1.0, a batch size of 10, LSTM layer dropout rate of 0.3, embeddings layer dropout rate of 0.3, and penultimate layer dropout rate of 0.5.

We experimented with two word embeddings: (1) the publicly available word embeddings GloVe [36], trained on Wikipedia and Gigaword 5 data (i.e. general domain); and (2) pre-trained word embeddings on the MIMIC clinical notes corpus (i.e. target domain) using word2vec. The reasons behind using the general word embeddings (GloVe) and the domain-specific word embeddings (MIMIC) were that: (1) they contain a large set of vocabularies;

thus, they are very likely to contain the vast majority of the English words; and (2) they are publicly available. Both word embeddings had a dimensionality of 100 and were trained using a window size of 10, a minimum vocabulary count of 5, and 15 iterations. Additional parameters of word2vec were the negative sampling and the model type, which were set to 10 and continuous bag-of-words, respectively. These were also optimised through a random search on the validation set [34]. The embeddings of out-of-vocabulary words were determined by returning a zero-vector.

2.3. Evaluation metrics

We considered the available annotations in the the i2b2 2012 temporal relation corpus as the gold standard when evaluating the models. We used the official i2b2 evaluation script provided with the data. It uses standard evaluation methods in information retrieval, i.e., precision, recall, and F-score metrics for each relation type (i.e., BEFORE, AFTER, OVERLAP). Macro-average and micro-average of the F-scores were also obtained for every relation type. Macro-average is a per-class metric that computes the metric independently for each class and then take the average (hence treating all classes equally), whereas the micro-average favours classes with a larger number of instances and aggregates the contributions of all classes to compute the average metric. The overall score of the model is measured using the micro-average F-score to overcome class imbalance issue in the data.

3. Results

Table 2 shows a breakdown of the results of each model for both DCT and EVENTs/TIMEXs relations. Overall, the proposed Att-BiLSTM(GloVe)

Relation type	**Att-BiLSTM(GloVe)**			**Att-BiLSTM(MIMIC)**		
	P	**R**	**F-score**	**P**	**R**	**F-score**
AFTER	0.500	0.470	0.485	0.429	0.449	0.439
BEFORE	0.904	0.853	0.878	0.910	0.821	0.863
OVERLAP	0.755	0.839	0.795	0.728	0.836	0.779
Micro	0.811	0.811	0.811	0.791	0.791	0.791
Macro	0.720	0.721	0.719	0.689	0.702	0.694

Table 2: Performance of Att-BiLSTM(GloVe) and Att-BiLSTM(MIMIC) for extracting temporal relations on the official benchmark test set of the i2b2 2012 temporal relation corpus.

model yielded a micro-average F-score of 0.811 and Att-BiLSTM(MIMIC) model yielded a micro-average F-score of 0.791.

We compared the Att-BiLSTM results to six state-of-the-art methods evaluated on the i2b2 2012 temporal relation corpus:

- **SVM-CRF** [13] is the top performing system in the i2b2 2012 shared task. It is a hybrid method consisting of SVM and CRFs as classifiers, and rules to resolve conflict cases. It achieved an F-score of 0.695.

- **SVM-rules** [37] composes of multiple supervised machine learning models and rule-based methods to extract TLINKs. It is on par with the best systems (i.e., SVM-CRF [13]) on the i2b2 2012 corpus with an F-score of 0.695.

- **SVM-KNN-rules** [38] is also a hybrid approach to relation extraction. It uses a set of hand-crafted rules to determine the relation between two entities. If the relation cannot be classified by any of the rules, they classify it using SVM and K-Nearest Neighbor (KNN) machine learning classifiers. It achieved an F-score of 0.702.

- **1d-CNN-BERT** [39] is a neural networks model. It uses a pre-trained model instead of word embeddings as an input to a one-dimensional convolutional neural network (1d-CNN). Then it combines the 1d-CNN with Bidirectional Encoder Representations from Transformers (BERT) and uses the 1-dCNN to fine-tune the parameters of the BERT model. It achieved an F-score of 0.709.

Two recent state-of-the-art methods, which focus on direct temporal relations (i.e., relations between EVENTs and TIMEXs only) in the i2b2 2012 temporal relation corpus, were also compared to our Att-BiLSTM results.

- **re-SVM-CRF** [40] is a re-implemented version of **SVM-CRF** [13], that is the top performing system in the i2b2 2012 shared task. It is re-trained on temporal relations between EVENTs and TIMEXs only and achieved an F-score of 0.557.

- **SVM** [41] is another hybrid approach composing of an SVM-based system tailored to relations between EVENTs and TIMEXs, and deterministic rules to fix common errors observed during the development period. It achieved an F-score of 0.638.

Table 3 shows a performance comparison of our models with these state-of-the-art methods. With only word vectors entity position indicators and without using additional NLP sources to extract high-level features, the Att-BiLSTM models demonstrate improved performance.

Model	Feature set	DCT F-score	E/T F-score	Overall F-score
SVM-CRF [13]	Entity position, bag-of-words, part-of-speech, tense, dependency features, time attributes, event attributes, conjunction and distance.	-	-	0.695
SVM-rules [37]	Linguistic features, discourse features and semantic features.	-	-	0.695
SVM-KNN-rules [38]	Part-of-speech, event attributes, UMLS features, dependency features, special words, section ID, conjunction and distance.	-	-	0.702
1d-CNN-BERT [39]	BERT with general embeddings (GloVe) and PubMed embeddings.	-	-	0.709
SVM-CRF-re [40]	Entity position, bag-of-words, part-of-speech, tense, dependency features, time attributes, event attributes, conjunction and distance.	-	0.557*	-
SVM [41]	BERT with pre-trained model.	-	0.638*	-
Att-BiLSTM (GloVe)	Entity position indicators as input and general word embeddings (GloVe).	0.852+	0.706*	**0.811**
Att-BiLSTM (MIMIC)	Entity position indicators as input and word embeddings from the clinical domain (MIMIC).	0.893+	0.738*	0.791

E/T: EVENTs/TIMEXs relations,

*F-score obtained from training and testing the Att-BiLSTM on DCT relations only,

+F-score obtained from training and testing the Att-BiLSTM on EVENTs/TIMEXs relations only.

Table 3: Summary comparison of different models evaluated on the i2b2 2012 temporal relation test corpus.

4. Discussion

We found that the attention-based model, with only word embeddings, is an effective approach for extracting temporal links from clinical notes. Its performance (0.811 F-score) compete with complex features-based and neural networks state-of-the-art temporal relation classification systems. The improvement may come from the attention over multiple instances, which is expected to reduce the weights of those noisy instances dynamically. The general word embedding (GloVe) and the domain-specific word embeddings (MIMIC) work on par with each other. The clinical-domain specific representation (i.e., MIMIC) shows a slight advantage (0.893 and 0.738 F-score for DCT and EVENTs/TIMEXs relations respectively) over GloVe (0.852 and 0.706 F-score for DCT and EVENTs/TIMEXs relations respectively) when evaluated on individual relations (Table 3).

One of the reasons behind using the attention mechanism in this study was its ability to capture the most important parts of the text to identify temporal relations. In fact, this issue has been raised as a result of error analysis in a previous study: Nikfarjam et al. in [14] realised that, for many misclassified TLINKs, there were temporal trigger phrases in the text (e.g., "history of", "continued", "repeat", "consecutive", "subsequently") that might provide important features if modelled properly. We noticed that using the attention mechanism positively impacted the model's ability to overcome such an issue. To demonstrate this, we applied our best-performed model (Att-BiLSTM(GloVe)) to the following two examples, which were misclassified previously (in [14]):

- "He was given **D50**, but continued to have **progressive respiratory**

failure"

- "A bone marrow biopsy revealed the transformation of **his CMML** to **acute myelogenous leukemia**"

We found that the Att-BiLSTM(GloVe) was able to successfully classify the relation between "**D50**" and "**progressive respiratory failure**" as OVERLAP, and the relation between "**his CMML**" and "**acute myelogenous leukemia**" as BEFORE.

4.1. Error Analysis

4.1.1. Statistical significant difference

To gain insight into the errors made by the two temporal relation classification models (i.e., Att-BiLSTM with GloVe and with MIMIC) and to evaluate any statistical significance of the differences between them, we performed an error analysis by running paired sample t-tests, with the differences considered significant if the P-value was <0.05. We found that there is a statistically significant difference between the two models (P-value = 3.7e-06), with the model using GloVe performing better than the domain-specific embeddings. Thus, applying word embeddings trained on corpora in a general domain seems useful for this type of clinical narrative. This result is consistent with, but more general than, the conclusion drawn by Wang et al. in [42]. This suggests that a lack of access to a domain-specific corpus is not necessarily a barrier for the use of word embeddings in implementations for specific document types.

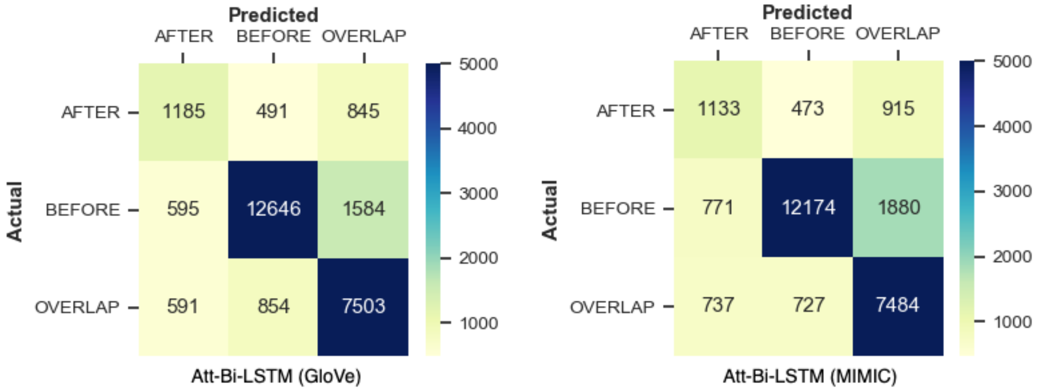

Figure 2: Confusion matrices for the performance of Att-BiLSTM models on the official benchmark test set of the i2b2 2012 temporal relation corpus.

4.1.2. Common classifications errors

We also constructed confusion matrices based on the gold standard and predicted relation types on the test set to determine the errors made by the models (see Figure 2). We found that in both models there are two common types of confusion that account for nearly 80% of the classification errors. These are the confusion between BEFORE and OVERLAP relations (accounting for 50% of the errors), and the confusion between AFTER and OVERLAP (accounting for 30% of the errors). Below we illustrate these types of confusion with examples.

Example (1):

> **Admission** Date: 07/10/1991
>
> Discharge Date: 07/18/1991
>
> The patient is an 85-year-old male with a history of **ischemic bowel status** post recent **admission** for **urosepsis** and **C. dif-**

> **ficile colitis**.

In this example, the system was able to correctly classify most of the TLINKs in the sentence, such as the relations between "ischemic bowel status" and "Admission" (BEFORE), "admission" and "the Admission" (BEFORE) and "admission" and "07/10/1991" (OVERLAP). However, the TLINK between "Admission" and the occurrence events "urosepsis" and "C. difficile colitis" was misclassified as OVERLAP, while the correct relation is BEFORE. It seems the confusion arises from both the mention of the word "admission" and the presence of the coordinating conjunction "and", which frequently appears together with the OVERLAP-ed events. In this example, determining whether the relationship should be BEFORE or OVERLAP requires understanding the narrative chains (including the disambiguation of two different admission references). One possible solution is that utilising phrase embeddings beside the word embeddings as phrases can be critical for capturing lexical meaning for many tasks [43].

Example (2):

> He was **able to communicate appropriately** as his level of **narcotic medications** waned in his blood.

In this example, event **"able to communicate appropriately"** happened AFTER the treatment event **"narcotic medications"**, but the relation is incorrectly identified as OVERLAP. The difficulty in correctly classifying this relation as AFTER arises from the fact that when events appear to occur simultaneously, they tend to have temporal synchronicity, and in this case, the entity type may not be important. However, for clinical events, when there is not temporal synchronicity (as in the above example), the entity type is of great significance. To illustrate that, in order to classify the relation type correctly, we may need to include the event type as a feature (for example, DRUG for treatment events, ADE for adverse drug events, TEST for clinical trials, etc.). This could give an understanding of the nature of the events as the relation cannot simply be inferred based on the expression pattern. If the system had the knowledge that once narcotic medications waned, a patient could communicate, then it could predict the right link type. Incorporating similar knowledge in NLP systems requires creating and incorporating comprehensive ontologies of clinical events is an ongoing research problem [44, 45, 46]. We also note that — if the treatment event has been recognised as **"narcotic medications waned in his blood"**, then the OVERLAP relation with **"able to communicate appropriately"** would be correct.

4.1.3. Training size effect on classifier

Training set characteristics such as class imbalance can significantly affect the performance of classifiers [47, 48]. The imbalanced data are characterised as having many more examples of certain classes than others. In such a case, classifiers tend to make a biased learning model that has a poorer predictive accuracy over the minority classes compared to the majority classes. Too few examples might result in low test accuracy, perhaps because the model overfits the training set or the training set is not sufficiently representative of the problem.

There are at least 30 points F-score difference between the AFTER and

	25%	**50%**	**75%**	**100%**
AFTER	0.358	0.433	0.467	0.485
BEFORE	0.863	0.849	0.880	0.878
OVERLAP	0.778	0.743	0.788	0.795
Overall	0.791	0.766	0.806	0.811

Table 4: Performance (Micro F-score) of Att-BiLSTM(GloVe) when trained on 25%, 50%, 75% and 100% per annotation type on the official benchmark test set of the i2b2 2012 temporal relation corpus.

other temporal relations (see Table 1). It is unclear if there is a relationship between the size and the F-score or not in the i2b2 2012 temporal relations dataset. One way to investigate this question is by evaluating the performance of the Att-BiLSTM(GloVe) model on training datasets of different size. We trained the model on variable sizes (25%, 50% and 75%) randomly selected of data, and we use the same test set for each different sized training dataset. Table 4 shows the performance of Att-BiLSTM(GloVe) when trained on variable sizes per class, and Figure 3 shows a line plot of the relationship between training set size and Att-BiLSTM(GloVe) model test set F-score. The line plot shows a slight improvement in the test accuracy as the training set increased for all classes, as we expect. The plot also shows small drops in the performances from 25% to 50% examples for all classes except the minority class, after which performances appear to level off. Thus, there is a direct relationship between training dataset size and F-score. When the training data set size increases, the model seems to learn more about the various temporal relations in the data.

Another approach for investigating the problem of class imbalance is

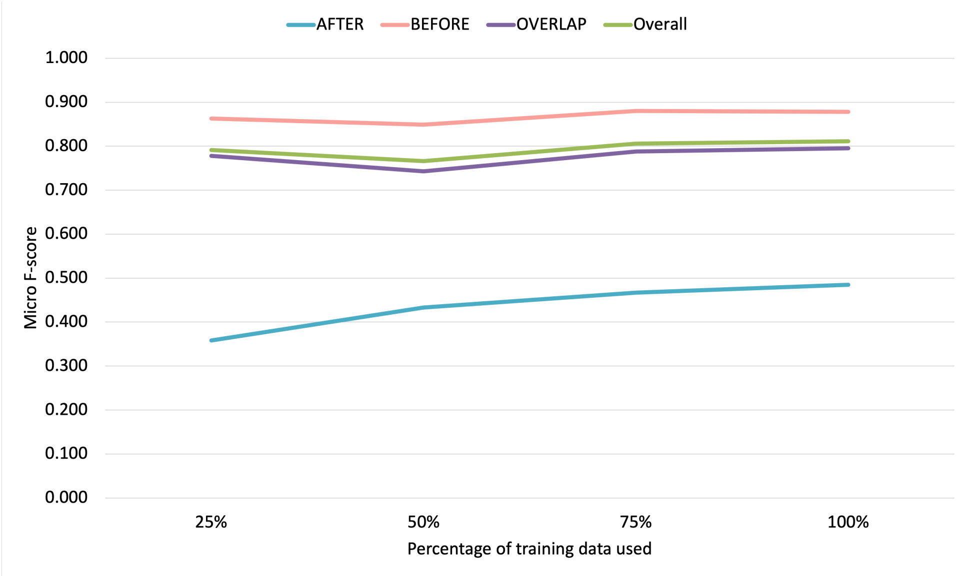

Figure 3: The relationship between training set size and Att-BiLSTM(GloVe) model test set F-score. 100% training set size refer to using all of the available training data.

to randomly re-sample the training dataset. The two main approaches to randomly re-sampling an imbalanced dataset are to delete examples from the majority class, called under-sampling, and to add examples to the minority class, called over-sampling. In order to under-sample the majority classes, we randomly extracted only 25% of the annotated examples from the BEFORE and the OVERLAP classes while keeping all available examples for the AFTER class. For over-sampling the minority class, we expanded the number of examples by using the transitive closure, which has proven to be effective in several studies [3, 15, 37, 49, 13, 50, 26, 51, 52, 53, 40]. Thus, we derived new implied AFTER relations from the existing labelled BEFORE relations. In other words, we copy all examples in the BEFORE class, we swap the relation entities, and then we change the relation from BEFORE to AFTER.

	Under-sampling (BEFORE and OVERLAP)	**Official dataset**	**Over-sampling (After)**
AFTER	0.200	0.485	0.389
BEFORE	0.381	0.878	0.857
OVERLAP	0.328	0.795	0.750
Overall	0.292	0.811	0.774

Table 5: Performance (Micro F-score) of Att-BiLSTM(GloVe) when applying different data sampling techniques.

Table 5 shows that under-sampling of the majority classes leads to significantly poorer results than the other data sampling techniques. This is most likely due to the fact that random under-sampling may lead to loss of vital information as some data points have been removed. Over-sampling with transitive closure performed better than the Under-sampling but not than the official dataset. This could be due to the fact that over-sampling has disturbed the data distribution (within the AFTER class) either by overfitting or by generating synthetic data points that do not follow the original class distribution as we have very little information about the minority class.

5. Conclusions

Temporal relation classification represents a special challenge for the field of clinical text analytics. The structure of clinical texts ranges from brief statements to long stories describing a patient's medical history, current condition, diagnostic analysis, and management plan. It is often impossible to interpret clinical texts without domain knowledge. It is even harder to extract and classify temporal relationships. Still, there have been several attempts to extract and classify temporal relation from clinical text. We have explored a neural network model (Att-BiLSTM) based on the attention mechanism that is integrated into a Bidirectional Long Short-Term Memory Network. The model does not rely on specific NLP pre-processing and uses raw text with entity position indicators as input, alongside word embeddings that have been generated either from a generic or domain-specific corpus. We demonstrate the effectiveness of the proposed method by evaluating it on 140 discharge summaries from the i2b2 2012 temporal relation corpus. The model achieved an F-score of 0.811, which is at least 10% better than state-of-the-art in the field. We show that the neural attention model can be remarkably effective at classifying temporal relations when provided with word embeddings trained from corpora in a general domain. Furthermore, we perform an extensive error analysis to gain insight into the errors made by the models.

As future work, we plan to investigate how different semantic features and ontologies contribute to the performance. We will also consider exploring variations of the attention mechanisms such as multi-head attention [54] and self-attention [55, 56], and other state-of-the-art text mining approaches that have not been used for TLINKs yet, including, for instance, clinical contextual models (e.g., BioBERT, clinicalBERT, etc.), bootstrapping, distance supervision [57] and minwise hashing [58].

Acknowledgements

This project is partially supported by the Saudi Arabian Ministry of education, the Saudi Arabian cultural bureau in London and the Healthcare Text Analytics Network (Healtex, grant EP/N027280/1, funded by UK EPSRC).

References

[1] D. B. C. L. D. of Health, A guide to the national programme for information technology., 2005. URL: https://web.archive.org/web/20051026213141/http://www.connectingforhealth.nhs.uk/all_images_and_docs/NPfIT%20brochure%20Apr%2005%20final.pdf.

[2] L. Zhou, G. Hripcsak, Temporal reasoning with medical data—a review with emphasis on medical natural language processing, Journal of biomedical informatics 40 (2007) 183–202. doi:10.1016/j.jbi.2006.12.009.

[3] G. Alfattni, N. Peek, G. Nenadic, Extraction of temporal relations from clinical free text: A systematic review of current approaches, Journal of Biomedical Informatics 108 (2020) 103488. URL: http://www.sciencedirect.com/science/article/pii/S1532046420301167. doi:https://doi.org/10.1016/j.jbi.2020.103488.

[4] W. Sun, A. Rumshisky, O. Uzuner, Annotating temporal information in clinical narratives, Journal of biomedical informatics 46 (2013) S5–s12. doi:10.1016/j.jbi.2013.07.004.

[5] S. Bethard, L. Derczynski, G. Savova, J. Pustejovsky, M. Verhagen, Semeval-2015 task 6: Clinical tempeval, in: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), 2015, pp. 806–814. doi:10.18653/v1/s15-2136.

[6] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, M. Verhagen, Semeval-2016 task 12: Clinical tempeval, in: Proceedings

of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1052–1062. doi:10.18653/v1/s16-1165.

[7] S. Bethard, G. Savova, M. Palmer, J. Pustejovsky, SemEval-2017 task 12: Clinical TempEval, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 565–572. URL: https://www.aclweb.org/anthology/S17-2093. doi:10.18653/v1/S17-2093.

[8] R. Gaizauskas, H. Harkema, M. Hepple, A. Setzer, Task-oriented extraction of temporal information: The case of clinical narratives, in: Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06), volume 2006, Ieee, 2006, pp. 188–195. doi:10.1109/time.2006.27.

[9] L. Zhou, S. Parsons, G. Hripcsak, The evaluation of a temporal reasoning system in processing clinical discharge summaries, Journal of the American Medical Informatics Association 15 (2008) 99–106. doi:10.1197/jamia.M2467.

[10] Y.-L. Yang, P.-T. Lai, R. T.-H. Tsai, A hybrid system for temporal relation extraction from discharge summaries, in: International Conference on Technologies and Applications of Artificial Intelligence, volume 8916, Springer, 2014, pp. 379–386. doi:10.1007/978-3-319-13987-6_35.

[11] E. P. Hernandez, A. P. Quimbaya, O. M. Munoz, Htl model: A model for extracting and visualizing medical events from narrative text in

electronic health records., in: ICT4AgeingWell, 2016, pp. 107–114. doi:10.1109/ColumbianCC.2016.7750768.

[12] A. A. Abdulsalam, S. Velupillai, S. Meystre, Utahbmi at semeval-2016 task 12: extracting temporal information from clinical text, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1256–1262. doi:10.18653/v1/s16-1195.

[13] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, H. Xu, A hybrid system for temporal information extraction from clinical text, Journal of the American Medical Informatics Association 20 (2013) 828–835. doi:10.1136/amiajnl-2013-001635.

[14] A. Nikfarjam, E. Emadzadeh, G. Gonzalez, Towards generating a patient's timeline: extracting temporal relationships from clinical notes, Journal of biomedical informatics 46 (2013) S40–s47. doi:10.1016/j.jbi.2013.11.001.

[15] H.-J. Lee, H. Xu, J. Wang, Y. Zhang, S. Moon, J. Xu, Y. Wu, Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1292–1297. doi:10.18653/v1/s16-1201.

[16] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: Proceedings of the 54th Annual Meeting of the Association for

Computational Linguistics (Volume 2: Short Papers), volume 2, 2016, pp. 207–212. doi:10.18653/v1/p16-2034.

[17] N. Bach, S. Badaskar, A review of relation extraction, Literature review for Language and Statistics II 2 (2007) 15. URL: https://www.researchgate.net/publication/265006408{%}0A{%}5C{%}5C.

[18] D. Dligach, T. Miller, C. Lin, S. Bethard, G. Savova, Neural temporal relation extraction, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, volume 2, 2017, pp. 746–751. doi:10.18653/v1/e17-2118.

[19] B. Waltl, G. Bonczek, F. Matthes, Rule-based information extraction: advantages, limitations, and perspectives, Jusletter IT (02 2018) (2018).

[20] C. Lin, T. Miller, D. Dligach, S. Bethard, G. Savova, Representations of time expressions for temporal relation extraction with convolutional neural networks, in: BioNLP 2017, 2017, pp. 322–327. doi:10.18653/v1/w17-2341.

[21] P. Li, H. Huang, Uta dlnlp at semeval-2016 task 12: deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1268–1273. doi:10.18653/v1/s16-1197.

[22] V. R. Chikka, Cde-iiith at semeval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques,

in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1237–1240. doi:10.18653/v1/s16-1192.

[23] D. Galvan, N. Okazaki, K. Matsuda, K. Inui, Investigating the challenges of temporal relation extraction from clinical text, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018, pp. 55–64. doi:10.18653/v1/w18-5607.

[24] Y. Long, Z. Li, X. Wang, C. Li, Xjnlp at semeval-2017 task 12: Clinical temporal information ex-traction with a hybrid model, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 1014–1018. doi:10.18653/v1/s17-2178.

[25] J. Tourille, O. Ferret, X. Tannier, A. Neveol, Limsi-cot at semeval-2017 task 12: Neural architecture for temporal information extraction from clinical narratives, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), 2017, pp. 597–602. doi:10.18653/v1/s17-2098.

[26] C. Lin, T. Miller, D. Dligach, H. Amiri, S. Bethard, G. Savova, Self-training improves recurrent neural networks performance for temporal relation extraction, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018, pp. 165–176. doi:10.18653/v1/w18-5619.

[27] J. Tourille, O. Ferret, A. Neveol, X. Tannier, 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, 2017, pp. 224–230. doi:10.18653/v1/P17-2035.

[28] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[29] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 4, 2016, pp. 2124–2133. doi:10.18653/v1/p16-1200.

[30] X. Huang, et al., Attention-based convolutional neural network for semantic relation extraction, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 2526–2536.

[31] S. Liu, L. Wang, V. Chaudhary, H. Liu, Attention neural model for temporal relation extraction, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 134–139.

[32] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850 (2013). URL: http://arxiv.org/abs/1308.0850. arXiv:1308.0850.

[33] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 challenge, Journal of the American Medical Informatics Association 20 (2013) 806–813. doi:10.1136/amiajnl-2013-001628.

[34] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of machine learning research 13 (2012) 281–305.

[35] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701 (2012). URL: http://arxiv.org/abs/1212.5701. arXiv:1212.5701.

[36] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.

[37] C. Lin, D. Dligach, T. A. Miller, S. Bethard, G. K. Savova, Multilayered temporal modeling for the clinical domain, Journal of the American Medical Informatics Association 23 (2015) 387–395. doi:10.1093/jamia/ocv113.

[38] J. D'Souza, V. Ng, Knowledge-rich temporal relation identification and classification in clinical notes, Database 2014 (2014). doi:10.1093/database/bau109.

[39] T. Chen, M. Wu, H. Li, A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning, Database 2019 (2019).

[40] H.-J. Lee, Y. Zhang, J. Xu, C. Tao, H. Xu, M. Jiang, Towards practical temporal relation extraction from clinical notes: an analysis of direct temporal relations, in: 2017 IEEE International Conference on Bioin-

formatics and Biomedicine (BIBM), volume 2017-Janua, Ieee, 2017, pp. 1272–1275. doi:10.1109/bibm.2017.8217842.

[41] H.-J. Lee, Y. Zhang, M. Jiang, J. Xu, C. Tao, H. Xu, Identifying direct temporal relations between time and events from clinical notes, BMC medical informatics and decision making 18 (2018) 49. doi:10.1186/s12911-018-0627-5.

[42] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, Journal of biomedical informatics 87 (2018) 12–20.

[43] Y. Wu, S. Zhao, W. Li, Phrase2vec: phrase embedding based on parsing, Information Sciences 517 (2020) 100–127.

[44] C. Tao, W.-Q. Wei, H. R. Solbrig, G. Savova, C. G. Chute, Cntro: a semantic web ontology for temporal relation inferencing in clinical narratives, in: AMIA annual symposium proceedings, volume 2010, American Medical Informatics Association, 2010, p. 787.

[45] C. Tao, H. R. Solbrig, C. G. Chute, Cntro 2.0: a harmonized semantic web ontology for temporal relation inferencing in clinical narratives, AMIA summits on translational science proceedings 2011 (2011) 64.

[46] F. Li, J. Du, Y. He, H.-Y. Song, M. Madkour, G. Rao, Y. Xiang, Y. Luo, H. W. Chen, S. Liu, et al., Time event ontology (teo): to support semantic representation and reasoning of complex temporal relations of

clinical events, Journal of the American Medical Informatics Association 27 (2020) 1046–1056.

[47] G. Foody, M. McCulloch, W. Yates, The effect of training set size and composition on artificial neural network classification, International Journal of Remote Sensing 16 (1995) 1707–1723.

[48] W. Zheng, M. Jin, The effects of class imbalance and training data size on classifier learning: an empirical study, SN Computer Science 1 (2020) 1–13.

[49] S. Jeblee, G. Hirst, Listwise temporal ordering of events in clinical notes, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018, pp. 177–182. doi:10.18653/v1/w18-5620.

[50] C. Cherry, X. Zhu, J. Martin, B. de Bruijn, A la recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 nlp challenge, Journal of the American Medical Informatics Association 20 (2013) 843–848. doi:10.1136/amiajnl-2013-001624.

[51] Y. Xu, Y. Wang, T. Liu, J. Tsujii, E. I.-C. Chang, An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge, Journal of the American Medical Informatics Association 20 (2013) 849–858. doi:10.1136/amiajnl-2012-001607.

[52] C. Grouin, N. Grabar, T. Hamon, S. Rosset, X. Tannier, P. Zweigenbaum, Eventual situations for timeline extraction from clinical reports, Journal

of the American Medical Informatics Association 20 (2013) 820–827. doi:10.1136/amiajnl-2013-001627.

[53] Y. Cheng, P. Anick, P. Hong, N. Xue, Temporal relation discovery between events and temporal expressions identified in clinical narrative, Journal of biomedical informatics 46 (2013) S48–S53.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017) 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[55] P. Verga, E. Strubell, A. McCallum, Simultaneously self-attending to all mentions for full-abstract biological relation extraction, arXiv preprint arXiv:1802.10569 (2018).

[56] J. Cheng, L. Dong, M. Lapata, Long short-term memory-networks for machine reading, arXiv preprint arXiv:1601.06733 (2016).

[57] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, Association for Computational Linguistics, 2009, pp. 1003–1011. doi:10.3115/1690219.1690287.

[58] D. S. Batista, R. Silva, B. Martins, M. J. Silva, A minwise hashing method for addressing relationship extraction from text, in: International

Conference on Web Information Systems Engineering, volume 8181 Lncs, Springer, 2013, pp. 216–230. doi:10.1007/978-3-642-41154-0_16.