A temporal expression recognition system for medical documents by taking help of news domain corpora

Naman Gupta[1] **Aditya Joshi**[1,2,3]
Pushpak Bhattacharyya[1]
[1]IIT Bombay, India, [2]Monash University, Australia
[3]IITB-Monash Research Academy, India
{adityaj,pb}@cse.iitb.ac.in
namanbbps@gmail.com

Abstract

A bottleneck for medical domain Temporal Expression Recognition (TER) is the availability of data. An open-domain TER system may not be able to capture domain-specific expressions, while domain-specific TER may be cumbersome to implement. We present a novel neural network based medical TER system that uses corpora from news and medical domains. Thus, it serves as a middle ground between an open-domain and a domain-specific TER. We show that our system outperforms state-of-art open-domain baselines, and gets close to domain-specific skylines. Thus, our system proves to be a promising alternative for domain specific TER for domains where data may be limited.

1 Introduction

Temporal Expression Recognition (TER) is the process of locating phrases that denote temporal information. Temporal expressions may be an expressed point in time, a duration or a frequency (Wikipedia, 2014). These expressions can be used in information extraction and question-answering to (a) answer time-specific queries, (b) arrange information in a chronological manner, etc. Early work in TER considers it as a part of named entity recognition (Bikel et al., 1999). TER, as a separate task, was introduced as Temporal Expression Recognition and Normalization (TERN). TER in general domain has been widely studied. Rule-based methods for specific domains were adopted by popular systems like Heideltime (Strötgen and Gertz, 2010), SUTime (Chang and Manning, 2012), MayoTime (Sohn et al., 2013). The rules were regular expressions over word or tokens. Supervised Classifiers like SVM, CRF using linguistic features have been explored (Adafre and de Rijke, 2005; Bethard, 2013). Joint inference-based classifiers like Markov Logic have also been reported (UzZaman and Allen, 2010). Medical domain TER (Sun et al., 2013; Bethard et al., 2015) has resulted in alternate methods and systems for detecting and normalizing temporal expressions. Our system uses a neural network based architecture which has hitherto not been used for TER. In addition, we also deal with a specific situation: In-domain data being difficult to obtain. Research in TER mostly deals with news domain text, arguably because of availability of large corpora and abundance of temporal expressions in news documents. In recent times, TER has also been applied to other domains like medical. Approaches for medical domain TER in the past have been either rule-based (Sohn et al., 2013; Jindal and Roth, 2013), statistical (Xu et al., 2013; Roberts et al., 2013) or hybrid (Lin et al., 2013).

However, a bottleneck for medical domain TER is the **availability of data**. Medical documents such as discharge summaries are of classified nature, and also must be de-identified (*i.e.*, anonymized) before being used. **Our paper is motivated by this limitation**. An open-domain TER system (*i.e.*, a TER not learned from medical domain data) may not be able to capture domain-specific expressions (for example, Latin acronyms like *bid, tid* that are used in medical documents). On the other hand, a domain-specific TER system is time-consuming to construct[1].We address the question:

Can a TER system that uses documents of two domains serve as a middle ground between an open-domain and a domain-specific TER, in case domain-specific data is difficult to obtain ?

The novelty of our work lies in: (a) A simple yet effective neural network based architecture for

[1]This holds irrespective of whether it is rule-based or statistical.

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 84–88, Trivandrum, India. December 2015.

TER, (b) Use of a combination of open-domain and domain-specific data. Thus, **our TER system combines information from medical and news domain, to perform TER of medical documents.** In the rest of the paper, we refer to news as out-of-domain corpora, and medical documents as in-domain corpora.

2 Our System: Neural Network based TER

In the past, TER has been modeled either as a sequence labeling (Bethard, 2013) or a classification task (Tissot et al., 2015). We choose the latter design. Our model takes as input a word and outputs the most probable tag. We have used 9 tags, namely *B-DATE, B-DURATION, B-FREQUENCY, B-TIME, I-DATE, I-DURATION, I-FREQUENCY, I-TIME and O*. For a temporal expression, B,I,O indicate beginning, inside and outside respectively.

Our three-layer neural network model is shown in Figure 1. It makes use of vector representation of words. Mikolov et al. (2013) proposed a computationally efficient method for learning distributed word representation such that words with similar meanings will map to similar vectors. We use the same approach for learning word vectors using word2vec (`https://code.google.com/p/word2vec/`). Table 1 shows nearest neighbors for four sample words that are commonly used as temporal expressions. We then create a *lookup table* $LT \in \mathbb{R}^{|C| \times d}$ to store a d-dimensional representation of every word in vocabulary C.

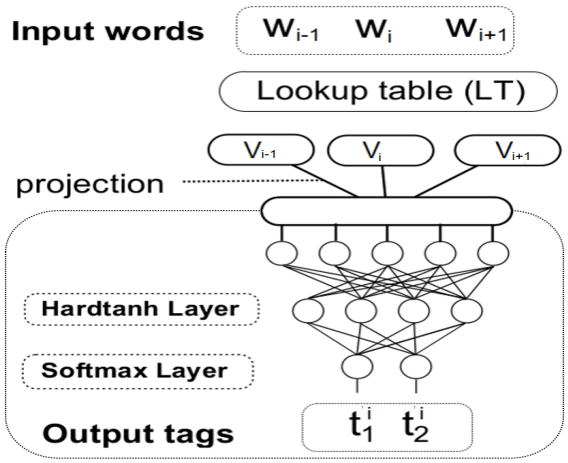

Figure 1: Our Neural network-based architecture

A neural network is trained with a word as a data unit. First, the size of a context window w_s is chosen. The word along with its context words forms a n-gram sequence S represented as $\{W_{i-1}, W_i, W_{i+1}\}$ in the network. Every word W_i is mapped to its corresponding word vector V_i using lookup table LT. Word vectors V_i's are projected onto the input layer. In order to preserve word order, projection concatenates the word vectors into single vector $X \in \mathbb{R}^{|w_s * d|}$ which are passed to non-linear *hardtanh* layer.

After applying the $hardtanh$ transformation, we get the H as the output of the hidden layer

$$H = hardtanh(W_1^T X + b_1) \quad (1)$$

We then transform the output of the hidden layer using a *softmax* layer.

$$O = W_2^T H + b_2 \quad (2)$$

The output layer $O \in \mathbb{R}^t$ has the dimensionality of number of tags t. Errors on cost are back propagated into the network using back-propagation algorithm (Russell and Norvig, 1967) to generate probability distribution over output tags t_1^i, t_2^i.

qd	postop	admission	tuesday
tid	post-operative	transfer	sunday
qid	post-op	discharge	saturday
qday	hospital	admit	monday
qam	day	preoperative	march
daily	life	0/0/0	august
qhs	number	report	thursday
q0s	post-day	summary	january

Table 1: Nearest neighbors for sample temporal expressions

3 Experimental Setup

We evaluate our systems in three settings: (1) **Overlap**, where overlapping spans are considered as match, (2) **Exact**, where precise matches are counted, and (3) **Partial**, where full credit is awarded for exact match, and half credit for overlapping match. All the systems (baseline, skyline and our system) are tested on a publicly available dataset from the i2b2 2012 Temporal Relation Challenge (Sun et al., 2013). This is a benchmark dataset, and consists of 120 discharge summaries from Partners Healthcare and Beth Israel Deaconess Medical Center.

Type	Sentences	Tokens	Vocabulary
Medical	20,125	481,601	12,142
News	24.445	717,698	30,527

Table 2: Statistics of the datasets

	Overlap			Partial			Exact		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
BL - SUTime (RB)	73.53	74.78	**74.15**	62.75	63.82	**63.28**	51.97	52.86	**52.41**
BL - Heideltime (RB)	79.92	56.43	**66.15**	72.10	50.91	**59.68**	64.28	45.38	**53.20**
BL - ClearTk (ST)	44.36	19.34	**26.94**	34.41	15.05	**20.95**	24.46	10.77	**14.95**
SL-Rule	87.91	92.25	**90.02**	79.86	83.98	**81.87**	71.81	75.71	**73.71**
SL-Stat	95.13	83.74	**89.07**	89.11	78.46	**83.45**	83.09	73.19	**77.83**
Our System: News	86.18	77.36	**81.53**	74.28	66.70	**70.29**	62.39	56.04	**59.04**
Our System: News + Medical	81.37	86.98	**84.08**	73.19	78.10	**75.57**	65.02	69.23	**67.06**

Table 3: Comparison of our system with baseline (BL-*) and skyline (SL-*) systems

3.1 Datasets

The datasets used for training word vectors were created as follows. The statistics are shown in Table 2.

1. **In-domain dataset**: Medical discharge summaries are collected from i2b2. The documents are pre-processed by removing markup tags and irrelevant information in the form of document numbers and codes.

2. **Out-of-domain dataset**: Out-of-domain word vectors are learned from Timebank, AQUAINT (Pustejovsky et al., 2003), and TE-3 silver dataset (UzZaman et al., 2012)

3.2 Baseline: Open-domain TER

Rule-based temporal taggers like Heideltime (Strötgen and Gertz, 2010) and SUTime (Chang and Manning, 2012) and Statistical tagger like ClearTk[2] were developed as a part of TempEval-2,3 challenges for news text. They are our baselines: BL-SUTime, BL-Heideltime, and BL-ClearTk.

3.3 Skyline: Medical TER

State-of-art rule-based (Sohn et al., 2013) and a statistical (Roberts et al., 2013) medical domain TER systems are chosen as skylines. These system (indicated by SL-Rule and SL-Stat respectively) were developed as a part of i2b2 2012 challenge, and trained on medical data. We call them as skyline because the availability of medical data itself is the best situation for medical TER.

[2] https://code.google.com/p/cleartk/wiki/ClearTKTimeML

4 Results

We now compare our results of our system with the existing systems, and then describe how proportion of in-domain data impacts the performance. Finally, we discuss a detailed error analysis.

4.1 Comparative performance against baseline and skyline

Table 3 compares the performance of the baseline **(BL-*)** (first three rows) and skyline **(SL-*)** approaches (next two rows) with our system, for overlap, partial and exact matches. For our system, we experiment with two settings: (1) Word vectors trained on out-of-domain dataset (indicated by **Our System: News**), and (2) Word vectors trained on both datasets (indicated by **Our System: News + Medical**). In case of overlap match, the best performance of baseline systems is 74.15% in case of BL-SUTime. When neural network architecture is used **even in absence of any in-domain data**, the F-score increases to 81.53%. This value rises to 84.08% when a combination of medical and news domain is used.

SL-Rule and SL-Stat were created for medical TER. In case of our premise, medical data is difficult to obtain. Our system shows that by mixing out-of-domain (news) data with in-domain (medical) data, we can get close to the skyline performance. This degradation in performance is likely to be because of the small size of medical domain corpora available for training word vectors. It must be noted that while our dataset has 12,142 (as shown in Table 2) unique tokens, the corresponding value is usually much higher. For example, the pre-trained vectors trained on Google

News dataset had a vocabulary size of 300K[3].

4.2 Impact of proportion of in-domain data

Availability of in-domain data is restricted for medical documents as compared to news text. To find how our system peforms if a combination of in-domain and out-of-domain data is to be used, we conduct experiments by incrementally adding in-domain data to learn word vectors. Figure 2 plots the F-score against the percentage of total in-domain data used during training. 10 on the X-axis indicates that 10% of the total available medical domain data (along with the complete news data) was used during training. For all three kinds of matches, the F-score stabilizes beyond 40% (which is 5K sentences). There is a dip in performance for all three matches when 10% medical data is used. This may be due to dilution of word vectors, since only a small portion of a new domain has been added to the training set.

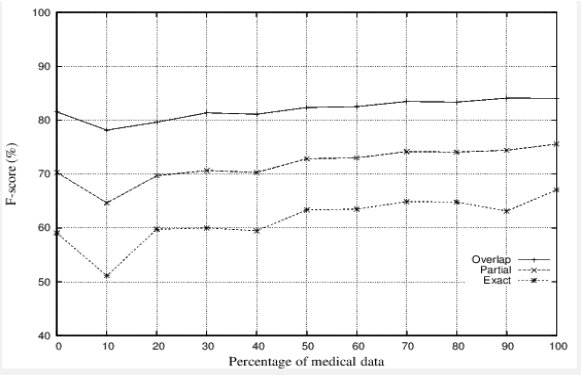

Figure 2: Performance of our system on addition of in-domain data

4.3 Error Analysis

We manually labeled 468 erroneous instances into one out of 11 broad categories. The distribution of these errors is shown in Figure 3.

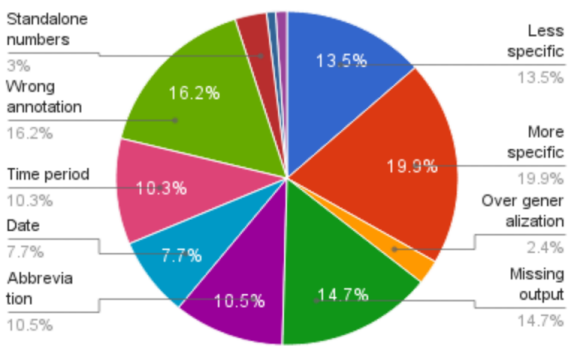

Figure 3: Distribution of errors made by our system

[3]https://code.google.com/p/word2vec/

'More specific' errors form 19.9% of total errors. This means that our system could extract temporal expressions that were more specific than the annotations. For example, our system tags '*one day prior to admission*' when only '*one day*' was expected. 16.2% of our errors arise due to wrong annotations. 'More specific' output is acceptable. 'Wrong annotations' are the ones where a second manual check revealed that the labels were disputable. Thus, **36% errors do not directly point to deficiencies in the system**.

'Less specific' errors (13.5%) are when our system leaves out a part of the temporal expression. 10.3% of errors are related to time periods and frequency-related words like '*daily*'. Over-generalization errors are said to occur when an extracted time expression contains some non-time-related words. Medical domain text is fraught with abbreviations (such as q.i.d.) leading to 10.5% errors, and peculiar date formats (for example, 12-20 as a date indicates 20th December) leading to 7.7% errors. 'Special' words are related to seasons and events like Halloween. Standalone numbers indicate situations like the '2' in 'see you at 2'. The 'Others' category includes errors due to garbled characters, relative days ('tomorrow'), ordinal numbers and WSD errors (two senses of 'may' can be derived out of 'this may', in absence of capitalized 'M')

5 Conclusion & Future Work

We presented a simple yet effective three-layer neural network based TER system for medical domain. Our system used out-of-domain news text to extract temporal expressions from medical documents. Our system, without any in-domain data at all, improves the F-score by 7% over three baseline systems, and to a greater degree when in-domain data is used. With a dataset of 5K medical domain sentences, we obtain a good performance . Our error analysis showed that the top three kinds of errors are: 'More specific output', 'Wrong annotations' and 'Missing output'. We, thus, show that our TER system can act as a middle ground between an open-domain and a domain-specific TER, in situations where in-domain data is difficult to obtain. A possible future work is to model TER as a sequence labeling task while still using a neural network based system.

References

Sisay Fissaha Adafre and Maarten de Rijke. 2005. Feature engineering and post-processing for temporal expression recognition using conditional random fields. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 9–16. Association for Computational Linguistics.

Steven Bethard, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics*.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 10–14.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns whats in a name. *Mach. Learn.*, 34(1-3):211–231, feb.

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, pages 3735–3740.

Prateek Jindal and Dan Roth. 2013. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics*, 46:S13–S19.

Yu-Kai Lin, Hsinchun Chen, and Randall A Brown. 2013. Medtime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46:S20–S28.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Kirk Roberts, Bryan Rink, and Sanda M Harabagiu. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, 20(5):867–875.

S Russell and P Norvig. 1967. The most popular method for learning in multilayer networks is called back-propagation. *Artif. Intel. Mod. Approach*, pages 201–218.

Sunghwan Sohn, Kavishwar B Wagholikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association*, 20(5):836–842.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics.

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.

Hegler Tissot, Cel Franc H dos Santos, Genevieve Gorrell, Angus Roberts, Leon Derczynski, and Marcos Didonet Del Fabro. 2015. Ufprsheffield: Contrasting rule-based and support vector machine approaches to time expression identification in clinical tempeval.

Naushad UzZaman and James F Allen. 2010. Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283. 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*.

Wikipedia. 2014. Temporal expressions — wikipedia, the free encyclopedia. [Online; accessed 12-June-2015].

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, I Eric, and Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):849–858.