Temporal Relation Extraction Using Expectation Maximization

Seyed Abolghasem Mirroshandel
Sharif University or Technology, Iran
mirroshandel@ce.sharif.edu

Gholamreza Ghassem-Sani
Sharif University or Technology, Iran
sani@sharif.edu

Abstract

The ability to accurately determine temporal relations between events is an important task for several natural language processing applications such as Question Answering, Summarization, and Information Extraction. Since current supervised methods require large corpora, which for many languages do not exist, we have focused our attention on approaches with less supervision as much as possible. This paper presents a fully generative model for temporal relation extraction based on the expectation maximization (EM) algorithm. Our experiments show that the performance of the proposed algorithm, regarding its little supervision, is considerable in temporal relation learning.

1 Introduction

Lately, the increasing attention to the practical NLP applications such as question answering, information extraction, and summarization have resulted in a growing demand of temporal information processing (Tatu and Srikanth, 2008). In question answering, one may expect the system to answer questions such as *"when an event occurred"*, or *"what is the chronological order of some desired events"*. In text summarization, especially in the multi-document type, knowing the order of events is a useful source of correctly merging related information.

Unlike problems such as part-of-speech tagging, morphological analysis, parsing, and named entity recognition which have been recently addressed with satisfactory results by combining statistical and symbolic methods (Mani et al., 2006), temporal relation extraction that requires deeper semantic analysis are yet to be worked on. One of recent efforts has disclosed that this task is a complicated task, even for human annotators (Mani et al., 2006).

Based on the type of corpora that different temporal relation learning methods use, these methods are divided into three major categories: supervised, semi-supervised, and unsupervised. Supervised methods normally rely on the correct temporal relations of training sentences of a manually tagged corpus. Semi-supervised methods often rely on a partially tagged corpus and need less supervision. Finally, unsupervised methods rely only on raw sentences without any temporal relation annotation. It is obvious that producing the necessary training data (corpora) of supervised and to a less extent semi-supervised methods is a time consuming, hard, and expensive work. Besides, it is very difficult to adapt such methods for new tasks, languages, and/or domains. Consequently, it is in fact the corpus availability that directs the research in this area. For mentioned reasons, we have focused on unsupervised and weakly supervised temporal relation learning.

This paper presents a novel usage of expectation maximization (EM) algorithm for temporal relation learning. The algorithm also employs Allen's interval algebra (Allen, 1984). Our experiments show that the performance of the proposed algorithm is acceptable with respect to little usage of tagged corpora which is used.

The remainder of the paper is organized as follows: section 2 is about previous works on temporal relation extraction. Section 3 explains our proposed method. Section 4 briefly presents the characteristic of the corpora that we have used. Section 5 demonstrates the evaluation of the proposed algorithm. Finally, section 6 includes our conclusions and some possible future works.

2 Temporal Relation Extraction

For a given ordered pair of components *(x_1, x_2)*, where x_1 and x_2 are times and/or events, a

Proceedings of Recent Advances in Natural Language Processing, pages 218–225,
Hissar, Bulgaria, 12-14 September 2011.

temporal information processing system identifies the type of relation that temporally links x_1 to x_2. The relation type can for instance be one of the 14 types proposed in TimeML (Pustejovsky et al., 2003). For example, in *"If all the debt is converted (e_7) to common, Automatic Data will issue (e_8) about 3.6 million shares; last Monday (t_{24}), the company had (e_{25}) nearly 73 million shares outstanding."*, taken from document *wsj_0541* of TimeBank (Pustejovsky et al., 2003), there are two temporal relations between pairs *(e_7, e_8)* and *(t_{24}, e_{25})*. The task of temporal relation extraction is to automatically tag these pairs respectively with the *BEFORE* and *INCLUDES* relations.

2.1 Related Work

There are numerous ongoing researches focused on temporal relation extraction. Existing methods of temporal relation learning, which are mainly fully supervised, can be divided into three categories: 1) Pattern based; 2) Rule based, and 3) Anchor based. These categories are respectively discussed in the next three sub-sections.

Pattern Based Methods

Pattern based methods extract some generic lexico-syntactic patterns for events co-occurrence. Extracting such patterns can be done manually or automatically.

Perhaps the simplest pattern based method is the one that was developed using a knowledge resource called VerbOcean (Chklovski and Pantel, 2005). VerbOcean has a small number of manually selected generic patterns. The style of patterns is in the form of <Verb-X> and then <Verb-Y>. Similar to other manual methods, a major drawback of this method is its tendency to have a high recall but a low precision. Several heuristics have been proposed to resolve the low precision problem (Chklovski and Pantel, 2005; Torisawa, 2006).

On the other hand, automatic methods try to learn a classifier from an annotated corpus, and attempt to improve classification accuracy by feature engineering. MaxEnt classifier is an example of this group (Mani et al., 2006). The state of the art of supervised methods in this group is very similar to the MaxEnt classifier (Chambers et al., 2007). This classifier tries to learn event attributes and event-event features in two consecutive stages. It also uses WordNet to find words' synsets.

Some of researches on pattern based temporal relation classification only work on corpora with specific characteristics, rather than general corpora such as TimeBank (Bethard and Martin, 2008; Bethard et al, 2007a; Lapata and Lascarides 2006; Bethard et al, 2007b; Bethard, 2007). There are also algorithms that work on only limited types of relations (Lapata and Lascarides 2006; Bethard, 2007; Bethard and Martin, 2007; Chambers and Jurafsky, 2008).

In another work, a weakly-supervised algorithm was proposed to classify temporal relation between events (Mirroshandel and Ghassem-Sani, 2010). In that work, it was shown that by applying a bootstrapping technique to some unlabeled documents that were related to the test documents and without any additional annotated data, temporal relations can be classified with satisfactory results.

Rule Based Methods

The common idea behind rule based methods is to design a number of rules for classifying temporal relations. In most existing works, these rules, which are manually defined, are based on Allen's interval algebra (Allen, 1984). One usage of these rules is enlarging the training set (Mani et al., 2006). Reasoning about the certainty of predicted temporal relations is the other utilization of these rules.

Anchor Based Methods

Anchor based methods use information of argument fillers (called anchors) of every event expression as a valuable clue for recognizing temporal relations. These methods rely on the distributional hypothesis (Harris, 1968), and by looking at a set of event expressions whose argument fillers have a similar distribution, try to recognize synonymous event expressions. Algorithms such as DIRT (Lin and Pantel, 2001), TE/ASE (Szpektor et al., 2004), and that of Pekar's system (Pekar, 2006) are examples of anchor based methods.

3 Using EM for Temporal Relation Learning

Due to appropriate results of the expectation maximization (EM) algorithm in some unsupervised tasks of natural language processing such as unsupervised grammar induction (Klein, 2005), unsupervised anaphora resolution (Cherry and Bergsma, 2005; Charniak and Elsner, 2009), and unsupervised coreference resolution (Ng, 2008), we decided to apply EM to temporal relation extraction. Currently, there is no reported work in temporal relation extraction based on EM. Here, we explain how EM can be successfully applied to the task of temporal relation extraction and show that the results are notable in this task. Before that, we first introduce definitions and notations that will be later used in subsequent sections.

3.1 Definitions

In temporal relation learning, system must be able to determine temporal relation *r* between two events e_1 and e_2. Here, we assume that events are annotated and the learner must find out the relation type *r*. In general, the relation type can be one of the 14 types proposed in TimeML (Pustejovsky et al., 2003) plus relation *NONE* (which indicates there is no temporal relation between respected pair of events). In this paper, *context* means the sentence (or sentences) containing pairs of examined events.

3.2 The Model

The proposed algorithm operates at the corpus level, inducing valid temporal clustering for all event pairs of a given corpus. More specifically, our algorithm, over a *corpus*, works in two steps: first, according to some temporal clustering distribution *P(TC)*, a temporal clustering *TC* is applied to the event pairs of the *corpus*, and then given that temporal clustering, the *corpus* is generated by using equation (1):

$$P(corpus, TC) = P(TC)P(corpus|TC) \qquad (1)$$

To easily incorporate linguistic constraints defined on event pairs, *corpus* is represented by its event pairs, *EventPairs(corpus)*. Now we can assume event pairs are independent and generated by using the following equation:

$$P(corpus|TC) = \prod_{e_i e_j \in EventPairs(corpus)} P(e_i e_j \mid TC_{ij}) \qquad (2)$$

where e_ie_j are event pairs, and TC_{ij} are the specified temporal relation type of e_ie_j. The marginal probability of *corpus* is computed as follows:

$$P(corpus) = \sum_{\text{All possile temporal clustering } TC} P(TC)P(corpus|TC) \qquad (3)$$

For inducing temporal relations, algorithm runs the EM algorithm on this model. We used a uniform distribution over *P(TC)*.

If we expand the equations, each e_ie_j can be represented by its features, which can potentially be used for determining temporal relation type between events e_i and e_j. Therefore, *P(corpus | TC)* is rewritten using equation (4). Where $e_ie_j^l$ is the value of the l^{th} feature of e_ie_j. These features, which are similar to those mentioned in (Chambers and Jurafsky, 2008), are shown in table 1.

$$\prod_{e_i e_j \in EventPairs(corpus)} P(e_ie_j^{\,1}, e_ie_j^{\,2}, ..., e_ie_j^{\,k} \mid TC_{ij}) \qquad (4)$$

Feature	Description
$Word_1$ & $Word_2$	The text of first and second events
$Lemma_1$ & $Lemma_2$	The lemmatized first and second events heads
$Synset_1$ & $Synset_2$	The WordNet synset for first and second events heads
POS_1 & POS_2	The POS of the first and second events
Event Government $Verb_1$ & $Verb_2$	The verbs that govern the first and second events
Event Government $Verb_1$ & $Verb_2$ POS	The verbs' POS that govern the first and second events
Auxiliary	Any auxiliary adverbs and verbs that modifies the governing verbs
$Class_1$ & $Class_2$	The Class of the first and second events
$Tense_1$ & $Tense_2$	The tense of the first and second events
$Aspect_1$ & $Aspect_2$	The aspect of the first and second events
$Modality_1$ & $Modality_2$	The modality of the first and second events
$Polarity_1$ & $Polarity_2$	The polarity of the first and second events
Tense Match	If two events have the same tense
Aspect Match	If two events have the same aspect
Class Match	If two events have the same class
Tense Pair	Pair of two events' tense
Aspect Pair	Pair of two events' aspect
Class Pair	Pair of two events' class
POS pair	Pair of two events' POS
$Preposition_1$	If first event is in a prepositional phrase or not
$Preposition_2$	If second event is in a prepositional phrase or not
Text order	If the first event occurs first in the document or not
Dominates	If the first event syntactically dominates second event or not
Entity Match	If an entity as an argument is shared between two events

Table 1: The features of events which are used in our algorithm for temporal relation learning

To reduce data sparseness and improve probability estimation, conditional independence assumption is made on these features' value generation. We only assume that *tense* and *aspect* are not independent (i.e., $tense_i$ and $aspect_i$ are dependent), because *tense* and *aspect* define temporal location and event structure, and considering these features together is a powerful source of information in any temporal relation extraction system. By conditional independence assumption, the value of *P(corpus | TC)* can be rewritten as

$$\prod_{e_i e_j \in EventPairs(corpus)} \prod_{All\ features\ l} P\left(e_i e_j^{\ l} \mid TC_{ij}\right) \qquad (5)$$

3.3 The Induction Algorithm

To induce a temporal clustering *TC* on a *corpus*, EM was applied to our proposed model. In the EM algorithm, *corpus* (its event pairs) and temporal clustering *TC* are respectively the observed and unobserved (the hidden) random variables. The EM algorithm includes the following two steps to iteratively estimate the parameters of the model, θ:

E-step: Fix current θ and obtain the conditional temporal clustering likelihoods $P(TC \mid corpus, \theta)$. As a result, for each event pair candidate, a temporal relation type will be selected based on current θ.

Due to inability to consider other relations in pairwise relation learning, some contradictions will be introduced in this step. For example, figure 1 shows an inconsistency in the relations between following events:

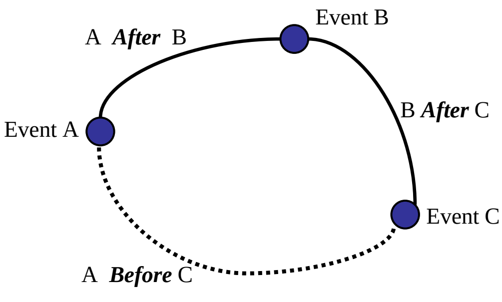

Figure 1: A contradiction in temporal relations between three events A, B, and C.

There are several ways for eliminating such inconsistencies (Mani et al., 2007; Tatu and Srikanth, 2008; Chambers and Jurafsky; 2008). In this paper, we propose a best-first greedy search strategy for temporal reasoning and removing inconsistencies among predicted relations.

First the contradictions in the connected graphs of the text will be discovered with applying a set of rules (e.g., $Before(x, y) \wedge Before(y, z) \rightarrow Before(x, z)$), which are based on Allen's interval algebra (1984). Then the inconsistent relations of each connected graph will be sorted in a list named *SL* based on computed confidence score ($P(TC \mid corpus, \theta)$). In *SL*, the first and the last elements are the most and the least confident relations, respectively.

Now, the algorithm starts from the first relation of *SL*, and pops off this relation and adds it to another list named *FL*. In adding a new relation (r_{new}) to *FL*, the algorithm verifies the consistency between relations of *FL*. If r_{new} is a relation between events e_i and e_j, which introduces an inconsistency into the graph, it will be replaced by the next confident relation between e_i and e_j. These replacements are repeated until *FL* relations will be consistent. When there are no more contradictions in *FL*, algorithm will try to move the next element of *SL* to *FL*. These operations are iterated until there will be no more relations in *SL*. Then the resultant consistent relations in *FL* can be used in the next stages of EM.

M-step: Find θ^{new} that maximizes the equation $\sum_{TC} P(TC \mid corpus, \theta^{old}) \log P(corpus, TC \mid \theta^{new})$ with fixed θ^{old}. In order to predict θ^{new}, different optimization algorithms such as conjugate gradient can be used. However, these methods are slow and costly. In addition, it is difficult to smooth these methods in a desired manner. Therefore, we used smoothed relative frequency estimates.

Now, the EM algorithm can either begin at the E-Step or the M-step, which we start the induction algorithm at the M-step. It is clear that $P(TC \mid corpus, \theta^{old})$ is not available in the first iteration of EM. Instead, an initial distribution over temporal clustering, *P(TC | Corpus)*, can be used. Now, there is an important question: how should we initialize *P(TC | Corpus)*?

Initialization is an important task in EM, because EM only guarantees to find a local maximum of the likelihood. The quality of such a local maximum is highly dependent on the initial start point. We tested three different ways of initialization: first, we used a uniform distribution over all temporal clustering. Second, we used a small part of a labeled corpus for setting *P(TC | Corpus)*. Third, we used some rules for initial estimation of temporal relation types and then used those types for the initial estimation to compute *P(TC | Corpus)*. The detailed accounts of the second and the third methods are discussed in subsection 5.1.

Like many statistical NLP tasks in which smoothing is required to alleviate the problem of data sparseness, smoothing is vital here, too. In particular, in the first few iterations, much more smoothing is required than in later iterations. In our experiments, we used an additive smoothing technique.

4 Corpus Description

In our experiments, we used two standard corpora which had been utilized in evaluation of most previous works: TimeBank (v. 1.2) and Opinion Corpus (Mani et al., 2006). TimeBank includes 183 newswire documents and 64077 words, and Opinion Corpus comprises 73 documents with 38709 words. These two datasets have been annotated based on TimeML (Pustejovsky et al., 2003). There are 14 temporal relations (Event-Event and Event-Time relations) in the TLink class of TimeML. Relation *NONE*, which indicates there is no temporal relation between respected event pairs, must also be considered. For the sake of alleviating the data sparseness problem, we used a converted version of these temporal relations, which contains only four following temporal relations:

BEFORE , AFTER , OVERLAP , NONE

As it was shown in (Bethard et al, 2007a), it is easy to convert 14 TimeML relations into just *BEFORE*, *AFTER*, and *OVERLAP* relations. Here, we merged *BEFORE* and *IBEFORE* relations into only *BEFORE* relations. Similarly *AFTER* and *IAFTER* relations were also merged into *AFTER* relations. All the remaining 10 relation types were collapsed in *OVERLAP* relations.

In our experiments, like several previous works, we merged Opinion and TimeBank to generate a single corpus, which is called OTC. Table 2 shows the converted TLink class distribution over TimeBank and OTC corpora for intra-sentential and general (intra- and inter-sentential) event pairs which are situated in the same document.

Relation Type	**TimeBank Corpus**		**OTC Corpus**	
	Intra-Sentential	**General**	**Intra-Sentential**	**General**
BEFORE	593	706	<u>1944</u>	2369
AFTER	549	692	810	1073
OVERLAP	<u>1225</u>	<u>2083</u>	1623	<u>2792</u>
NONE	11309	353401	16768	543918
Total	**13676**	**356882**	**21145**	**550152**

Table 2: The converted TLink class distribution in TimeBank and OTC for intra-sentential and general event pairs.

5 Evaluation

5.1 Experimental Setup

We applied our algorithm to both TimeBank and OTC corpora, using the five-fold cross validation method. The results were evaluated by measuring accuracy. One important point that we should mention is the parameter initialization of EM.

As it was mentioned in section 3.3, we used three different initializations: first, a uniform distribution over all temporal clustering was used; therefore, all temporal clustering in the first step had equal probability. Second, we used a small part of labeled corpora (10% of each relation type) for setting *P(TC | Corpus)*. Relations were selected randomly. Third, we used some rules for initial estimation of temporal relation types and used this initial estimation for computing *P(TC | Corpus)*. The rules were the combination of *GTag* rules (Mani et al., 2006), *VerbOcean* (Chklovski and Pantel, 2005), and some rules derived from certain signal words (e.g., *"on"*, *"during"*, *"when"*, and *"if"*) of the text.

5.2 Results and Discussions

As it is shown in table 2 (in General columns), *NONE* relations dwarf all other relations. As a result, temporal relation learning, because of heavy bias of learner to *NONE* relations, will be very hard (even useless). Regarding this problem, we set up two different types of experiments:

1) Algorithms were applied only for intra-sentential event pairs, considering all relation types (including *NONE*). The results of these experiments are shown in table 3.

2) The *NONE* relations were removed, and algorithms were applied to both intra-sentential and general (intra- and inter-sentential) event pairs. Table 4 shows the results of experiments without considering *NONE* relations.

One important issue in the results of table 3 is that in our experiments, all four mentioned relation types (*BEFORE*, *AFTER*, *OVERLAP*, and *NONE*) have been considered, but in reporting the results, we have reported the aggregated accuracy of only *BEFORE*, *AFTER*, and *OVERLAP* relations, and excluded the accuracy results of *NONE* relations. That is because by considering *NONE*, one could design a simple system which tags all relations to *NONE*, and would get a very high accuracy. But, in that case the comparison would be inappropriate.

In our evaluations, both table 3 and 4, the baselines have been the majority classes for event pair relations ignoring *NONE* relations of the evaluated corpora (i.e., *BEFORE* and *OVERLAP* relations as it is depicted in table 2). The Mani's method is in fact a supervised method which exclusively uses gold-standard features (Mani et al., 2007). The Chambers' method is similar to Mani's, except that it uses some external resources such as WordNet (Chambers et al., 2007). The Mani and Chambers results are different from (or even lower than) their reported results, because of two differences: first, we considered only three temporal relation types while in their experiments, there were six relation types. Second, the results of table 3 are reported by considering *NONE* relations, but in their original works, there was not any *NONE* relation.

Method Type	TimeBank	OTC Corpus
Baseline	51.75	44.41
Mani	31.77	47.24
Chambers	36.03	48.86
EM_1	23.76 (22.10)	32.48 (32.21)
EM_2	28.65 (26.31)	38.68 (36.45)
EM_3	29.81 (27.13)	39.92 (39.28)

Table 3: The results of proposed method for intra-sentential event pairs on all mentioned relation types including *NONE* relations

Method Type	TimeBank Corpus		OTC Corpus	
	Intra-Sentential	General	Intra-Sentential	General
Baseline	51.75	59.83	44.41	44.79
Mani	54.80	61.55	60.86	60.58
Chambers	62.31	66.79	63.57	62.94
EM_1	41.67 (39.02)	42.09 (40.92)	43.86 (43.75)	42.94 (43.02)
EM_2	46.11 (45.28)	49.54 (48.31)	49.34 (48.35)	50.52 (49.34)
EM_3	48.03 (46.53)	50.88 (47.86)	50.27 (48.23)	49.98 (48.78)

Table 4: The results of different methods for intra-sentential and general event pairs by ignoring *NONE* relations.

EM_1, EM_2, and EM_3 are the results of our proposed method with three different initializations. The initializations of EM_1, EM_2, and EM_3 were random, with little supervision (10%), and by using a number of rules, respectively. For EM_1, one question is how this method can determine the label of different classes. In our experiments, EM_1, depending on the type of experiment, only determines three or four different classes ($Class_1$, $Class_2$, $Class_3$, and/or $Class_4$). To label these unlabeled classes, using annotated data, we assigned the labels in such a way that resulted in maximum similarity between predicted and annotated temporal relation types for each event pair.

In tables 3 and 4, the numbers inside parentheses show the results of our proposed algorithm without applying temporal reasoning.

As it is shown in tables 3, all mentioned methods generally demonstrate a weak performance. That is due to the problem's nature. As distribution of different columns of table 2 shows, the number of *NONE* relations, even in the intra-sentential case, is about 7 to 10 times greater than other relations. Therefore, it is very hard for a learning algorithm to precisely determine the relation types. On the other hand, results of table 4, which ignores *NONE* relations, are satisfactory. Comparing proposed method with the baseline, shows that in the cases that supervised methods can beat the baseline method, our weakly supervised method can also work better than the baseline or close to it.

It should be noted that the Chambers' method, which is the most successful method of tables 3 and 4, is in fact the state of the art supervised method, while our proposed method is, based on the initialization approaches, unsupervised or weakly supervised. Among different settings of the proposed method, EM_3 achieved the best results except for the general case of OTC in table 4, where EM_2 achieved better results.

The results show that EM_1 is not very efficient in either first or second type of experiments. It seems that randomized initialization in this hard problem, may cause some divergence in the probability distribution. On the other hand, both EM_2 and EM_3 showed satisfactory results in these problems. Therefore, initialization is a critical factor in our EM method, and some little source of supervision seems crucial for achieving better results.

Comparison of the results of proposed EM algorithm with and without utilization of temporal reasoning shows that using temporal reasoning can be effective on the accuracy of the algorithm. By using temporal reasoning, some inconsistencies are removed in step E of the algorithm and the predicted relations will be more reliable. Then in step M, the update of parameters will be performed more accurately and thus the accuracy of the algorithm iteratively will increase.

Another important point in the comparison of accuracy results is the existence of *NONE* relations. As it is shown in tables 3 and 4, the accuracies in table 3 is much lower than that of in table 4. These differences are all due to the existence of *NONE* relations, which makes problem hard. Figure 2 demonstrates the effects of *NONE* relations on the accuracy of our proposed algorithm. All the experiments have been performed using OTC. We repeated our experiments for different percentage of *NONE* relations. As it is shown, *NONE* relations have had a great impact on the accuracy of the system.

The larger gap between the accuracy of ignoring and consideration of *NONE* relations on TimeBank (in contrast that of OTC) implies that *NONE* relations would have an even greater impact on the accuracy of the algorithm if applied to TimeBank.

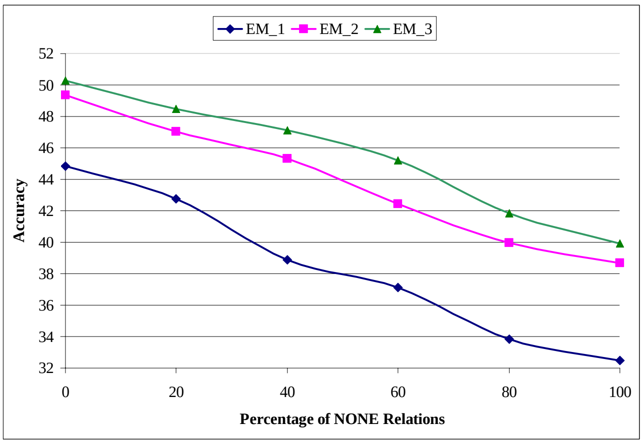

Figure 2: The effect of NONE relations on the accuracy

Figure 2 shows the impact of *NONE* relations on the accuracy (or recall) of the algorithm. Our experiments showed that this impact is even more substantial on the precision of the proposed algorithm. That is because although the algorithm can determine *BEFORE*, *AFTER*, and *OVERLAP* relations with an acceptable rate, but a lot of *NONE* relations will also be recognized. As a result, the precision will substantially decrease. Due to lack of space, we have not reported the precision of the algorithm.

6 Conclusion

In this paper, we have addressed the problem of learning temporal relations between event pairs, which is an interesting topic in natural language processing. Building a suitable corpus is a hard, expensive, and time consuming task. Therefore, we focused on unsupervised and weakly supervised types of learning. We proposed a novel generative model that uses the EM algorithm with some interval algebra reasoning for temporal relation learning. We compared our work with some of successful supervised methods. Our experiments showed that the result of the proposed algorithm, considering its little supervision, is satisfactory.

We think but have not yet verified that using other source of information like narrative information, global relationship between events and times, time expressions, and/or some other useful features of related documents might even further improve the accuracy of the new algorithm.

References

James Allen. 1984. *Towards a General Theory of Action and Time.* Artificial Intelligence, 23, 2, 123-154.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007a. *Finding Temporal Structure in Text: Machine Learning of Syntactic Temporal Relations.* Journal of Semantic Computing, 1, 4.

Steven Bethard, James H. Martin, and Sara Klingenstein. 2007b. *Timelines from Text: Identification of Syntactic Temporal Relations.* Proceeding of ICSC, 11-18.

Steven Bethard and James H. Martin. 2007. *CU-TMP: Temporal Relation Classification Using Syntactic and Semantic Features.* Proceeding of SemEval-2007, 129-132.

Steven Bethard. 2007. *Finding Event, Temporal and Causal Structure in Text: A Machine Learning Approach.* PhD thesis, University of Colorado at Boulder.

Steven Bethard and James H. Martin. 2008. *Learning Semantic Links from a Corpus of Parallel Temporal and Causal Relations*. Proceeding of ACL-2008, 177-180.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. *Classifying Temporal Relations between Events.* Proceeding of ACL-2007, 173-176.

Nathanael Chambers and Dan Jurafsky. 2008. *Jointly Combining Implicit Constraints Improves Temporal Ordering.* Proceeding of EMNLP-2008, 698-706.

Eugene Charniak and Micha Elsner. 2009. *EM Works for Pronoun Anaphora Resolution.* Proceedings of EACL-2009, 148-154.

Colin Cherry and Shane Bergsma. 2005. *An Expectation Maximization Approach to Pronoun Resolution*. Proceeding of CoNLL 2005. 88-95.

Timothy Chklovski and Patrick Pantel. 2005. *Global Path-based Refinement of Noisy Graphs Applied to Verb Semantics.* Proceeding of IJCNLP-05, 792-803.

Zellig Harris. 1968. *Mathematical Structure of Language*. John Wiley Sons, New York, 1968.

Dan Klein. 2005. *The Unsupervised Learning of Natural Language Structure.* Ph.D. Thesis, Department of Computer Science, Stanford University.

Mirella Lapata and Alex Lascarides. 2006. *Learning Sentence-Internal Temporal Relations.* Journal of Artificial Intelligence Research, 27, 85-117.

Dekang Lin and Patrick Pantel. 2001. *Dirt-Discovery of Inference Rules From Text*. Proceeding of ACM SIGKDD-2001, 323-328.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong M. Lee, and James Pustejovsky. 2006. *Machine Learning of Temporal Relations*. Proceedings of ACL-2006, 753-760.

Inderjeet Mani, Ben Wellner, Marc Verhagen, and James Pustejovsky. 2007. *Three Approaches to Learning Tlinks in TimeML. Technical Report CS-07-268*. Brandeis University, Waltham, USA.

Seyed Abolghasem Mirroshandel, Gholamreza Ghassem-Sani, and Mahdy Khayyamian. 2009. *Using Tree Kernels for Classifying Temporal Relations between Events.* Proceedings of PACLIC-2009, 355-364.

Seyed Abolghasem Mirroshandel and Gholamreza Ghassem-Sani. 2010. *Temporal Relations Learning with a Bootstrapped Cross-document Classifier.* Proceedings of ECAI-2010, 829-834.

Vincent Ng, 2008. *Unsupervised Models for Coreference Resolution.* Proceedings of EMNLP-2008, 640-649.

Victor Pekar. 2006. *Acquisition of Verb Entailment from Text.* Proceeding of NAACL/HLT-2006, 49-56.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro and Marcia Lazo. 2003. *The TimeBank Corpus.* Corpus Linguistics, 647-656.

Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. *Scaling Web-based Acquisition of Entailment Relations.* Proceeding of EMNLP-2004, 41-48.

Marta Tatu and Munirathnam Srikanth. 2008. *Experiments with Reasoning for Temporal Relations between Events.* Proceeding of Coling-2008, 857-864.

Kentaro Torisawa. 2006. *Acquiring Inference Rules with Temporal Constraints by Using Japanese Coordinated Sentences and Noun-Verb Co-Occurrences.* Proceedings of NAACL-2006, 57-64.

Puscasu G. Wvali. 2007. *Temporal Relation Identification by Syntactico-Semantic Analysis.* Proceeding of SemEval-2007, 484–487.