Toward an Automated Time-Event Anchoring System

Chong Min Lee and Graham Katz
Department of Linguistics, Georgetown University
37th and O Streets, Washington, DC 20036, USA
cml54@georgetown.edu, egk7@georgetown.edu

Abstract

Temporal information processing is an essential step to get deeper understanind of text. This research focuses on getting a reference information of a sentence by anchoring an event to a explicit time expression in a sentence. Even though previous research considered an anchoring as one step process, we adopt two step processes such as link detection and link classification. And, we try to find optimal feature set with hand coded data first, then we compare it with feature set extracted from NLP tools. Through our experiments, we show that a bottle neck exists in link classification, and syntactic information is crucial to the link detection task.

1 Introduction

Even though language processing techniques have produced useful methods for natural language understanding, the need for deeper understanding of text is increasing. When a system understands temporal ordering of events in a text, it can answer questions on which event occurs at a certain time or before an event. Moreover, the notion of time and unfolding of events is central to all narratives (Mani & Pustejovsky 2004). When the temporal knowledge is represented in a summarization system, it can generate better summarization of a narrative.

The temporal interpretation of a text has long been an important area in linguistics research (Bennett and Partee, 1972; Kamp and Reyle, 1993). In recent years, with the advent of the TimeML markup language (Pustejovsky et al., 2003) and the creation of the TimeBank resource (Pustejovsky et al., 2003), the interest has focussed on the application of machine learning techniques to this task (Mani et al., 2006; Bramsen et al., 2006; Chambers et al., 2007), with a recent competition pitting groups against one another (Verhagen et al., 2007).

In this paper we present a set of experiments in which we investigate aspects of a machine learning solution to the problem of determining the temporal relations that hold among the events and times referred to in a sentence (the *event anchoring* problem of TempEval Task A). The paper has two goals: first to determine which of the large range of potentially relevant linguistic features are useful to the anchoring task, and secondarily to determine the degree to which NLP techniques can replace manual annotation in this task.

(1) In the last twenty four hours, the value of the Indonesian stock market has fallen by twelve percent.

When we are asked about when the Indonesian stock market value was fallen, we need to figure out a time expression that stock-market-falling event can be anchored to. The anchoring process can be assumed as consisting of two subtasks. The first subtask is to determine whether an event expression can be anchored to a time expression (link detection), and the second one is to identify an appropriate relation between them (link classification). Previous research didn't make the distinction and tried to find the best feature set

for the anchoring process. This research will show what is the best feature set for each subtask.

As mentioned in (Boguraev and Ando, 2005), "TimeML can be used as the first pass in the syntax-semantics interface of a temporal resolustion framework." So, the goal of this research is to construct an analysis compatible with TimeML specification. To get the goal, we use TimeBank 1.2 that is built in TimeML 1.2.

In section 2, we will summarize previous works. And, the design of our experiments will be discussed in Section 3. Section 4 will give the results of the experiments. Then, we will make a discussion on the results in Section 5.

2 Previous Work

(Boguraev and Ando, 2005) used a machine learning tool to recognize events and detect links between a event and a time expression in a sentence. In the research, the following was used as features: POS and tokens in five partitions, and syntactic relations such that they appear in a clause or not. Then, the performance was evaluated based on the distances between an event and a time.

(Mani et al., 2006) compared the performance between a hand-coded rule approach and a machine learning approach. In the research, they used attribute values available in TimeBank as a feature set. The research identified every possible pair of events and times. So, it made links between an event and a time even though they do not appear in a sentence.

Six teams competed In TempEval-2007 ((Verhagen et al., 2007)) for three tasks: (1) the identification of the temporal relations "holding between time and event expressions that occur within the same sentence", (2) the identification of the temporal relations "holding between the Document Creation Time and event expressions", and (3) the identification of the temporal relations "between the main events of adjacent sentences." In the first task competition, the number of relations to be identified was reduced into 6 relations such as BEFORE, OVERLAP, AFTER, BEFORE-OR-OVERLAP, OVERLAP-OR-AFTER, and VAGUE. And, event terms were restricted into the terms that occurrs over twenty times in TimeBank.

The previous research did not try to find most influential feature set with hand-coded corpus in figuring out temporal relations between an event and a time in a sentence. (Mani et al., 2006) adopted every attribute value that can be extracted from TimeBank tags. The other research extracted features from NLP tools, even though the performance could be influenced by the performance of the tools.

And, the research ((Mani et al., 2006; Verhagen et al., 2007)) ignored the distinction between identifying the possibility in linking an event to a time and identifying the relation of the link whenever the link is possible. The feature set that showed the best performance is not guranteed as the best one when the performance of NLP tools is improved.

3 Overall Experiment Description

We assume that the task of intra-sentential temporal interpretation is composed of two subtasks: *link detection* is the task of determining whether or not a given event-expression is to be related to a given temporal expression; *link classification* is the task of determining for those event-time pairs that are to be related which temporal relation holds between them (*before, begins, ends, overlaps, includes*, or *simultaneity*).

We experiment two things. First, we try to find the best feature sets for the subtasks with the information from gold standard corpus. Then, we compare the performance with the performance when the information is from NLP tools.

3.1 Features

There are a number of features that, intuitively speaking, might contribute to getting the right temporal interpretation: lexical features, such as the presence in the sentence of the preposition *in*, or the tense and aspect of the verb; syntactic features, such as the syntactic path from the main verb to the temporal expression (PP>S>VP>V), and semantic features, such as meaning of the time expression and the type of verb involved. In (2) we list the factors that we chose to investigate.

(2) Feature sets

- Lexical features: modal, tense, aspect, lemma of event word, words of time expression, normalized time expression, signals (event and

time expression), part-of-speech

- Relational features: syntactic path, distance between two elements, number of events or time expressions between two elements

- Semantic features: event class, WordNet meaning of event, class of time expression, value of time expression

We use TimeBank 1.2 (Pustejovsky et al., 2003) as training and test data. And it is assumed that the temporal expressions and event-denoting expressions of interest were marked in the input. Naive Baysian classifier in WEKA (Witten and Frank, 2005) is used in the experiments.

3.2 Data

In this research, I used TimeBank version 1.2 which consists of 186 documents and 64,077 words. Events and temporal expressions of news articles are marked in TimeBank, based on TimeML specification (Pustejovsky et al., 2003). EVENT, TIMEX3, s (sentence boundary), and TLINKtags in TimeBank are used in the research. Sentence boundaries of a manually tagged corpus are employed in the first experiment. In the second experiment, sentence boundaries are automatically computed with Alembic sentence splitter.

EVENTs are defined as a cover term for situations that happen or occur. Events can be expressed by verbs, nouns, adjectives, predicative clauses, or prepositional phrases (Pustejovsky et al., 2003). Whenever complex expressions are expressed as events, TimeML usually marks head words of the expressions. We use the event attributes such as event class, tense, aspect, parts-of-speech, polarity, and modality from TimeBank corpus in the experiment 1. Then, the similar information with the attributes is extracted from several NLP methods in the experiment 2.

Temporal expressions are marked with TIMEX3 tags. Information on temporal expressions is specified as attributes such as value, type, beginPoint, endpoint, quant, freq, functionInDocument, mod, and temporalFunction. Each document has a special TIMEX3 at its head. The TIMEX3 represents speech time of the document. This research extracts some of the attributes that can be achieved through NLP tools such as TimexTag (Ahn et al., 2007).

relType	WSJ part	Whole TimeBank
BEFORE	41	61
BEGINS	44	48
ENDS	134	147
IBEFORE	3	3
INCLUDES	652 (70.2%)	822 (71.6%)
SIMULTANEOUS	55	67
Total	929	1148

Table 1: Counts of each relType when linked EVENT and TIMEX3 are in the same sentence

TLINK (Temporal Link) represents the temporal relationship holding between events, times, or between an event and a time (Pustejovsky et al. 2005). Possible types of this relationship are SIMULTANEOUS, BEFORE, AFTER, INCLUDES, IS_INCLUDED, DURING, IMMEDIATELY AFTER (IAFTER), IMMEDIATELY BEFORE (IBEFORE), IDENTITY, BEGINS, ENDS, BEGUN_BY, ENDED_BY, and DURING_INV. The size of TimeBank is too small to classify the types by means of machine-learning algorithms. Also, some of the relations such as INCLUDES and IS_INCLUDED are redundant. Therefore, I merge 14 relations into 6 relations: SIMULTANEOUS, BEFORE, IBEFORE, BEGINS, ENDS, and INCLUDES. The distribution of the relations is given in Table 1.

3.3 Experiment 1

In the first experiment with gold standard information, the Wall Street Journal portion of the TimeBank 1.2 is used. The restriction is due to the availabilty of syntactic parsed trees from Penn TreeBank. An obstacle in the experiment is the mismatches in sentence boundaries between TimeBank 1.2 and Penn TreeBank. The mismatches are resolved based on the boundaries of PTB because we need to extract syntactic path information.

For link detection, targets are marked as 1 for the existence of a link and 0 for no link. Each relation name is used with the relation identification task as targets. We start from the whole lexical features. Then, each lexical feature is removed to evaluate its influence on the subtasks.

After the evaluation, the performance of relational features and the combination of lexical and relational features is tested. First, the performance of relational feature set is evaluated. Secondly, each relational feature is removed from the relational set. After checking the performance of relational features, the combination of relational features with lexical features are tested. Each relational feature is added into the lexical feature set. Then, we evaluate the combination of relational feature set with lexical feature set. As the final step, the influence of semantic information and the combinations of lexical, relational, and semantic information is evaluated. 10 folds-cross validation is used as the evaluation measure.

3.4 Experiment 2

The aim of the experiment is to estimate the potential of an automated system in each subtask. Sentence boundaries are tagged with the Alembic sentence splitter. The generated sentence boundaries do not match those in the corpus which makes the number of instances different from each other. Therefore, results of the experiment cannot be directly compared with the first experiment, but, they can demonstrate the relative ability of the NLP system.

- Off-the-shelf NLP tools: Charniak parser, TimeTag, Alembic sentence splitter, morpha
- Purpose-built software: tense, aspect, and modal extractors.

NLP tools that are used in the experiment can be categorized into off-the-shelf tools and purpose-built software. There are no available tools for the extraction of sentence tense, sentence aspect, and modal auxiliary verb. Therefore, three finite state machines for the information are implemented by ourselves. After the extraction of features, we follow the steps in the experiment 1.

4 Evaluations and Results

Two different measures, accuracy and F-measure, are used in the experiments because the targets of link detection and link classification show different characteristics. In link detection, system performance is evaluated with F-measure while accuracy is used for temporal relation classification. Link detection system should show good performance in detecting links between events and temporal expressions instead of showing good performance for non-link detection. When we use accuracy measure in the detection process, it shows the system performance on how good the system predicts links and no links. But, no links take big portion of the data. Therefore, accuracy measure shows higher values than the actual performance. In a relation detection system, however, a non-relation target is not part of target relations. The system performance of every target category should be evaluated which means that F-measure is not appropriate measure to evaluate the system because F-measure is the evaluation on a target form.

Feature Set	Link Detection		Relation Identification	
	Gold	NLP	Gold	NLP
Lex	29.4%	28.2%	66.8%	67.2%
Rel	72.0%	72.5%	69.9%	69.2%
Sem	17.1%	14.2%	68.7%	69.8%
Lex+Rel	71.9%	72.6%	65.6%	66.1%
Lex+Sem	34.2%	32.2%	65.0%	64.4%
Rel+Sem	73.2%	73.8%	67.2%	69.0%
Lex+Rel+Sem	71.3%	69.9%	64.8%	63.8%

Table 2: Results of Category Combinations

In table 2, we can see the rise of the performance measure whenever relational feature set is added to the combinations in link detection task. Moreover, when relational feature set is only used, 72% F-measuere is observed in link detection task. But, we can find little variation of accuracy measure in relation identification. Moreover, 'includes' takes 70% portion of TLINKs in a sentence. When we consider it as a baseline. the performances in the link classification task are lower than the baseline.

The best performance feature set in link detection with gold standard information is the combination of lexical feature set, relational feature set, and EVENT and TIMEX3 class information except EVENT and TIMEX3 words. When we use lexical feature set without EVENT word as feature vector for training, link detection system shows the worst performance, 10.4% F-measure. The best feature set in the link detection system that is trained with features from NLP tools is different from gold standard one. Rela-

	Link Detection		Link Classification	
	Featureset	F-measure	Featureset	Accuracy
Best Gold	Lex+Rel w classes - words	73.8%	Lex wo event word	70.6%
Best NLP	Rel w classes	74.2%	Lex wo event word	70.2%
Worst Gold	Lex wo event word	10.4%	Lex+Rel w Timex value	64.2%
Worst NLP	Lex wo event word	10.8%	Lex w Event class	63.7%

Table 3: Best and Worst Performance

tional features with EVENT and TIMEX3 class information is the best set. In link classification task, lexical feature set without EVENT words is the best feature set. Moreover, the feature set is only over the baseline measure.

5 Discussion

These experiments show the importance of relational features such as distance, parse tree path and number of events in the link attachment decision. When only relational features are evaluated, the link detection system shows 72% F-measure. This value is a relatively high score when it is compared with the 29% of lexical level features and 17% of semantic features. This can be support for the influence of relational information over the linking attachment decision when an event and a temporal expression are in the same sentence.

The performance of link classification system was not successful in the experiments. The skewed distribution of relation types can cause poor performance. The distribution of every other relation type except INCLUDES and ENDS is less than 6%. There are two reasons for the low performance. First, skewed data generates skewed model. The link classification system that is trained with the best feature set has 83.8% F-measure (74.8% precision and 95.4% recall) with the WSJ. Second, the training data size around 800 is too small to yield good performance with 6 types that are not distributed evenly. One solution for the low performance is the construction of a rule-based system for the types that have small amount of instances.

6 Conclusions

Our experiments show that syntactic information is crucial to the link detection task. The link classification task was clearly more difficult, and none of the features we made use of were resulted in particularly good system performance. This clearly indicates that for this task, other sources of information are required. We were encouraged, however, by the fact that using automatic processing such as parsing and event-type tagging, to deliver training data, yielded results comparable to gold-standard training. Finally, we are currently working to extend this evaluation to intersentential links and to the problem of integrating link coherence.

References

Michael Bennett and Barbara Partee. 1972. Toward the logic of tense and aspect in English. *Technical report, System Development Corporation*. Santa Monica, CA

Branimir Boguraev and Rie Kubota Ando. 2005. TimeML-Compliant Text Analysis for Temporal Reasoning. *Preceedings of IJCAI-05*, 997–1003.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing Temporal Graphs *Proceedings of EMNLP 2006*, 189–198.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *IWCS-5, Fifth International Workshop on Computational Semantics*.

Nathanael Chambers, Shan Wang, and Dan Jurafsky. 2007. Classifying Temporal Relations Between Events *Proceedings of ACL 2007*, 173–176.

David Ahn and Joris van Rantwijk and Maarten de Rijke. 2007. A Cascaded Machine Learning Approach to Interpreting Temporal Expressions *Proceedings of NAACL-HLT 2007*.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to modeltheoretic semantics of natural language*. Kluwer Academic, Boston.

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

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, David Day, Lisa Ferro, Robert Gaizauska, Marcia Lazo, Andrea Setzer, Beth Sundheim. 2003. The

TIMEBANK corpus. *Proceedings of Corpus Linguistics 2003*, 647–656.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. *Proceedings of SemEval-2007*, 75–80.

Ian D. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA.