NavyTime: Event and Time Ordering from Raw Text

Nathanael Chambers
United States Naval Academy
Annapolis, MD 21401, USA
nchamber@usna.edu

Abstract

This paper describes a complete event/time ordering system that annotates raw text with events, times, and the ordering relations between them at the SemEval-2013 Task 1. Task 1 is a unique challenge because it starts from raw text, rather than pre-annotated text with known events and times. A working system first identifies events and times, then identifies which events and times should be ordered, and finally labels the ordering relation between them. We present a split classifier approach that breaks the ordering tasks into smaller decision points. Experiments show that more specialized classifiers perform better than few joint classifiers. The NavyTime system ranked second both overall and in most subtasks like event extraction and relation labeling.

1 Introduction

The SemEval-2013 Task 1 (TempEval-3) contest is the third instantiation of an event ordering challenge. However, it is the first to start from raw text with the challenge to create an end-to-end algorithm for event ordering. Previous challenges included the individual aspects of such a system, including event extraction, timex extraction, and event/time ordering (Verhagen et al., 2007; Verhagen et al., 2010). However, neither task was dependent on the other. This paper presents NavyTime, a system inspired partly by this previous breakup of the tasks. We focus on breaking up the event/time ordering task further, and show that 5 classifiers yield better performance than the traditional 3 (or even 1).

The first required steps to annotate a document are to extract its events and time expressions. This paper describes a new event extractor with a rich set of contextual features that is a top performer for event attributes at Tempeval-3. We then explore additions to SUTime, a top rule-based extractor for time expressions (Chang and Manning, 2012). However, the core challenge is to link these extracted events and times together. We describe new models for these difficult tasks: (1) identifying ordered pairs, and (2) labeling the ordering relations.

Relation identification is rarely addressed in the literature. Given a set of events, which pairs of events are temporally related? Almost all previous work assumes we are given the pairs, and the task is to label the relation (before, after, etc.). Raw text presents a new challenge: extract the relevant pairs before labeling them. We present some of the first results that compare rule-based approaches to trained probabilistic classifiers. These are the first such comparisons to our knowledge.

Finally, after relation identification, we label relations between the pairs. This is the traditional event ordering task, although we now start from noisy pairs. Our main contribution is to build independent classifiers for intra-sentence event/time pairs. We show improved performance when training these split classifiers. NavyTime’s approach is highly competitive, achieving 2nd place in relation labeling (and overall).

2 Dataset

All models are developed on the TimeBank (Pustejovsky et al., 2003) and AQUAINT corpora (Mani
et al., 2007). These labeled newspaper articles have fueled many years of event ordering research. TimeBank includes 183 documents and AQUAINT includes 73. The annotators of each were given different guidance, so they provide unique distributions of relations. Development of the algorithms in this paper were solely on 10-fold cross validation on the union of the two corpora.

The SemEval-2013 Task 1 (TempEval-3) provides unseen raw text to then evaluate the final systems. Final results are from this set of unseen newspaper articles. They were annotated by a different set of people who annotated TimeBank and AQUAINT.

3 Event Extraction

The first stage to processing raw text is to extract the event mentions. We treat this as a binary classification task, classifying each token as either event or not-event. Events are always single tokens in the TimeBank/AQUAINT corpora, so a document with $n$ tokens requires $n$ classifications. Further, each event is marked up with its tense, aspect, and class.

We used a maximum entropy classification framework based on the lexical and syntactic context of the target word. The same features are used to first identify events (binary decision), and then three classifiers are trained for the tense, aspect, and class. The following features were used:

- **Token N-grams**: Standard n-gram context that includes the target token (1,2,3grams), as well as the unigrams and bigrams that occur directly before and after the target token.
- **Part of Speech n-grams**: The POS tag of the target, and the bigram and trigram ending with the target.
- **Lemma**: The lemmatized token in WordNet.
- **WordNet-Event**: A binary feature, true if the token is a descendent of the Event synset in WordNet.
- **Parse Path**: The tree path from the token’s leaf node to the root of the syntactic parse tree.
- **Typed Dependencies**: The typed dependency triple of any edge that begins or ends with the target.

We used 10-fold cross validation on the combined corpora of TimeBank and AQUAINT to develop the above features, and then trained one classifier on the entire dataset. Our approach was the 2nd best event extraction system out of 8 submission sites on the unseen test set from TempEval-3. Detailed results are given in Figure 1.

Results on event attribute extraction were also good (Figure 1). We again ranked 2nd best in both Tense and Aspect. Only with the Class attribute did we fare worse (4th of 8). We look forward to comparing approaches to see why this particular attribute was not as successful.

4 Temporal Expression Extraction

As with event extraction, time expressions need to be identified from the raw text. Recent work on time extraction has suggested that rule-based approaches outperform others (Chang and Manning, 2012), so we adopted the proven SUTime system for this task. SUTime is a rule-based system that extracts phrases and normalizes them to a TimeML time. However, we improved it with some TimeBank specific rules.

We observed that the phrases ‘a year ago’ and ‘the latest quarter’ were often inconsistent with standard TimeBank annotations. These tend to involve fiscal quarters, largely due to TimeBank’s heavy weight on the financial genre. For these phrases, we first determine the current fiscal quarter, and adjust the normalized time to include the quarter, not just the year (e.g., 2nd quarter of 2012, rather than just 2012). Further, the generic phrase ‘last year’ should normalize to just a year, and not include a more specific month or quarter. We added rules to strip off months.

SUTime was the best system for time extraction, and our usage matched its performance as one would hope. Full credit goes to SUTime, and its extraction is not a contribution of this paper. However, NavyTime outperformed SUTime by over 3.5 F1 points on time normalization. Our additional rulebank appears to have helped significantly, allowing NavyTime to be the 2nd best in this category behind HeidelTime. We recommend users to use either HeidelTime or SUTime with the NavyTime rulebank.

5 Temporal Relation Extraction

After events and time expressions are identified, it remains to create temporal links between them. A temporal link is an ordering relation that occurs in four possible entity pairings: event-event, event-time, time-time, and event-DCT (DCT is the document creation time).
It is unrealistic to label all possible pairs in a document. Many event/time pairs have ambiguous orderings, and others are simply not labeled by the annotators. We propose a two-stage approach where we first identify likely pairs (relation identification), and then independently decide what specific ordering relation holds between them (relation labeling).

5.1 Relation Identification

TempEval-3 defined the set of possible relations to exist in particular configurations: (1) any pairs in the same sentence, (2) event-event pairs of main events in adjacent sentences, and (3) event-DCT pairs. However, the training and test corpora do not follow these rules. Many pairs are skipped to save human effort. This task is thus a difficult balance between labeling all true relations, but also matching the human annotators. We tried two approaches to identifying pairs: rule-based, and data-driven learning.

Rule-Based: We extract all event-event and event-time pairs in the same sentence if they are adjacent to each other (no intervening events or times). We also extract main event pairs of adjacent sentences. We identify main events by finding the highest VP in the parse tree.

Data-Driven: This approach treats it as a binary classification task. Given a pair of entities, determine if they are ordered or not-ordered. We condense the training corpora’s TLINK relations into ordered, and label all non-labeled pairs as not-ordered. We tried a variety of classifiers for each event/time pair type: (1) intra-sentence event-event, (2) intra-sentence event-time, (3) inter-sentence event-event, and (4) event-DCT.

The data-driven features are shown in Figure 2. After labeling pairs of entities, the ordered pairs are then labeled with specific relations, described next.

5.2 Relation Labeling

This is the traditional ordering task. Given a set of entity pairs, label each with a temporal relation. TempEval-3 uses the full set of 12 relations.

Traditionally, ordering research trains a single classifier for all event-event links, and a second for all event-time links. We experimented with more...
specific classifiers, observing that two events in the
same sentence share a syntactic context that does not
exist between two events in different sentences. We
must instead rely on discourse cues and word seman-
tics for the latter. We thus propose using different
classifiers to learn better feature weights for these
unique contexts. Splitting into separate classifiers is
largely unexplored on TimeBank, and just recently
applied to a medical domain (Xu et al., 2013).

We train two MaxEnt classifiers for event-event
links (inter and intra-sentence), and two for event-
time links. The event-DCT links also have their own
classifier for a total of 5 classifiers. We use the same
features (Figure 2) as in relation identification.

5.3 Experiments and Results
All models were created by using 10-fold cross val-
ification on TimeBank+AQUAINT. The best model
was then trained on the entire set. Features seen
only once were trimmed from training. The relation
labeling confidence threshold was set to 0.3. Final
results are reported on the held out test set provided
by SemEval-2013 Task 1 (TempEval-3).

Our first experiments focus on relation labeling.
This is a simpler task than identification in that we
start with known pairs of entities, and the task is to
assign a label to them (Task C-relation at SemEval-
2013 Task 1). Table 1 gives the results. Our system
initially ranked second with 46.83.

The next task is both relation identification and
relation labeling combined (Task C). This is unfor-
nunately a task that is difficult to define. Without a
completely labeled graph of events and times, it is
not about true extraction, but matching human la-
beling decisions that were constrained by time and
effort. We experimented with rule-based vs data-
driven extractors. We held our relation labeling
model constant, and swapped different identification
models in and out. Our best configuration was eval-
uated on test. Results are shown in Table 2. Navy-
Time is the third best performer.

Finally, the full task from raw text requires all
stages of this paper, starting from event and tem-
poral extraction, then applying relation ID and la-
beling. Results are shown in Table 3. Our system
ranked 2nd of 4 systems.

Our best performing setup uses trained classi-
fiers for relation identification of event-event and
event-DCT links, but deterministic rules for event-
time links (Sec 5.1). It then uses trained classi-
fiers for relation labeling of all pair types. Training
with TimeBank+AQUAINT outperformed just
TimeBank. The split classifier approach for intra
and inter-sentence event-event relations also outper-
formed a single event-event classifier. We cannot
give more specific results due to space constraints.

6 Discussion
Our system was 2nd in most of the subtasks and
overall (Task ABC). Split-classifiers for inter and
intra-sentence pairs are beneficial. Syntactic fea-
tures help event extraction. Compared to cleartk,
NavyTime was better in event and time extraction
individually, but worse overall. Our approach to re-
lation identification is likely the culprit.

We urge future work to focus on relation identifi-
cation. Event and time performance is high, and re-
lation labeling is covered in the literature. For iden-
tification, it is not clear that TimeBank-style corpora
are appropriate for evaluation. Human annotators do
not create connected graphs. How can we evaluate
systems that do? Do we want systems that mimic
imperfect, but testable human effort? Accurate eval-
uation on raw text requires fully labeled test sets.

| UTTime Best | 56.45 |
| NavyTime (TimeBank+AQUAINT) | 46.83 |
| NavyTime (TimeBank) | 43.92 |
| JU-CSE Best | 34.77 |

Table 1: Task Crel, F1 scores of relation labeling.

cleartk Best | 36.26 |
| UTTime-5 | 34.90 |
| NavyTime (TimeBank+AQUAINT) | 31.06 |
| JU-CSE Best | 26.41 |
| NavyTime (TimeBank) | 25.84 |
| KUL | 24.83 |

Table 2: Task C, F1 scores of relation ID and labeling.

cleartk Best | 30.98 |
| NavyTime (TimeBank+AQUAINT) | 27.28 |
| JU-CSE | 24.61 |
| NavyTime (TimeBank) | 21.99 |
| KUL | 19.01 |

Table 3: Task ABC, Extraction and labeling raw text.
References

Angel Chang and Christopher D. Manning. 2012. Su- time: a library for recognizing and normalizing time expressions. In Proceedings of the Language Re- sources and Evaluation Conference.

Inderjeet Mani, Ben Wellner, Marc Verhagen, and James Pustejovsky. 2007. Three approaches to learning tlinks in timeml. Technical Report CS-07-268, Brandeis University.

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.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In Proceedings of the 4th Interna- tional Workshop on Semantic Evaluations, pages 75– 80. Association for Computational Linguistics.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 57–62. As- sociation for Computational Linguistics.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. Journal of the American Medi- cal Informatics Association.