ClearTK-TimeML: A minimalist approach to TempEval 2013

Steven Bethard
Center for Computational Language and Education Research
University of Colorado Boulder
Boulder, Colorado 80309-0594, USA
steven.bethard@colorado.edu

Abstract

The ClearTK-TimeML submission to TempEval 2013 competed in all English tasks: identifying events, identifying times, and identifying temporal relations. The system is a pipeline of machine-learning models, each with a small set of features from a simple morpho-syntactic annotation pipeline, and where temporal relations are only predicted for a small set of syntactic constructions and relation types. ClearTK-TimeML ranked 1st for temporal relation F1, time extent strict F1 and event tense accuracy.

1 Introduction

The TempEval shared tasks (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013) have been one of the key venues for researchers to compare methods for temporal information extraction. In TempEval 2013, systems are asked to identify events, times and temporal relations in unstructured text.

This paper describes the ClearTK-TimeML system submitted to TempEval 2013. This system is based off of the ClearTK framework for machine learning (Ogren et al., 2008)[1], and decomposes TempEval 2013 into a series of sub-tasks, each of which is formulated as a machine-learning classification problem. The goals of the ClearTK-TimeML approach were:

- To use a small set of simple features that can be derived from either tokens, part-of-speech tags or syntactic constituency parses.
- To restrict temporal relation classification to a subset of constructions and relation types for which the models are most confident.

Thus, each classifier in the ClearTK-TimeML pipeline uses only the features shared by successful models in previous work (Bethard and Martin, 2006; Bethard and Martin, 2007; Llorens et al., 2010; UzZaman and Allen, 2010) that can be derived from a simple morpho-syntactic annotation pipeline[2]. And each of the temporal relation classifiers is restricted to a particular syntactic construction and to a particular set of temporal relation labels. The following sections describe the models, classifiers and datasets behind the ClearTK-TimeML approach.

2 Time models

Time extent identification was modeled as a BIO token-chunking task, where each token in the text is classified as being at the B(eginning) of, I(nside) of, or O(utside) of a time expression. The following features were used to characterize tokens:

- The token's text
- The token's stem
- The token's part-of-speech
- The unicode character categories for each character of the token, with repeats merged (e.g. *Dec28* would be 'LuLlNd')
- The temporal type of each alphanumeric sub-token, derived from a 58-word gazetteer of time words
- All of the above features for the preceding 3 and following 3 tokens

Time type identification was modeled as a multi-class classification task, where each time is classified

[1] http://cleartk.googlecode.com/

[2] OpenNLP sentence segmenter, ClearTK PennTreebankTokenizer, Apache Lucene Snowball stemmer, OpenNLP part-of-speech tagger, and OpenNLP constituency parser

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

as DATE, TIME, DURATION or SET. The following features were used to characterize times:

- The text of all tokens in the time expression
- The text of the last token in the time expression
- The unicode character categories for each character of the token, with repeats merged
- The temporal type of each alphanumeric sub-token, derived from a 58-word gazetteer of time words

Time value identification was not modeled by the system. Instead, the TimeN time normalization system (Llorens et al., 2012) was used.

3 Event models

Event extent identification, like time extent identification, was modeled as BIO token chunking. The following features were used to characterize tokens:

- The token's text
- The token's stem
- The token's part-of-speech
- The syntactic category of the token's parent in the constituency tree
- The text of the first sibling of the token in the constituency tree
- The text of the preceding 3 and following 3 tokens

Event aspect identification was modeled as a multi-class classification task, where each event is classified as PROGRESSIVE, PERFECTIVE, PERFECTIVE-PROGRESSIVE or NONE. The following features were used to characterize events:

- The part-of-speech tags of all tokens in the event
- The text of any verbs in the preceding 3 tokens

Event class identification was modeled as a multi-class classification task, where each event is classified as OCCURRENCE, PERCEPTION, REPORTING, ASPECTUAL, STATE, I-STATE or I-ACTION. The following features were used to characterize events:

- The stems of all tokens in the event
- The part-of-speech tags of all tokens in the event

Event modality identification was modeled as a multi-class classification task, where each event is classified as one of WOULD, COULD, CAN, etc. The following features were used to characterize events:

- The text of any prepositions, adverbs or modal verbs in the preceding 3 tokens

Event polarity identification was modeled as a binary classification task, where each event is classified as POS or NEG. The following features were used to characterize events:

- The text of any adverbs in the preceding 3 tokens

Event tense identification was modeled as a multi-class classification task, where each event is classified as FUTURE, INFINITIVE, PAST, PASTPART, PRESENT, PRESPART or NONE. The following features were used to characterize events:

- The last two characters of the event
- The part-of-speech tags of all tokens in the event
- The text of any prepositions, verbs or modal verbs in the preceding 3 tokens

4 Temporal relation models

Three different models, described below, were trained for temporal relation identification. All models followed a multi-class classification approach, pairing an event and a time or an event and an event, and trying to predict a temporal relation type (BEFORE, AFTER, INCLUDES, etc.) or NORELATION if there was no temporal relation between the pair.

While the training and evaluation data allowed for 14 possible relation types, each of the temporal relation models was restricted to a subset of relations, with all other relations mapped to the NORELATION type. The subset of relations for each model was selected by inspecting the confusion matrix of the model's errors on the training data, and removing relations that were frequently confused and whose removal improved performance on the training data.

Event to document creation time relations were classified by considering (event, time) pairs where each event in the text was paired with the document creation time. The classifier was restricted to the relations BEFORE, AFTER and INCLUDES. The following features were used to characterize such relations:

- The event's aspect (as classified above)
- The event's class (as classified above)
- The event's modality (as classified above)
- The event's polarity (as classified above)
- The event's tense (as classified above)
- The text of the event, only if the event was identified as having class ASPECTUAL

Event to same sentence time relations were classified by considering (event, time) pairs where the syntactic path from event to time matched a regular expression of syntactic categories and up/down movements through the tree: ˆ((NP|PP|ADVP)↑)* ((VP|SBAR|S)↑)* (S|SBAR|VP|NP) (↓(VP|SBAR|S))* (↓(NP|PP|ADVP))*$. The classifier relations were restricted to INCLUDES and IS-INCLUDED. The following features were used to characterize such relations:

- The event's class (as classified above)
- The event's tense (as classified above)
- The text of any prepositions or verbs in the 5 tokens following the event
- The time's type (as classified above)
- The text of all tokens in the time expression
- The text of any prepositions or verbs in the 5 tokens preceding the time expression

Event to same sentence event relations were classified by considering (event, event) pairs where the syntactic path from one event to the other matched ˆ((VP↑|ADJP↑|NP↑)? (VP|ADJP|S|SBAR) (↓(S|SBAR|PP))* ((↓VP|↓ADJP)*|(↓NP)*)$. The classifier relations were restricted to BEFORE and AFTER. The following features were used to characterize such relations:

- The aspect (as classified above) for each event
- The class (as classified above) for each event
- The tense (as classified above) for each event
- The text of the first child of the grandparent of the event in the constituency tree, for each event
- The path through the syntactic constituency tree from one event to the other
- The tokens appearing between the two events

5 Classifiers

The above models described the translation from TempEval tasks to classification problems and classifier features. For BIO token-chunking problems, Mallet[3] conditional random fields and LIBLINEAR[4] support vector machines and logistic regression were applied. For the other problems, LIBLINEAR, Mallet MaxEnt and OpenNLP MaxEnt[5] were applied. All classifiers have hyper-parameters that must be tuned during training – LIBLINEAR has the classifier type and the cost parameter, Mallet CRF has the iteration count and the Gaussian prior variance, etc.[6]

The best classifier for each training data set was selected via a grid search over classifiers and parameter settings. The grid of parameters was manually selected to provide several reasonable values for each classifier parameter. Each (classifier, parameters) point on the grid was evaluated with a 2-fold cross validation on the training data, and the best performing (classifier, parameters) was selected as the final model to run on the TempEval 2013 test set.

6 Data sets

The classifiers were trained using the following sources of training data:

TB The TimeBank event, time and relation annotations, as provided by the TempEval organizers.

AQ The AQUAINT event, time and relation annotations, as provided by the TempEval organizers.

SLV The "Silver" event, time and relation annotations, from the TempEval organizers' system.

BMK The verb-clause temporal relation annotations of (Bethard et al., 2007). These relations are added on top of the original relations.

PM The temporal relations inferred via closure on the TimeBank and AQUAINT data by Philippe Muller[7]. These relations replace the original ones, except in files where no relations were inferred (because of temporal inconsistencies).

7 Results

Table 1 shows the performance of the ClearTK-TimeML models across the different tasks when trained on different sets of training data. The "Data" column of each row indicates both the training data sources (as in Section 6), and whether the events and times were predicted by the models ("system") or taken from the annotators ("human"). Performance is reported in terms of strict precision (P), Recall (R) and F1 for event extents, time extents and temporal relations, and in terms of Accuracy (A) on the correctly identified extents for event and time attributes.

[3] http://mallet.cs.umass.edu/

[4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[5] http://opennlp.apache.org/

[6] For BIO token-chunking tasks, LIBLINEAR also had a parameter for how many previous classifications to use as features.

[7] https://groups.google.com/d/topic/tempeval/LJNQKwYHgL8

Data		Event						Time					Relation		
annotation	events	extent			class	tense	aspect	extent			value	type	type		
sources	& times	F1	P	R	A	A	A	F1	P	R	A	A	F1	P	R
TB+BMK	system	77.3	81.9	73.3	84.6	**80.4**	91.0	**82.7**	85.9	79.7	71.7	93.3	**31.0**	34.1	28.4
TB	system	77.3	81.9	73.3	84.6	**80.4**	91.0	**82.7**	85.9	79.7	71.7	93.3	29.8	**34.5**	26.2
TB+AQ	system	78.8	81.4	76.4	86.1	78.2	90.9	77.0	83.2	71.7	69.9	92.9	28.6	30.9	26.6
TB+AQ+PM	system	78.8	81.4	76.4	86.1	78.2	90.9	77.0	83.2	71.7	69.9	92.9	28.5	29.7	27.3
*TB+AQ+SLV	system	80.5	**82.1**	78.9	88.4	71.6	91.2	80.0	**91.6**	71.0	73.6	91.5	27.8	26.5	29.3
Highest in TempEval		81.1	82.0	80.8	89.2	80.4	91.8	82.7	91.4	80.4	86.0	93.7	31.0	34.5	34.4
TB+BMK	human	-	-	-	-	-	-	-	-	-	-	-	**36.3**	37.3	35.2
TB	human	-	-	-	-	-	-	-	-	-	-	-	35.2	**37.6**	33.0
TB+AQ	human	-	-	-	-	-	-	-	-	-	-	-	34.1	33.3	35.0
TB+AQ+PM	human	-	-	-	-	-	-	-	-	-	-	-	35.9	35.2	36.6
*TB+AQ+SLV	human	-	-	-	-	-	-	-	-	-	-	-	**37.7**	34.9	41.0
Highest in TempEval		-	-	-	-	-	-	-	-	-	-	-	36.3	37.6	65.6

Table 1: Performance across different training data. Systems marked with * were tested after the official evaluation. Scores in bold are at least as high as the highest in TempEval.

Training on the AQUAINT (AQ) data in addition to the TimeBank (TB) hurt times and relations. Adding the AQUAINT data caused a -2.7 drop in extent precision, a -8.0 drop in extent recall, a -1.8 drop in value accuracy and a -0.4 drop in type accuracy, and a -3.6 to -4.3 drop in relation recall.

Training on the "Silver" (SLV) data in addition to TB+AQ data gave mixed results. There were big gains for time extent precision (+8.4), time value accuracy (+3.7), event extent recall (+2.5) and event class accuracy (+2.3), but a big drop for event tense accuracy (-6.6). Relation recall improved (+2.7 with system events and times, +6.0 with manual) but precision varied (-4.4 with system, +1.6 with manual).

Adding verb-clause relations (BMK) and closure-inferred relations (PM) increased recall but lowered precision. With system-annotated events and times, the change was +2.2/-0.4 (recall/precision) for verb-clause relations, and +0.7/-1.2 for closure-inferred relations. With manually-annotated events and times, the change was +2.2/-0.3 for verb-clause relations, and (the one exception where recall improved) +1.5/+1.9 for closure-inferred relations.

8 Discussion

Overall, the ClearTK-TimeML ranked 1st in relation F1, time extent strict F1 and event tense accuracy.

Analysis across the different ClearTK-TimeML runs showed that including annotations from the AQUAINT corpus hurt model performance across a variety of tasks. A manual inspection of the AQUAINT corpus revealed many annotation errors, suggesting that the drop may be the result of attempting to learn from inconsistent training data. The AQUAINT corpus may thus have to be partially re-annotated to be useful as a training corpus.

Analysis also showed that adding more relation annotations increased recall, typically at the cost of precision, even though the added annotations were highly accurate: (Bethard et al., 2007) reported agreement of 90%, and temporal closure relations were 100% deterministic from the already-annotated relations. One would expect that adding such high-quality relations would only improve performance. But not all temporal relations were annotated by the TempEval 2013 annotators, so the system could be marked wrong for a finding a true temporal relation that was not noticed by the annotators. Further analysis is necessary to investigate this hypothesis.

Acknowledgements

Thanks to Philippe Muller for providing the closure-inferred relations. The project described was supported in part by Grant Number R01LM010090 from the National Library Of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library Of Medicine or the National Institutes of Health.

References

[Bethard and Martin2006] Steven Bethard and James H. Martin. 2006. Identification of event mentions and their semantic class. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 146154. (Acceptance rate 31%).

[Bethard and Martin2007] Steven Bethard and James H. Martin. 2007. CU-TMP: temporal relation classification using syntactic and semantic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 129–132, Prague, Czech Republic. Association for Computational Linguistics.

[Bethard et al.2007] Steven Bethard, James H. Martin, and Sara Klingenstein. 2007. Finding temporal structure in text: Machine learning of syntactic temporal relations. *International Journal of Semantic Computing*, 01(04):441.

[Llorens et al.2010] Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, page 284291, Uppsala, Sweden, July. Association for Computational Linguistics.

[Llorens et al.2012] Hector Llorens, Leon Derczynski, Robert Gaizauskas, and Estela Saquete. 2012. TIMEN: an open temporal expression normalisation resource. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Ogren et al.2008] Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard. 2008. ClearTK: A UIMA toolkit for statistical natural language processing. In *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*, 5.

[UzZaman and Allen2010] Naushad UzZaman and James Allen. 2010. TRIPS and TRIOS system for TempEval-2: extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, page 276283, Uppsala, Sweden, July. Association for Computational Linguistics.

[UzZaman et al.2013] Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3 evaluating time expressions, events, and temporal relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*. Association for Computational Linguistics, June.

[Verhagen et al.2007] 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 International Workshop on Semantic Evaluations*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

[Verhagen et al.2010] 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*, page 5762, Uppsala, Sweden, July. Association for Computational Linguistics.