ATT1: Temporal Annotation Using Big Windows and Rich Syntactic and Semantic Features

Hyuckchul Jung and **Amanda Stent**
AT&T Labs - Research
180 Park Ave
Florham Park, NJ 07932, USA
hjung, stent@research.att.com

Abstract

In this paper we present the results of experiments comparing (a) rich syntactic and semantic feature sets and (b) big context windows, for the TempEval time expression and event segmentation and classification tasks. We show that it is possible for models using only lexical features to approach the performance of models using rich syntactic and semantic feature sets.

1 Introduction

TempEval-3 Temporal Annotation Task (UzZaman et al., 2012) has three subtasks:

A *Time expression extraction and classification* - extract time expressions from input text, and determine the type and normalised value for each extracted time expression.

B *Event extraction and classification* - extract event mentions from input text, and determine the class, tense and aspect features for each extracted event.

C *Temporal link identification* - identify and categorise temporal links between events in the same or consecutive sentences, events and time expressions in the same sentence, and events and the document creation time of the input text.

Here we report results for the first two tasks.

Previous TempEval competitions have shown that rich syntactic and semantic feature sets can lead to good performance on event and time expression extraction and classification tasks (*e.g.* (Llorens et al., 2010; UzZaman and Allen, 2010)). In this work, we show that with large windows of context, it is possible for models using only lexical features to approach the performance of models using rich syntactic and semantic feature sets.

	Type	Files	EVENT	TIMEX
AQUAINT	gold	73	4431	579
TimeBank	gold	183	6698	1243
TE3-Silver	silver	2452	81329	12739

Table 1: Frequency of event and time expressions in the text portions of the TempEval-3 data sets

2 Data

Using the gold and silver data distributed by the TempEval-3 task organizers (see Table 1), we processed each input file with the Stanford CoreNLP (Stanford Natural Language Processing Group, 2012) and SENNA (Collobert et al., 2011) open-source NLP tools. From the Stanford CoreNLP tools we obtained a tokenization of the input text, the lemma and part of speech (POS) tag for each token, and dependency and constituency parses for each sentence. From SENNA, we obtained a semantic role labelling for each sentence.

3 Approach

We were curious to explore the tradeoff between additional context on the one hand, and additional layers of representation on the other, for the event and time expression extraction tasks. Researchers have investigated the impacts of different sets of features (Adafre and de Rijke, 2005; Angeli et al., 2012;

Feature type	Features	Used in
Lexical 1	token	ATT1, ATT2, ATT3
Lexical 2	lemma	ATT1, ATT2
Part of speech	POS tag	ATT1, ATT2
Dependency	governing verb, governing verb POS, governing preposition, phrase tag, path to root of parse tree, head word, head word lemma, head word POS	ATT1, ATT2
Constituency parse	governing verb, governing verb POS, governing preposition, phrase tag, path to root of parse tree	ATT1, ATT2
Semantic role	semantic role label, semantic role labels along path to root of parse tree	ATT1

Table 2: Features used in our models

Tag type	Tags
time expression extraction tags	B_DATE, B_DURATION, B_SET, B_TIME, I_DATE, I_DURATION, I_SET, I_TIME, O
Event expression extraction tags	B_ACTION, B_ASPECTUAL, B_ACTION, B_OCCURRENCE, B_PERCEPTION, B_REPORTING, B_STATE, O
Event tense	FUTURE, INFINITIVE, PAST, PASTPART, PRESENT, PRESPART, NONE, O
Event aspect	PROGRESSIVE, PREFECTIVE_PROGRESSIVE, PERFECTIVE, NONE, O
Event polarity	NEG, POS
Event modality	'D, CAN, CLOSE, COULD, DELETE, HAVE TO, HAVE_TO, LIKELIHOOD, MAY, MIGHT, MUST, NONE, O, POSSIBLE, POTENTIAL, SHOULD, SHOULD HAVE TO, TO, UNLIKELY, UNTIL, WOULD, WOULD HAVE TO

Table 3: Tags assigned by our classifiers for TempEval-3 tasks A and B

Rigo and Lavelli, 2011). In particular, (Rigo and Lavelli, 2011) also examined performance based on different sizes of n-grams in a small scale (n=1,3).

In this work, we intended to systematically investigate the performance of various models with different layers of representation (based on much larger sets of rich syntactic/semantic features) as well as additional context. For each time expression/event segmentation/classification task, we trained twelve models exploring these two dimensions, three of which we submitted for TempEval-3.

Additional layers of representation We trained three types of model: (ATT1) STANFORD+SENNA, (ATT2) STANFORD and (ATT3) WORDS ONLY. The basic features used in each type of model are given in Table 2: ATT1 models include lexical, syntactic and semantic features, ATT2 models include only lexical and syntactic features, and ATT3 models include only lexical features. For the ATT1 models we had 18 basic features per token, for the ATT2 models we had 16 basic features per token, and for the ATT3 models we had one basic feature per token.

Additional context We experimented with context windows of 0, 1, 3, and 7 words preceding and following the token to be labeled (*i.e.* window sizes of 1, 3, 7, and 15). For each window size, we trained ATT1, ATT2 and ATT3 models. The ATT1 models had 18 basic features per token in the context window, for up to 15 tokens, so up to 270 basic features for each token to be labeled. The ATT2 models had 16 basic features per token in the context

window, so up to 240 basic features for each token to be labeled. The ATT3 models had 1 basic feature per token in the context window, so up to 15 basic features for each token to be labeled.

Model training For event extraction and classification, time expression extraction and classification, and event feature classification, we used the machine learning toolkit LLAMA (Haffner, 2006). LLAMA encodes multiclass classification problems using binary MaxEnt classifiers to increase the speed of training and to scale the method to large data sets. We also used a front-end to LLAMA that builds unigram, bigram and trigram extended features from basic features; for example, from the basic feature "go there today", it would build the features "go", "there", "today", "go there", "there today", and "go there today". We grouped our basic features (see Table 2) by type rather than by token, and the LLAMA front-end then produced ngram features. We chose LLAMA primarily because of the proven power of the ngram feature-extraction front-end for NLP tasks.

4 Event and Time Expression Extraction

For event and time expression extraction, we trained BIO classifiers. A BIO classifier tags each input token as either Beginning, In, or Out of an event/time expression. Our classifier for events simultaneously assigns a B, I or O to each token, and classifies the class of the event for tokens that Begin or are In an event. Our time expression classifier simultaneously assigns a B, I, or O to each token, and classifies the type of the time expression for tokens that Begin or are In a time expression (see Table 3).

A BIO model may sometimes be inconsistent; for example, a token may be labeled as Inside a segment of a particular type, while the previous token may be labeled as Out of any segment. We considered the two most likely labels for each token (as long as each had likelihood at least 0.9), choosing the one most consistent with the context.

5 Event Feature Classification

We determined the event features for each extracted event using four additional classifiers, one each for tense, aspect, polarity and modality. These classifiers were trained only on tokens identified as part of event expressions. Since the event expressions were single words for all but a few (erroneous) cases in the silver data, for determining the event features, we used the same features as before, with the single addition of the event class (during testing, we used the dynamically assigned event class from the event segmentation classifier). As before, we experimented with ATT1, ATT2, and ATT3 models. TempEval-3 only includes evaluation of tense and aspect features, so we only report for those. The tags assigned by each classifier are listed in Table 3.

6 Time Normalization

To compute TIMEX3 standard based values for extracted time expressions, we used the TIMEN (Llorens et al., 2012) and TRIOS (UzZaman and Allen, 2010) time normalizers. Values from the normalizers were validated in post-processing (*e.g.* "T2445" is invalid) and, when the normalizers returned different non-nil values, TIMEN's values were selected without further reasoning. Time normalization was out of scope in our research for this evaluation, but it remains as part of our future work.

7 Results and Discussion

Our results for event segmentation/classification on the TempEval-3 test data are provided in Table 4. The absence of semantic features causes only small changes in F1. The absence of syntactic features causes F1 to drop slightly (less than 2.5% for all but the smallest window size), with recall decreasing while precision improves somewhat. Attribute F1 is also impacted minimally by the absence of semantic features, and about 2-5% by the absence of syntactic features for all but the smallest window size.[1]

Our results for time expression extraction and classification on the TempEval-3 test data are provided in Table 5. Here, the performance drops more in the absence of semantic and syntactic features; however, there is an interaction between length of time expression and performance drop which we may be able to ameliorate in future work by handling consistency issues in the BIO time expression extraction model better.

[1] In Tables 4 and 5, we present results that are slightly different from our submission due to a minor fix in our models by removing some redundant feature values used twice.

Features	Window size	F1	P	R	Class	Tense	Aspect
STANFORD+SENNA	15 (ATT1)	81.16	81.49	80.83	71.60	59.62	73.76
	7	81.08	81.74	80.43	71.49	59.05	73.78
	3	80.35	81.23	79.49	71.41	58.67	73.17
	1	80.94	80.77	81.10	72.37	58.06	73.71
STANFORD	15 (ATT2)	80.86	81.02	80.70	71.05	59.10	73.34
	7	81.30	81.90	80.70	71.57	59.01	74.14
	3	80.87	81.58	80.16	71.94	58.96	73.70
	1	80.78	80.72	80.83	71.80	57.47	73.41
WORDS ONLY	15 (ATT3)	78.58	81.95	75.47	69.5	55.27	70.76
	7	78.40	82.21	74.93	69.14	55.54	70.27
	3	78.14	82.44	74.26	69.39	52.75	70.38
	1	73.55	79.78	68.23	66.33	44.94	63.15

Table 4: Event extraction results (F1, P and R, strict match); feature classification results (attribute F1)

Features	Window size	F1	P	R	Type	Value
STANFORD+SENNA	15 (ATT1)	80.17 (85.95)	93.27 (100)	70.29 (75.36)	77.69	65.29
	7	76.99 (83.68)	91.09 (99.01)	66.67 (72.46)	75.31	64.44
	3	75.52 (83.82)	88.35 (98.06)	65.94 (73.19)	75.52	63.07
	1	66.12 (83.27)	75.70 (95.33)	58.70 (73.91)	72.65	59.59
STANFORD	15 (ATT2)	78.69 (85.25)	90.57 (98.11)	69.57 (75.36)	76.23	65.57
	7	78.51 (84.30)	91.35 (98.08)	68.84 (73.91)	76.03	63.64
	3	78.19 (84.77)	90.48 (98.10)	68.84 (74.64)	75.72	64.20
	1	67.48 (83.74)	76.85 (95.37)	60.14 (74.64)	73.17	59.35
WORDS ONLY	15 (ATT3)	72.34 (80.85)	87.63 (97.94)	61.59 (68.84)	74.04	60.43
	7	72.34 (80.85)	87.63 (97.94)	61.59 (67.84)	74.04	59.57
	3	74.48 (82.85)	88.12 (98.02)	64.49 (71.74)	75.31	61.09
	1	44.62 (82.87)	49.56 (92.04)	40.58 (75.36)	70.92	39.84

Table 5: Time expression extraction results (F1, P and R, strict match with relaxed match in parentheses); attribute F1 for type and value features

A somewhat surprising finding is that both event and time expression extraction are subject to relatively tight constraints from the lexical context. We were surprised by how well the ATT3 (WORDS ONLY) models performed, especially in terms of precision. We were also surprised that the words only models with window sizes of 3 and 7 performed as well as the models with a window size of 15. We think these results are promising for "big data" text analytics, where there may not be time to do heavy preprocessing of input text or to train large models.

8 Future Work

For us, participation in TempEval-3 is a first step in developing a temporal understanding component for text analytics and virtual agents. We now intend to appy our best performing models to this task. In future work, we plan to evaluate our initial results with larger data sets (e.g., cross validation on the tempeval training data) and experiment with hybrid/ensemble methods for performing time expression and temporal link extraction.

Acknowledgments

We thank Srinivas Bangalore, Patrick Haffner, and Sumit Chopra for helpful discussions and for supplying LLAMA and its front-end for our use.

References

S. F. Adafre and M. 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*.

G. Angeli, C. D. Manning, and D. Jurafsky. 2012. Parsing time: Learning to interpret time expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12.

P. Haffner. 2006. Scaling large margin classifiers for spoken language understanding. *Speech Communication*, 48(3–4).

H. Llorens, E. Saquete, and B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

H. Llorens, L. Derczynski, R. Gaizauskas, and E. Saquete. 2012. Timen: An open temporal expression normalisation resource. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

S. Rigo and A. Lavelli. 2011. Multisex - a multilanguage timex sequential extractor. In *Proceedings of Temporal Representation and Reasoning (TIME)*.

Stanford Natural Language Processing Group. 2012. Stanford CoreNLP. `http://nlp.stanford.edu/software/corenlp.shtml`.

N. UzZaman and J. F. Allen. 2010. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

N. UzZaman, H. Llorens, J. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. `http://arxiv.org/abs/1206.5333v1`.