Distinguishing Past, On-going, and Future Events: The EventStatus Corpus

Ruihong Huang
Texas A&M University
huangrh@cse.tamu.edu

Ignacio Cases
Stanford University
cases@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Cleo Condoravdi
Stanford University
cleoc@stanford.edu

Ellen Riloff
University of Utah
riloff@cs.utah.edu

Abstract

Determining whether a major societal event has already happened, is still on-going, or may occur in the future is crucial for event prediction, timeline generation, and news summarization. We introduce a new task and a new corpus, *EventStatus*, which has 4500 English and Spanish articles about civil unrest events labeled as PAST, ON-GOING, or FUTURE. We show that the temporal status of these events is difficult to classify because local tense and aspect cues are often lacking, time expressions are insufficient, and the linguistic contexts have rich semantic compositionality. We explore two approaches for event status classification: (1) a feature-based SVM classifier augmented with a novel induced lexicon of *future-oriented* verbs, such as "threatened" and "planned", and (2) a convolutional neural net. Both types of classifiers improve event status recognition over a state-of-the-art TempEval model, and our analysis offers linguistic insights into the semantic compositionality challenges for this new task.

1 Introduction

When a major societal event is mentioned in the news (e.g., civil unrest, terrorism, natural disaster), it is important to understand whether the event has already happened (PAST), is currently happening (ON-GOING), or may happen in the future (FUTURE). We introduce a new task and corpus for studying the temporal/aspectual properties of major events. The **EventStatus corpus** consists of 4500 English and Spanish news articles about *civil unrest events*, such as protests, demonstrations, marches, and strikes, in which each event is annotated as PAST, ON-GOING, or FUTURE (sublabeled as PLANNED, ALERT or POSSIBLE). This task bridges event extraction research and temporal research in the tradition of TIMEBANK (Pustejovsky et al., 2003) and TempEval (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013). Previous corpora have begun this association: TIMEBANK, for example, includes temporal relations linking events with Document Creation Times (DCT). But the EventStatus task and corpus offers several new research directions.

First, major societal events are often discussed before they happen, or while they are still happening, because they have the potential to impact a large number of people. News outlets frequently report on impending natural disasters (e.g., hurricanes), anticipated disease outbreaks (e.g., Zika virus), threats of terrorism, and plans or warnings of potential civil unrest (e.g., strikes and protests). Traditional event extraction research has focused primarily on recognizing events that have already happened. Furthermore, the linguistic contexts of on-going and future events involve complex compositionality, and features like explicit time expressions are less useful. Our results demonstrate that a state-of-the-art TempEval system has difficulty identifying on-going and future events, mislabeling examples like these:

(1) *The metro workers' strike in Bucharest has entered the fifth day.* (On-Going)
(2) *BBC unions demand more talks amid threat of new strikes.* (Future)
(3) *Pro-reform groups have called for nationwide protests on polling day.* (Future)

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 44–54,
Austin, Texas, November 1-5, 2016.

Second, we intentionally created the EventStatus corpus to concentrate on one particular event frame (class of events): civil unrest. In contrast, previous temporally annotated corpora focus on a wide variety of events. Focusing on one frame (semantic depth instead of breadth) makes this corpus analogous to domain-specific event extraction data sets, and therefore appropriate for evaluating rich tasks like event extraction and temporal question answering, which require more knowledge about event frames and schemata than might be represented in large broad corpora like TIMEBANK (UzZaman et al., 2012; Llorens et al., 2015).

Third, the EventStatus corpus focuses on specific instances of high-level events, in contrast to the low-level and often non-specific or generic events that dominate other temporal datasets.[1] Mentions of specific events are much more likely to be realized in non-finite form (as nouns or infinitives, such as "the strike" or "to protest") than randomly selected event keywords. In breadth-based corpora like the EventCorefBank (ECB) corpus (Bejan and Harabagiu, 2008), 34% of the events have non-finite realization; in TIMEBANK, 45% of the events have non-finite realization. By contrast, in a frame-based corpus like ACE2005 (ACE, 2005), 59% of the events have non-finite forms. In the EventStatus corpus, 80% of the events have non-finite forms. Whether this is due to differences in labeling or to intrinsic properties of these events, the result is that they are much harder to label because tense and aspect are less available than for events realized as finite verbs.

Fourth, the EventStatus data set is multilingual: we collected data from both English and Spanish texts, allowing us to compare events representing the same event frame across two languages that are known to differ in their typological properties for describing events (Talmy, 1985).

Using the new EventStatus corpus, we investigate two approaches for recognizing the temporal status of events. We create a SVM classifier that incorporates features drawn from prior TempEval work (Bethard, 2013; Chambers et al., 2014; Llorens et al., 2010) as well as a new automatically induced lexicon of 411 English and 348 Spanish "future-oriented" matrix verbs—verbs like "threaten" and "fear" whose complement clause or nominal direct object argument is likely to describe a future event. We show that the SVM outperforms a state-of-the-art TempEval system and that the induced lexicon further improves performance for both English and Spanish. We also introduce a Convolutional Neural Network (CNN) to detect the temporal status of events. Our analysis shows that it successfully models semantic compositionality for some challenging temporal contexts. The CNN model again improves performance in both English and Spanish, providing strong initial results for this new task and corpus.

2 The EventStatus Corpus

For major societal events, it can be very important to know whether the event has ended or if it is still in progress (e.g., are people still rioting in the streets?). And sometimes events are anticipated before they actually happen, such as labor strikes, marches and parades, social demonstrations, political events (e.g., debates and elections), and acts of war. The EventStatus corpus represents the *temporal status* of an event as one of five categories:

Past: An event that has started and has ended. There should be no reason to believe that it may still be in progress.
On-going: An event that has started and is still in progress or likely to resume[2] in the immediate future. There should be no reason to believe that it has ended.
Future Planned: An event that has not yet started, but a person or group has planned for or explicitly committed to an instance of the event in the future. There should be near certainty it will happen.
Future Alert: An event that has not yet started, but a person or group has been threatening, warning, or advocating for a future instance of the event.
Future Possible: An event that has not yet started, but the context suggests that its occurrence is a live possibility (e.g., it is anticipated, feared, hinted at, or is mentioned conditionally).

The three subtypes of future events are important

[1] For example in TIMEBANK almost half the annotated events (3720 of 7935) are hypothetical or generic, i.e., PERCEPTION, REPORTING, ASPECTUAL, I_ACTION, STATE or I_STATE rather than the specific OCCURRENCE.

[2] For example, demonstrators have gone home for the day but are expected to return in the morning.

	Past
[EN]	Today's *demonstration* ended without violence.
	An estimated 2,000 people *protested* against the government in Peru.
[SP]	Terminó la *manifestación* de los kurdos en la UNESCO de París.
	On-going
[EN]	Negotiations continue with no end in sight for the 2 week old *strike*.
	Yesterday's rallies have caused police to fear more today.
[SP]	Pacifistas latinoamericanos no cesan sus *protestas* contra guerra en Irak.
	Future Planned
[EN]	77 percent of German steelworkers voted to *strike* to raise their wages.
	Peace groups have already started organizing mass *protests* in Sydney.
[SP]	Miedo en la City en víspera de masivas *protestas* que la toman por blanco.
	Future Alert
[EN]	Farmers have threatened to hold *demonstrations* on Monday.
	Nurses are warning they intend to walkout if conditions don't improve.
[SP]	Indigenas hondureños amenazan con declararse en *huelga* de hambre.
	Future Possible
[EN]	Residents fear *riots* if the policeman who killed the boy is acquitted.
	The military is preparing for possible protests at the G8 summit.
[SP]	Policía Militar analiza la posibilidad de decretar una *huelga* nacional.

Table 1: Examples of event status categories for civil unrest events, showing two examples in English [EN] and one in Spanish [SP].

in marking not just temporal status but also what we might call predictive status. Events very likely to occur are distinguished from events whose occurrence depends on other contingencies (Future Planned vs. Alert/Possible). Warnings or mentions of a potential event by a likely actor are further distinguished from events whose occurrence is more open-ended (Future Alert vs. Possible). The status of future events is not due just to lexical semantics or local context but also other qualifiers in the sentence (e.g. "may"), the larger discourse context, and world knowledge. The annotation guidelines are formulated with that in mind. The categories for future events are not incompatible with one another but are meant to be informationally ordered (e.g. "future alert" implies "future possible"). Annotators are instructed to go for the strongest implication supported by the overall context. Table 1 presents examples of each category in news reports about civil unrest events, with the event keywords in *italics*.

2.1 EventStatus Annotations

The EventStatus dataset consists of English and Spanish news articles. We manually identified 6 English words[3] and 13 Spanish words[4] and phrases associated with civil unrest events, and added their morphological variants. We then randomly selected 2954 and 1491[5] news stories from the English Gigaword 5th Ed. (Parker et al., 2011) and Spanish Gigaword 3rd Ed. (Mendon et al., 2011) corpora, respectively, that contain at least one civil unrest phrase. Events of a specific type are very sparsely distributed in a large corpus like the Gigaword, so we used keyword matching just as a first pass to identify candidate event mentions.

[3]The English keywords are "protest", "strike", "march", "rally", "riot" and "occupy". These correspond to the most frequent words in the relevant frame in the Media Frames corpus (Card et al., 2015). Because "march" most commonly refers to the month, we removed the word itself and only kept its other morphological variations.

[4]Spanish keywords: "marchar", "protestar", "amotinar(se)", "manifestar(se)", "huelga", "manifestación", "disturbio", "motín", "ocupar * la calle", "tomar * la calle", "salir * las calles", "lanzarse a las calles", "cacerolas vacías", "cacerolazo", "cacerolada". Asterisks could be replaced by up to 4 words. The last three terms are common expressions for protest marches in many countries of Latin America and Spain.

[5]46 (out of 3000) and 9 (out of 1500) stories were removed due to keyword errors.

	Past	Ongoing	Future (Plan,Alert,Possible)	Multiple	Not Event
EN	1735	583	292 (197,48,47)	28	186
SP	1545	739	360 (279,61,30)	21	72

Table 2: Counts of Temporal Status Labels in EventStatus.

Because many keyword instances don't refer to a specific event, primarily due to lexical ambiguity and generic descriptions (e.g., *"Protests are often facilitated by ..."*), we used a two-stage annotation process. First, we extracted sentences containing at least one key phrase, and had three human annotators judge whether the sentence describes a specific civil unrest event. Next, for each sentence that mentions a specific event, the annotators assigned an event status to every civil unrest key phrase in that sentence. In both annotation phases, we asked the annotators to consider the context of the entire article.

In the first annotation phase, the average pairwise inter-annotator agreement (Cohen's κ) among the annotators was $\kappa = 0.84$ on the English data and 0.70 on the Spanish data. We then assigned the majority label among the three annotators to each sentence. In the English data, of the 5085 sentences with at least one key phrase, 2492 (49%) were judged to be about a specific civil unrest event. In the Spanish data, 3249 sentences contained at least one key phrase and 2466 (76%) described a specific event.

In the second phase, the annotators assigned one of the five temporal status categories listed in Section 2 to each event keyword in a relevant sentence. In addition, we provided a *Not Event* label.[6] Occasionally, a single instance of a keyword can refer to multiple events (e.g., *"Both last week's and today's protests..."*), so we permitted multiple labels to be assigned to an event phrase. However this happened for only 28 cases in English and 21 cases in Spanish.

The average pairwise inter-annotator agreement among the three human annotators for the temporal status labels was κ=.78 for English and κ=.80 for Spanish. We used the majority label among the three annotators as the gold status. In total, 2907 English and 2807 Spanish event phrases exist in the relevant sentences and were annotated. However there were 83 English cases (≈2.9%) and 70 Spanish cases (≈2.5%) where the labels among the three annotators were all different, so we discarded these cases. Table 2 shows the final distribution of labels in the EventStatus corpus. The EventStatus corpus[7] is available through the LDC.

[6] A sentence can contain multiple keyword instances. So even in a relevant sentence, some instances may not refer to a specific event.

2.2 Linguistic Properties of Event Mentions

Next, we investigated the linguistic properties of the event status categories, lumping together the 3 future subcategories. Table 3 shows the distribution of syntactic forms of the event mentions in two commonly used event datasets, ACE2005 (ACE, 2005) and EventCorefBank (Bejan and Harabagiu, 2008), and our new EventStatus corpus. In the introduction, we mentioned the high frequency of non-finite event expressions; Table 3 provides the evidence: non-finite forms (nouns and infinitives) constitute 59% in ACE2005, 34% in EventCorefBank, and a very high 80% of the events in the EventStatus dataset. The distribution is even more skewed for future events, which are 95% (English) and 96% (Spanish) realized by non-finite surface forms.

	Finite Verbs	Nouns	Inf. Verbs	Other
ACE Dataset				
	2201 (41)	2566 (48)	352 (7)	243 (5)
ECB Dataset				
	1151 (66)	488 (28)	77 (4)	25 (1)
EventStatus, English Section				
PA	331 (19)	1295 (75)	103 (6)	6 (0)
OG	58 (10)	476 (82)	29 (5)	20 (3)
FU	15 (5)	245 (84)	32 (11)	0 (0)
EventStatus, Spanish Section				
PA	315 (20)	1145 (74)	84 (5)	1 (0)
OG	41 (6)	685 (93)	12 (2)	1 (0)
FU	14 (4)	309 (86)	36 (10)	1 (0)

Table 3: Number and % (in parentheses) of event mentions by syntactic form. PA = Past; OG = On-going; FU = Future

2.3 Future Oriented Verbs

We observed that many future event mentions are preceded by a set of lexical (non-aux) verbs that we call *future oriented verbs*, such as "threatened" in (4) and "fear" in (5). These verbs project the events in the lower clause into the future.

[7] http://faculty.cse.tamu.edu/huangrh/EventStatus_corpus.html

(4) *They threatened to protest if Kmart does not acknowledge their request for a meeting.*

(5) *People fear renewed rioting during the coming days.*

Categories of future oriented verbs include mental activity ("anticipate", "expect"), affective ("fear", "worry"), planning ("plan", "prepare", "schedule"), threatening ("threaten", "advocate", "warn"), and inchoative verbs ("start", "initiate", and "launch"). We found that these categories correlate with the predictive status of the events they embed. We drew on these insights to induce a lexicon of future oriented verbs.

We harvested matrix verbs whose complement unambiguously describes a future event using two heuristics. One heuristic looks for examples with a tense conflict between the matrix verb and its complement: a matrix verb in the past tense (like "planned" below) whose complement event is an infinitive verb or deverbal noun modified by a future time expression (like "tomorrow" or "next week"), hence in the future (e.g., "strike" below): [8]

(6) *The union planned to strike next week.*

Future events are often marked by conditional clauses, so the second heuristic considers an event to be future if it was post-modified by a conditional clause (beginning with "if" or "unless"):

(7) *The union threatened to strike if their appeal was rejected.*

Finally, to increase precision, we only harvested a verb as future-oriented if it functioned as a matrix both in sentences with an embedded future time expression and in sentences with a conditional clause.

Future Oriented Verb Categories: We ran the algorithm on the English and Spanish Gigaword corpora (Parker et al., 2011; Mendon et al., 2011), obtaining 411 English verbs and 348 Spanish verbs. To better understand the structure of the learned lexicon, we mapped each English verb to Framenet (Baker et al., 1998); 86% (355) of the English verbs occurred in Framenet, in 306 unique frames. We clustered these into 102 frames[9] and grouped the Spanish verbs following English Framenet, identifying 67 categories. (Some learned verbs, such as "poise" , "slate" , "compel" and "hesitate", had a clear future orientation but didn't exist in Framenet.) Table 4 shows examples of learned verbs for English and their categories.

Commitment: threaten, vow, promise, pledge, commit, declare, claim, volunteer, anticipate
Coming_to_be: enter, emerge, plunge, kick, mount reach, edge, soar, promote, increase, climb, double
Purpose: plan, intend, project, aim, object, target
Permitting: allow, permit, approve, subpoena
Experiencer_subj: fear, scare, hate
Waiting: expect, wait
Scheduling: arrange, schedule
Deciding: decide, opt, elect, pick, select, settle
Request: ask, urge, order, encourage, demand, appeal, request, summon, implore, advise, invite
Evoking: raise, press, back, recall, pressure, force, rush, pull, drag, respond

Table 4: Examples from Future Oriented Verb Lexicon

In the next sections we propose two classifiers, an SVM classifier using standard TempEval features plus our new future-oriented lexicon, and a Convolutional Neural Net, as a pilot exploration of what features and architecture work well for the EventStatus task. For these studies we combine the Future Planned, Future Alert and Future Possible categories into a single Future event status because we first wanted to establish how well classifiers can detect the primary temporal distinctions between Past vs. Ongoing vs. Future. The future subcategories are, of course, relatively smaller and we expect that the most effective approach will be to design a classifier that sits on top of the primary classifier to further subcategorize the Future instances. We leave the task of subcategorizing future events for later work.

[8] For English, we extract events linked by the "xcomp" dependency using the Stanford dependency parser (Marneffe et al., 2006), with a future time expression attached to the second event with the "tmod" relation. For Spanish, we consider two events related if they are at most 5 words apart, and the second event is modified by a time expression, at most 5 words apart.

[9] By merging frames that share frame elements (e.g., "Purpose" and "Project" share the frame element "plan")

3 SVM Event Status Model

Our first classifier is a linear SVM classifier.[10] We trained three binary classifiers (one per class) using one-vs.-rest, and label an event mention with the class that assigned the highest score to the mention. We used features inspired by prior TempEval work and by the previous analysis, including words, tense and aspect features, time expressions, and the new future-oriented verb lexicon. We also experimented with other features used by TempEval systems (including bigrams, POS tags, and two-hop dependency features), but they did not improve performance.[11]

Bag-Of-Words Features: For bag-of-words unigram features we used a window size of 7 (7 left and 7 right) for the English data and 6 for the Spanish data; this size was optimized on the tuning sets.

Tense, Aspect and Time Expressions: Because these features are known to be the most important for relating events to document creation time (Bethard, 2013; Llorens et al., 2010), we used TIPSem (Llorens et al., 2010) to generate the tense and aspect of events and find time expressions in both languages. TIPSem infers the tense and aspect of nominal and infinitival event mentions using heuristics without relying on syntactic dependencies. For the English data set, we also generated syntactic dependencies using Stanford CoreNLP (Marneffe et al., 2006) and applied several rules to create additional tense and aspect features based on the governing words of event mentions[12]. Time indication features are created by comparing document creation time to time expressions linked to an event mention detected by TIPSem. If TIPSem detects no linked time expressions for an event mention, we take the nearest time expression in the same sentence.

Governing Words: Governing words have been useful in prior work. Our version of the feature pairs the governing word of an event mention with the dependency relation in between. We used Stanford CoreNLP (Marneffe et al., 2006) to generate dependencies for the English data. For the Spanish data, we used Stanford CoreNLP to generate Part-of-Speech tags[13] and then applied the MaltParser (Nivre et al., 2004) to generate dependencies.

4 Convolutional Neural Network Model

Convolutional neural networks (CNNs) have been shown to be effective in modeling natural language semantics (Collobert et al., 2011). We were especially keen to find out whether the convolution operations of CNNs can model the semantic compositionality needed to detect temporal-aspectual status. For our experiments, we trained a simple CNN with one convolution layer followed by one max pooling layer (Kim, 2014; Collobert et al., 2011),

The convolution layer has 300 hidden units. In each unit, the same affine transformation is applied to every consecutive 5 words (a filter instance) in the input sequence of words. A different affine transformation is applied to each hidden unit. After each affine transformation, a Rectified Linear Units (ReLU) (Nair and Hinton, 2010) non-linearity is applied. For each hidden unit, the max pooling layer selects the maximum value from the pool of real values generated from each filter instance.

After the max pooling layer, a *softmax* classifier predicts probabilites for each of the three classes, Past, Ongoing and Future. To alleviate overfitting of the CNN model, we applied dropout (Hinton et al., 2012) on the convolution layer and the following pooling layer with a keeping rate of 0.5.

Our experiments used the 300-dimension English word2vec embeddings[14] trained on 100 billion words of Google News. We trained our own 300-dimension Spanish embeddings, running word2vec (Mikolov et al., 2013) over both Spanish Gigaword (Mendon et al., 2011)— tokenized using Stanford CoreNLP SpanishTokenizer (Manning et al., 2014)— and the pre-tokenized Spanish Wikipedia dump (Al-Rfou et al., 2013). The vectors were then tuned during backpropagation for our specific task.

[10] Trained using LIBSVM (Chang and Lin, 2011) with linear kernels (polynomial kernels yielded worse performance).

[11] Previous TempEval work reported that those additional features were useful when computing temporal relations between two events but not when relating an event to the Document Creation Time, for which tense, aspect, and time expression features were the most useful (Llorens et al., 2010; Bethard, 2013).

[12] We did not imitate this procedure for Spanish because the quality of our generated Spanish dependencies is poor.

[13] Stanford CoreNLP has no support for generating syntactic dependencies for Spanish.

[14] docs.google.com/uc?id=0B7XkCwpI5KDYNlNUTTlSS21pQmM.

Row	Method	PA	OG	FU	Macro	Micro
1	**TIPSem**	26/80/39	8/32/13	4/23/7	13/45/20	20/68/31
2	**TIPSem with transitivity**	75/76/75	14/22/17	4/21/7	31/40/35	55/67/61
3	**SVM with all features**	91/81/86	33/47/39	45/58/51	56/62/59	75/75/75
4	**SVM with BOW features only**	88/80/84	37/46/41	40/53/45	55/60/57	72/72/72
5	**+Tense/Aspect/Time**	89/81/85	40/50/44	42/52/46	57/61/59	73/73/73
6	**+Governing Word**	90/81/85	43/56/48	42/55/47	58/64/61	75/75/75
7	**+Future Oriented Lexicon**	90/82/86	44/56/49	48/62/54	61/66/63	76/76/76
8	**Convolutional Neural Net**	91/83/**87**	46/57/**51**	49/67/**57**	62/69/**65**	77/77/**77**

Table 6: Experimental Results on English Data. Each cell shows Recall/Precision/F-score.

Row	Method	PA	OG	FU	Macro	Micro
1	**TIPSem**	19/84/31	14/38/20	4/53/8	12/58/20	16/65/25
2	**TIPSem with transitivity**	69/70/70	40/35/37	12/62/20	40/56/47	54/59/56
3	**SVM with all features**	84/77/80	48/51/49	42/57/48	58/62/60	69/69/69
4	**SVM with BOW features only**	82/75/78	53/56/54	34/52/41	56/61/59	68/68/68
5	**+Tense/Aspect/Time**	82/77/79	55/57/56	45/61/52	61/65/63	70/70/70
6	**+Governing Word**	83/75/79	51/56/53	42/58/49	59/63/61	69/69/69
7	**+Future Oriented Lexicon**	82/77/79	55/57/56	47/63/**54**	61/65/63	70/70/70
8	**Convolutional Neural Net**	84/80/**82**	60/58/**59**	44/59/50	62/66/**64**	72/72/**72**

Table 7: Experimental Results on Spanish Data. Each cell shows Recall/Precision/F-score.

	PA	OG	FU
English	1385 (68%)	427 (21%)	233 (11%)
Spanish	1251 (59%)	589 (28%)	280 (13%)

Table 5: Label Distributions in the Test Set

5 Evaluations

For all subsequent evaluations, we use gold event mentions. We randomly sampled around 20% of the annotated documents as the parameter tuning set and used the rest as the test set. Rather than training once on a distinct training set, all our experiment results are based on 10-fold cross validation on the test set, (1191 Spanish documents, 2364 English documents; see Table 5 for the distribution of event mentions).

5.1 Comparing with a TempEval System

We begin with a baseline: applying a TempEval system to classify each event. Most of our features are already drawn from TempEval, but our goal was to see if an off-the-shelf system could be directly applied to our task. We chose TIPSem (Llorens et al., 2010), a CRF system trained on TimeBank that uses linguistic features, has achieved top performance in TempEval competitions for both English and Spanish (Verhagen et al., 2010), and can compute the relation of each event with the Document Creation Time. We applied TIPSem to our test set, mapping the DCT relations to our three event status classes[15].

Row 1 of Tables 6 and 7 shows TIPSem results. The columns show results for each category separately, as well as macro-average and micro-average results across the three categories. Each cell shows the Recall/Precision/F-score numbers. Since TIPSem linked relatively few event mentions to the DCT, we next leveraged the transitivity of temporal relations (UzZaman et al., 2012; Llorens et al., 2015), linking an event to a DCT if the temporal relation between another event in the same sentence and the DCT is transferable. For instance, if event A is AFTER its DCT, and event B is AFTER event A, then event B is also AFTER the DCT.[16] Row 2 shows the results of TIPSem with temporal transitivity.

Even augmented by transitivity, TIPSem fails to detect many Ongoing (**OG**) and Future (**FU**) events; most mislabeled OG and FU events were nominal. Confusion matrices (Table 8) show that most of the

[15]We used the obvious mappings from TIPSem relations: "BEFORE" to "PA", "AFTER" to "FU" , and "INCLUDES" (for English) and "OVERLAP" (for Spanish) to "OG".

[16]Some transitivity rules are ambiguous: if event A is AFTER DCT, event B INCLUDES event A, event B can be AFTER or INCLUDES DCT. We ran experiments and chose rules that improved performance the most for TipSem.

missed OG events were labeled as Past (PA) while FU events were commonly mislabeled as both PA and OG. Below are some examples of OG and FU events mislabeled as PA:

(8) Jego said Sunday on arriving in Guadeloupe that he would stay as long as it took to bring an end to the strike organised by the Collective against Extreme Exploitation (LKP). (OG)

(9) A massive protest planned for Kathmandu on Tuesday has been re-baptised a victory parade. (FU)

	Predicted (EN)			Predicted (SP)		
	PA	**OG**	**FU**	**PA**	**OG**	**FU**
Gold **PA**	718	96	15	653	231	6
Gold **OG**	156	35	11	196	160	10
Gold **FU**	72	30	7	78	72	26

Table 8: Confusion Matrices for TIPSem (with transitivity).

SVM Results Next, we compare TIPSem's results with our SVM classifier. An issue is that TIPSem identifies only 72% and 78% of the gold event mentions, for English and Spanish respectively[17]. To have a fair comparison, we applied the SVM to only the event mentions that TipSem recognized. Row 3 shows these results for the SVM classifier using its full feature set. The SVM outperforms TipSem on all three categories, for both languages, with the largest improvements on Future events.

Next, we ran ablation experiments with the SVM to evaluate the impact of different subsets of its features. For these experiments, we applied the SVM to all gold event mentions, thus Rows 1-3 of Tables 6 and 7 report on fewer event mentions than rows 4-8. Row 4 shows results using only bag-of-words features[18]. Row 5 shows results when additionally including the tense, aspect, and time features provided by TIPSem (Llorens et al., 2010). Unsurprisingly, in both languages[19] these features improve over just bag-of-word features.

Row 6 further adds governing word features. These improve English performance, especially for On-Going events. For Spanish, governing word features slightly decrease performance, likely due to the poor quality of the Spanish dependencies.

Row 7 adds the future oriented lexicon features[20]. For both English and Spanish, the future oriented lexicon increased overall performance, and (as expected) especially for Future events.

CNN Results Row 8 shows the results using CNN models. For English and Spanish, the same window (7 words for English, 6 words for Spanish) was used to compute bag-of-word features for SVMs as for training the CNN models. For English, the CNN model further increased recall and precision across all three classes. The CNN improved Spanish performance on both Past and On-going events, but the SVM outperformed the CNN for Future events when the future oriented lexicon features were included.

6 Analysis

To better understand whether the CNN model's strong performance was related to handling compositionality, we examined some English examples that were correctly recognized by the CNN model but mislabeled by the SVM classifier with bag-of-words features. The examples below (event mentions are in *italics*) suggest that the CNN may be capturing the compositional impact of local cues like "possibility" or "since":

(10) Raising the possibility of a *strike* on New Year's Eve, the president of New York City's largest union is calling for a 30 percent raise over three years. (FU)

(11) The lockout was announced in the wake of a go-slow and partial *strike* by the union since July 12 after management turned down its demand. (OG)

We also conducted an error analysis by randomly sampling and then analyzing 50 of the 473 errors by the CNN model. Many cases (26/50) are ambiguous from the sentence alone, requiring discourse information. The first example below is caused by the well-known "double access" ambiguity of the complement of a communication verb (Smith, 1978; Abusch, 1997; Giorgi, 2010).

(12) Chavez also said he discussed the *strike* with UN Secretary General Kofi Annan and told him the strike organizers were "terrorists." (OG)

[17] We were not able to decouple TipSem's event recognition component and force it to process all event mentions.

[18] Replacing each word feature with a word2vec embedding resulted in slightly worse performance.

[19] We always obtain even recall and precision for the micro average metric because we only apply classifiers to event mentions that refer to a civil unrest event.

[20] For Spanish, we removed the governing word features because of the poor quality of the Spanish dependencies.

(13) Students and teachers *protest* over education budget (PA)

In 9/50 cases, the contexts that imply temporal status are complex and fall out of our ± 7 word range, e.g.,:

(14) Protesters on Saturday also *occupied* two gymnastics halls near Gorleben which are to be used as accommodation for police. They were later forcibly dispersed by policemen. (PA)

The remaining 15/50 cases contain enough local cues to be solvable by humans, but both the CNN and SVM models nonetheless failed:

(15) Eastern leaders have grown weary of the *protest* movement led mostly by Aymara. (OG)

7 Related Work

Our work overlaps with two communities of tasks and corpora: the task of classifying temporal order between event mentions and Document Creation Time (DCT) in TempEval (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), and the task of extracting events, associated with corpora such as ACE2005 (ACE, 2005) and the EventCorefBank (ECB) (Bejan and Harabagiu, 2008). By studying the events in a particular frame (civil unrest), but focusing on their temporal status, our work has the potential to draw these communities together. Most event extraction work (Freitag, 1998; Appelt et al., 1993; Ciravegna, 2001; Chieu and Ng, 2002; Riloff and Jones, 1999; Roth and Yih, 2001; Zelenko et al., 2003; Bunescu and Mooney, 2007) has focused on extracting event slots or frames for past events and assigning dates. The TempEval task of linking events to DCT has not focused on events that tend to have non-finite realizations, nor has it focused on subtypes of future events. Our work, including the corpus and the future-oriented verb lexicon, has the potential to benefit related tasks like generating event timelines from news articles (Allan et al., 2000; Yan et al., 2011) or social media sources (Li and Cardie, 2014; Ritter et al., 2012), or exploring the psychological implications of future oriented language (Nie et al., 2015; Schwartz et al., 2015).

8 Conclusions

We have proposed a new task of recognizing the past, on-going, or future temporal status of major events, introducing a new resource for studying events in two languages. Besides its importance for studying time and aspectuality, the EventStatus dataset offers a rich resource for any future investigation of information extraction from major societal events.

The strong performance of the convolutional net system suggests the power of latent representations to model temporal compositionality, and points to extensions of our work using deeper and more powerful networks.

Finally, our investigation of the role of context and semantic composition in conveying temporal information also has implications for our understanding of temporality and aspectuality and their linguistic expression. Many of the errors made by our CNN system are complex ambiguities, like the double access readings, that cannot be solved without information from the wider discourse context. Our work can thus also be seen as a call for the further use of rich discourse information in the computational study of temporal processing.

9 Acknowledgments

We want to thank the Stanford NLP group and especially Danqi Chen for valuable inputs, and Michael Zeleznik for helping us refine the categories and for masterfully orchestrating the annotation efforts. We also thank Luke Zettlemoyer and all our reviewers for providing useful comments. This work was partially supported by the National Science Foundation via NSF Award IIS-1514268, by the Defense Advanced Research Projects Agency (DARPA) Deep Exploration and Filtering of Text (DEFT) Program under Air Force Research Laboratory (AFRL) contract no. FA8750-13-2-0040, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: the views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, NSF, IARPA, DoI/NBC, or the U.S. Government.

References

Dorit Abusch. 1997. Sequence of tense and temporal *de re*. *Linguistics & Philosophy*, 20:1–50.

ACE. 2005. NIST ACE evaluation website. In *http://www.nist.gov/speech/tests/ace/2005*.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

J. Allan, V. Lavrenko, D. Malin, and R. Swan. 2000. Detections, Bounds, and Timelines: Umass and TDT-3. In *Proceedings of Topic Detection and Tracking Workshop*.

D. Appelt, J. Hobbs, J. Bear, D. Israel, and M. Tyson. 1993. FASTUS: a Finite-state Processor for Information Extraction from Real-world Text. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI)*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *In Proceedings of COLING/ACL*, pages 86–90.

C. Bejan and S. Harabagiu. 2008. A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

S. Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM)*.

R. Bunescu and R. Mooney. 2007. Learning to Extract Relations from the Web using Minimal Supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Nathanael Chambers, Bill McDowell, Taylor Cassidy, and Steve Bethard. 2014. Dense event ordering with a multi-pass architecture. In *Transactions of the Association for Computational Linguistics (TACL)*.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

H.L. Chieu and H.T. Ng. 2002. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. In *Proceedings of the 18th National Conference on Artificial Intelligence*.

F. Ciravegna. 2001. Adaptive Information Extraction from Text by Rule Induction and Generalisation. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuglu, and P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. In *Journal of Machine Learning Research*.

Dayne Freitag. 1998. Toward General-Purpose Learning for Information Extraction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*.

Alessandra Giorgi. 2010. *About the speaker: towards a syntax of indexicality*. Oxford University Press, Oxford.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. 2012. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. In *arXiv preprint arXiv:1207.0580*.

Y. Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of 2014 the Conference on Empirical Methods in Natural Language Processing (EMNLP-2014)*.

J. Li and C. Cardie. 2014. Timeline Generation: Tracking Individuals on Twitter. In *Proceedings of the 23rd International Conference on World Wide Web*.

H. Llorens, E. Saquete, and B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

H. Llorens, N. Chambers, N. UzZaman, Mostafazadeh N., J. Allen, and J. Pustejovsky. 2015. Semeval-2015 Task 5: QA TempEval - Evaluating Temporal Information Understanding with Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–60.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC-2006)*.

Angelo Mendon ca, Daniel Jaquette, David Graff, and Denise DiPersio. 2011. Spanish Gigaword Third Edition. In *Linguistic Data Consortium*.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*.

V. Nair and G. E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of 27th International Conference on Machine Learning*.

A. Nie, J. Shepard, J. Choi, B. Copley, and P. Wolff. 2015. Computational Exploration of the Linguistic Structures of Future-Oriented Expression: Classification and Categorization. In *Proceedings of the NAACL Student Research Workshop (NAACL-SRW'15)*.

J. Nivre, J. Hall, and J. Nilsson. 2004. Memory-Based Dependency Parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword. In *Linguistic Data Consortium*.

J. Pustejovsky, P. Hanks, R. Saur, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*.

E. Riloff and R. Jones. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.

A. Ritter, Mausam, O. Etzioni, and S. Clark. 2012. Open Domain Event Extraction from Twitter. In *The 18th ACM SIGKDD Knowledge Discovery and Data Mining Conference*.

D. Roth and W. Yih. 2001. Relational Learning via Propositional Algorithms: An Information Extraction Case Study. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 1257–1263, Seattle, WA, August.

A. Schwartz, G. Park, M. Sap, E. Weingarten, J. Eichstaedt, M. Kern, J. Berger, M. Seligman, and L. Ungar. 2015. Extracting Human Temporal Orientation in Facebook Language. In *Proceedings of the The 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*.

Carlota Smith. 1978. The syntax and interpretation of temporal expressions in English. *Linguistics & Philosophy*, 2:43–99.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. In Timothy Shopen, editor, *Language Typology and Syntactic Description, Volume 3*. Cambridge University Press.

N. UzZaman, H. Llorens, and J. Allen. 2012. Evaluating Temporal Information Understanding with Temporal Question Answering. In *Proceedings of IEEE International Conference on Semantic Computing*.

N. UzZaman, H. Llorens, J. Allen, L. Derczynski, M. Verhagen, and J. 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)*.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*.

M. Verhagen, R. Sauri, T. Caselli, and J. Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

R. Yan, L. Kong, C. Huang, X. Wan, X. Li, and Y. Zhang. 2011. Timeline Generation through Evolutionary Trans-temporal Summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel Methods for Relation Extraction. *Journal of Machine Learning Research*, 3.