Exploiting Discourse Analysis for Article-Wide Temporal Classification

Jun-Ping Ng[1], **Min-Yen Kan**[1,2], **Ziheng Lin**[3], **Wei Feng**[4], **Bin Chen**[5], **Jian Su**[5], **Chew-Lim Tan**[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Interactive and Digital Media Institute, National University of Singapore, Singapore
[3]Research & Innovation, SAP Asia Pte Ltd, Singapore
[4]Department of Computer Science, University of Toronto, Canada
[5]Institute for Infocomm Research, Singapore
junping@comp.nus.edu.sg

Abstract

In this paper we classify the temporal relations between pairs of events on an article-wide basis. This is in contrast to much of the existing literature which focuses on just event pairs which are found within the same or adjacent sentences. To achieve this, we leverage on discourse analysis as we believe that it provides more useful semantic information than typical lexico-syntactic features. We propose the use of several discourse analysis frameworks, including 1) Rhetorical Structure Theory (RST), 2) PDTB-styled discourse relations, and 3) topical text segmentation. We explain how features derived from these frameworks can be effectively used with support vector machines (SVM) paired with convolution kernels. Experiments show that our proposal is effective in improving on the state-of-the-art significantly by as much as 16% in terms of F_1, even if we only adopt less-than-perfect automatic discourse analyzers and parsers. Making use of more accurate discourse analysis can further boost gains to 35%.

1 Introduction

A good amount of research had been invested in understanding temporal relationships within text. Particular areas of interest include determining the relationship between an event mention and a time expression (timex), as well as determining the relationship between two event mentions. The latter, which we refer to as event-event (*E-E*) temporal classification is the focus of this work.

For a given event pair which consists of two events $e1$ and $e2$ found *anywhere* within an article, we want to be able to determine if $e1$ happens before $e2$ (BEFORE), after $e2$ (AFTER), or within the same time span as $e2$ (OVERLAP).

Consider this sentence[1]:

> At least 19 people were **killed** and 114 people were **wounded** in Tuesday's southern Philippines airport blast, officials **said**, but reports said the death toll could climb to 30. (1)

Three event mentions found within the sentence are bolded. We say that there is an OVERLAP relationship between the "**killed** – **wounded**" event pair as these two events happened together after the airport blast. Similarly there is a BEFORE relationship between both the "**killed** – **said**", and "**wounded** – **said**" event pairs, as the death and injuries happened before reports from the officials.

Being able to infer these temporal relationships allows us to build up a better understanding of the text in question, and can aid several natural language understanding tasks such as information extraction and text summarization. For example, we can build up a temporal characterization of an article by constructing a temporal graph denoting the relationships between all events within an article (Verhagen et al., 2009). This can then be used to help construct an event timeline which layouts sequentially event mentions in the order they take place (Do et al., 2012). The temporal graph can also be used in text summarization, where temporal order can be used to improve sentence ordering and thereby the eventual generated summary (Barzilay et al., 2002).

Given the importance and value of temporal relations, the community has organized shared tasks

[1]From article AFP_ENG_20030304.0250 of the ACE 2005 corpus (ACE, 2005).

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 12–23,
Seattle, Washington, USA, 18-21 October 2013.

to spur research efforts in this area, including the TempEval-1, -2 and -3 evaluation workshops (Verhagen et al., 2009; Verhagen et al., 2010; Uzzaman et al., 2012). Most related work in this area have focused primarily on the task defintitions of these evaluation workshops. In the task definitions, *E-E* temporal classification involves determining the relationship between events found within the same sentence, or in adjacent sentences. For brevity we will refer to this loosely as intra-sentence *E-E* temporal classification in the rest of this paper.

This definition however is limiting and insufficient. It was adopted as a trade-off between completeness, and the need to simplify the evaluation process (Verhagen et al., 2009). In particular, one deficiency is that it does not allow us to construct the complete temporal graph we seek. As illustrated in Figure 1, being able to perform only intra-sentence *E-E* temporal classification may result in a forest of disconnected temporal graphs. A sentence $s3$ separates events C and D, as such an intra-sentence *E-E* classification system will not be able to determine the temporal relationship between them. While we can determine the relationship between A and C in the figure with the use of temporal transitivity rules (Setzer et al., 2003; Verhagen, 2005), we cannot reliably determine the relationship between say A and D.

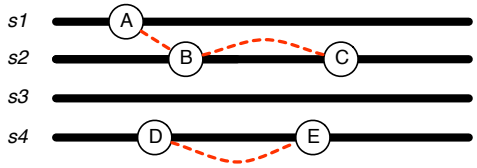

Figure 1: A disconnected temporal graph of events within an article. Horizontal lines depict sentences $s1$ to $s4$, and the circles identify events of interest.

In this work, we seek to overcome this limitation, and study what can enable effective article-wide *E-E* temporal classification. That is, we want to be able to determine the temporal relationship between two events located anywhere within an article.

The main contribution of our work is going beyond the surface lexical and syntactic features commonly adopted by existing state-of-the-art approaches. We suggest making use of semantically motivated features derived from discourse analysis instead, and show that these discourse features are superior.

While we are just focusing on *E-E* temporal classification, our work can complement other approaches such as the joint inference approach proposed by Do et al. (2012) and Yoshikawa et al. (2009) which builds on top of event-timex (*E-T*) and *E-E* temporal classification systems. We believe that improvements to the underlying *E-T* and *E-E* classification systems will help with global inference.

2 Related Work

Many researchers have worked on the *E-E* temporal classification problem, especially as part of the TempEval series of evaluation workshops. Bethard and Martin (2007) presented one of the earliest supervised machine learning systems, making use of support vector machines (SVM) with a variety of lexical and syntactic features. Kolya et al. (2010) described a conditional random field (CRF) based learner making use of similar features. Other researchers including Uzzaman and Allen (2010) and Ha et al. (2010) made use of Markov Logic Networks (MLN). By leveraging on the transitivity properties of temporal relationships (Setzer et al., 2003), they found that MLNs are useful in inferring new temporal relationships from known ones.

Recognizing that the temporal relationships between event pairs and time expressions are related, Yoshikawa et al. (2009) proposed the use of a joint inference model and showed that improvements in performance are obtained. However this gain is attributed to the joint inference model they had developed, making use of similar surface features.

To the best of our knowledge, the only piece of work to have gone beyond sentence boundaries and tackle the problem of article-wide *E-E* temporal classification is by Do et al. (2012). Making use of integer linear programming (ILP), they built a joint inference model which is capable of classifying temporal relationships between any event pair within a given document. They also showed that event co-reference information can be useful in determining these temporal relationships. However they did not make use of features directed specifically at determining the temporal relationships of event pairs

across different sentences. Other than event coreference information, they adopted the same mix of lexico-syntactic features.

Underlying these disparate data-driven methods for similar temporal processing tasks, the reviewed works all adopted a similar set of surface features including vocabulary features, part-of-speech tags, constituent grammar parses, governing grammar nodes and verb tenses, among others. We argue that these features are not sufficiently discriminative of temporal relationships because they do not explain how sentences are combined together, and thus are unable to properly differentiate between the different temporal classifications. Supporting our argument is the work of Smith (2010), where she argued that syntax cannot fully account for the underlying semantics beneath surface text. D'Souza and Ng (2013) found out as much, and showed that adopting richer linguistic features such as lexical relations from curated dictionaries (*e.g.* Webster and WordNet) as well as discourse relations help temporal classification. They had shown that the Penn Discourse TreeBank (PDTB) style (Prasad et al., 2008) discourse relations are useful. We expand on their study to assess the utility of adopting additional discourse frameworks as alternative and complementary views.

3 Making Use of Discourse

To highlight the deficiencies of surface features, we quote here an example from Lascarides and Asher (1993):

> [A] Max opened the door. The room was pitch dark.
> [B] Max switched off the light. The room was pitch dark. (2)

The two lines of text A and B in Example 2 have similar syntactic structure. Given only syntactic features, we may be drawn to conclude that they share similar temporal relationships. However in the first line of text, the events temporally OVERLAP, while in the second line they do not. Clearly, syntax alone is not going to be useful to help us arrive at the correct temporal relations.

If existing surface features are insufficient, what is sufficient? Given a *E-E* pair which crosses sentence boundaries, how can we determine the temporal relationship between them? We take our cue from the work of Lascarides and Asher (1993). They suggested instead that discourse relations hold the key to interpreting such temporal relationships.

Building on their observations, we believe that discourse analysis is integral to any solution for the problem of article-wide *E-E* temporal classification. We thus seek to exploit a series of different discourse analysis studies, including 1) the Rhetorical Structure Theory (RST) discourse framework, 2) Penn Discourse Treebank (PDTB)-styled discourse relations based on the lexicalized Tree Adjoining Grammar for Discourse (D-LTAG), and 3) topical text segmentation, and validate their effectiveness for temporal classification.

RST Discourse Framework. RST (Mann and Thompson, 1988) is a well-studied discourse analysis framework. In RST, a piece of text is split into a sequence of non-overlapping text fragments known as elementary discourse units (EDUs). Neighboring EDUs are related to each other by a typed relation. Most RST relations are *hypotactic*, where one of the two EDUs participating in the relationship is demarcated as a *nucleus*, and the other a *satellite*. The nucleus holds more importance, from the point of view of the writer, while the satellite's purpose is to provide more information to help with the understanding of the nucleus. Some RST relations are however *paratactic*, where the two participating EDUs are both marked as nuclei. A discourse tree can be composed by viewing each EDU as a leaf node. Nodes in the discourse tree are linked to one another via the discourse relations that hold between the EDUs.

RST discourse relations capture the semantic relation between two EDUs, and these often offer a clue to the temporal relationship between events in the two EDUs too. As an example, let us refer once again to Example 2. Recall that in the second line of text "**switched off**" happens BEFORE "**dark**". The RST discourse structure for the second line of text is shown on the left of Figure 2. We see that the two sentences are related via a "*Result*" discourse relation. This fits our intuition that when there is causation, there should be a BEFORE/AFTER relationship. The RST discourse relation in this case is very useful in helping us determine the relationship between the two events.

PDTB-styled Discourse Relations. Another widely adopted discourse relation annotation is the PDTB framework (Prasad et al., 2008). Unlike the RST

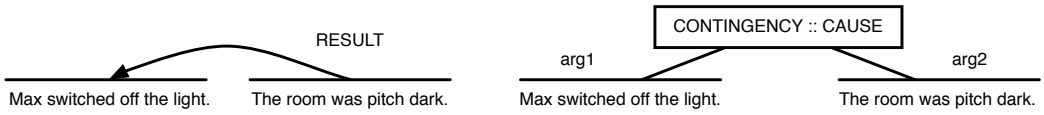

Figure 2: RST and PDTB discourse structures for the second line of text in Example 2. The structure on the left is the RST discourse structure, while the structure on the right is for PDTB.

framework, the discourse relations in PDTB build on the work on D-LTAG by Webber (2004), a lexicon-grounded approach to discourse analysis. Practically, this means that instead of starting from a pre-identified set of discourse relations, PDTB-styled annotations are more focused on detecting possible connectives (can be either explicit or implicit) within the text, before identifying the text fragments which they connect and how they are related to one another.

Applied again to the second line of text we have in Example 2, we get a structure as shown on the right side of Figure 2. From the figure we can see that the two sentences are related via a "*Cause*" relationship. Similar to what we have explained earlier for the case of RST, the presence of a causal effect here strongly hints to us that events in the two sentences share a BEFORE/AFTER relationship.

At this point we want to note the differences between the use of the RST framework and PDTB-styled discourse relations in the context of our work. The theoretical underpinnings behind these two discourse analysis are very different, and we believe that they can be complementary to each other. First, the RST framework breaks up text within an article linearly into non-overlapping EDUs. Relations can only be defined between neighboring EDUs. However this constraint is not found in PDTB-styled relations, where a text fragment can participate in one discourse relation, and a subsequence of it participate in another. PDTB relations are also not restricted only to adjacent text fragments. In this aspect, the flexibility of the PDTB relations can complement the seemingly more rigid RST framework.

Second, with PDTB-styled relations not every sentence needs to be in a relation with another as the PDTB framework does not aim to build a global discourse tree that covers all sentence pairs. This is a problem when we need to do an article-wide analysis. The RST framework does not suffer from this limitation however as we can build up a discourse tree connecting all the text within a given article.

Topical Text Segmentation. A third complementary type of inter-sentential analysis is topical text segmentation. This form of segmentation separates a piece of text into non-overlapping segments, each of which can span several sentences. Each segment represents passages or topics, and provides a coarse-grained study of the linear structure of the text (Skorochod'Ko, 1972; Hearst, 1994). The transition between segments can represent possible topic shifts which can provide useful information about temporal relationships.

Referring to Example 3[2], we have delimited the different lines of text into segments with parentheses along with a subscript. Segment (1) talks about the casualty numbers seen at a medical centre, while Segment (2) provides background information that informs us a bomb explosion had taken place. The segment boundary signals to us a possible temporal shift and can help us to infer that the bombing event took place BEFORE the deaths and injuries had occurred.

(The Davao Medical Center, a regional government hospital, recorded 19 deaths with 50 wounded. Medical evacuation workers however said the injured list was around 114, spread out at various hospitals.)$_1$
(A powerful bomb tore through a waiting shed at the Davao City international airport at about 5.15 pm (0915 GMT) while another explosion hit a bus terminal at the city.)$_2$ (3)

4 Methodology

Having motivated the use of discourse analysis for our problem, we now proceed to explain how we can make use of them for temporal classification. The different facets of discourse analysis that we are exploring in this work are structural in nature. RST

[2]From article AFP_ENG_20030304.0250 of the ACE 2005 corpus.

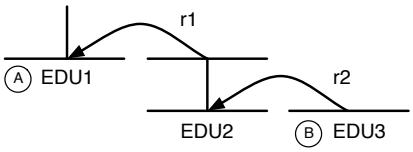

Figure 3: A possible RST discourse tree. The two circles denote two events A and B which we are interested in.

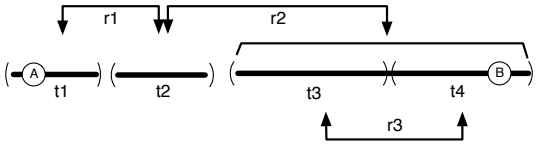

Figure 4: A possible PDTB-styled discourse annotation where the circles represent events we are interested in.

and PDTB discourse relations are commonly represented as graphs, and we can also view the output of text segmentation as a graph with individual text segments forming vertices, and the transitions between them forming edges.

Considering this, we propose the use of support vector machines (SVM), adopting a convolution kernel (Collins and Duffy, 2001) for its kernel function (Vapnik, 1999; Moschitti, 2006). The use of convolution kernels allows us to do away with the extensive feature engineering typically required to generate flat vectorized representations of features. This process is time consuming and demands specialized knowledge to achieve representations that are discriminative, yet are sufficiently generalized. Convolution kernels had also previously been shown to work well for the related problem of *E-T* temporal classification (Ng and Kan, 2012), where the features adopted are similarly structural in nature.

We now describe our use of the discourse analysis frameworks to generate appropriate representations for input to the convolution kernel.

RST Discourse Framework. Recall that the RST framework provides us with a discourse tree for an entire input article. In recent years several automatic RST discourse parsers have been made available. In our work, we first make use of the parser by Feng and Hirst (2012) to obtain a discourse tree representation of our input. To represent the meaningful portion of the resultant tree, we encode path information between the two sentences of interest.

We illustrate this procedure using the example discourse tree illustrated in Figure 3. EDUs including $EDU1$ to $EDU3$ form the vertices while discourse relations $r1$ and $r2$ between the EDUs form the edges. For a *E-E* pair, $\{A, B\}$, we can obtain a feature structure by first locating the EDUs within which A and B are found. A is found inside $EDU1$ and B is found within $EDU3$. We trace the shortest path between $EDU1$ and $EDU3$, and use this path as the feature structure for the *E-E* pair, *i.e.* $\{r1 \rightarrow r2\}$.

PDTB-styled Discourse Relations. We make use of the automatic PDTB discourse parser from Lin et al. (2013) to obtain the discourse relations over an input article. Similar to how we work with the RST discourse framework, for a given *E-E* pair, we retrieve the relevant text fragments and use the shortest path linking the two events as a feature structure for our convolution kernel classifier.

An example of a possible PDTB-styled discourse annotation is shown in Figure 4. The horizontal lines represent different sentences in an article. The parentheses delimit text fragments, $t1$ to $t4$, which have been identified as arguments participating in discourse relations, $r1$ to $r3$. For a given *E-E* pair $\{A, B\}$, we use the trace of the shortest path between them *i.e.* $\{r1 \rightarrow r2\}$ as a feature structure.

We take special care to regularize the input (as, unlike EDUs in RST, arguments to different PDTB relations may overlap, as in $r2$ and $r3$). We model each PDTB discourse annotation as a graph and employ Dijkstra's shortest path algorithm. The graph resulting from the annotation in Figure 4 is given in Figure 5. Each text fragment t_i maps to a vertex n_i in the graph. PDTB relations between text fragments form edges between corresponding vertices. As $r2$ relates $t2$ to both $t3$ and $t4$, two edges link up $n2$ to the corresponding vertices $n3$ and $n4$ respectively. By doing this, Dijkstra's algorithm will always allow us to find the desired shortest path.

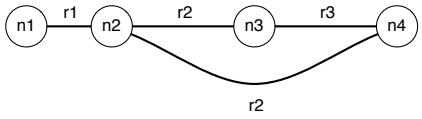

Figure 5: Graph derived from discourse annotation in Figure 4.

Topical Text Segmentation. Taking as input a complete text article, we make use of the state-of-the-art text segmentation system from Kazantseva and Szpakowicz (2011). The output of the system is a series of non-overlapping, linear text segments, which we can number sequentially.

In Figure 6 the horizontal lines represent sentences. Parentheses with subscripts mark out the segment boundaries. We can see two segments $s1$ and $s2$ here. Given a target *E-E* pair $\{A, B\}$ (represented as circles inside the figure), we identify the segment number of the corresponding segment in which each of A and B is found. We build a feature structure with the identified segment numbers, *i.e.* $\{s1 \rightarrow s2\}$ to capture the segmentation.

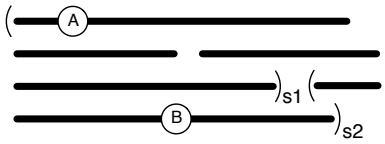

Figure 6: A possible segmentation of three sentences into two segments.

5 Results

We conduct a series of experiments to validate the utility of our proposed features.

Data Set. We make use of the same data set built by Do et al. (2012). The data set consists of 20 newswire articles which originate from the ACE 2005 corpus (ACE, 2005). Initially, the data set consist of 324 event mentions, and a total of 375 annotated *E-E* pairs. We perform the same temporal saturation step as described in Do et al. (2012), and obtained a total of 7,994 *E-E* pairs[3].

A breakdown of the number of instances by each temporal classes is shown in Table 1. Unlike earlier data sets such as that for TempEval-2 where more than half (about 55%) of test instances belong to the OVERLAP class, OVERLAP instances make up just 10% of the data set.

This difference is due mainly to the fact that our data set consists not only of intra-sentence *E-E* pairs, but also of article-wide *E-E* pairs. Figure 7 shows the number of instances for each temporal class broken down by the number of sentences (*i.e.* sentence gap) that separate the events within each *E-E* pair. We see that as the sentence gap increases, the proportion of OVERLAP instances decreases. The intuitive explanation for this is that when event mentions are very far apart in an article, it becomes more unlikely that they happen within the same time span.

Class	AFTER	BEFORE	OVERLAP
# *E-E* pairs	3,588 (45%)	3,589 (45%)	815 (10%)

Table 1: Number of *E-E* pairs in data set attributable to each temporal class. Percentages shown in parentheses.

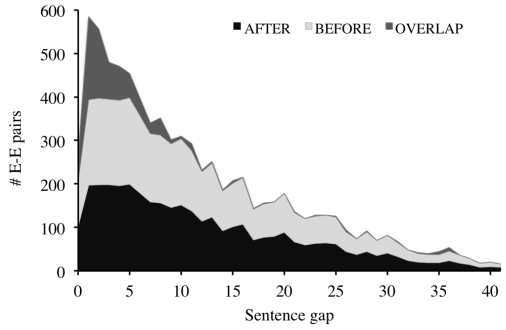

Figure 7: Breakdown of number of *E-E* pairs for each temporal class based on sentence gap.

Experiments. The work done in Do et al. (2012) is highly related to our experiments, and so we have reported the relevant results for local *E-E* classification in Row 1 of Table 2 as a reference. While largely comparable, note that a direct comparison is not possible because 1) the number of *E-E* instances we have is slightly different from what was reported, and 2) we do not have access to the exact partitions they have created for 5-fold cross-validation.

As such, we have implemented a baseline adopting similar surface lexico-syntactic features used in previous work (Mani et al., 2006; Bethard and Martin, 2007; Ng and Kan, 2012; Do et al., 2012), including 1) part-of-speech tags, 2) tenses, 3) dependency parses, 4) relative position of events in article,

[3]Though we have obtained the data set from the original authors, there was a discrepancy in the number of *E-E* pairs. The original paper reported a total of 376 annotated *E-E* pairs. Besides this, we also repeated the saturation steps iteratively until no new relationship pairs are generated. We believe this to be an enhancement as it ensures that all inferred temporal relationships are generated.

	System	Precision	Recall	F_1
(1)	DO2012	43.86	52.65	47.46
(2)	BASE	59.55	38.14	46.50
(3)	BASE + RST + PDTB + TOPICSEG	71.89	41.99	53.01
(4)	BASE + RST + PDTB + TOPICSEG + COREF	75.23	43.58	55.19
(5)	BASE + O-RST + PDTB + O-TOPICSEG + O-COREF	78.35	54.24	64.10

Table 2: Macro-averaged results obtained from our experiments. The difference in F_1 scores between each successive row is statistically significant, but a comparison is not possible between rows (1) and (2).

5) the number of sentences between the target events and 6) VerbOcean (Chklovski and Pantel, 2004) relations between events. This baseline system, and the subsequent systems we will describe, comprises of three separate one-vs-all classifiers for each of the temporal classes. The result obtained by our baseline is shown in Row 2 (*i.e.* BASE) in Table 2. We note that our baseline is competitive and performs similarly to the results obtained by Do et al. (2012). However as we do not have the raw judgements from Do's system, we cannot test for statistical significance.

We also implemented our proposed features and show the results obtained in the remaining rows of Table 2. In Row 3, RST denotes the RST discourse feature, PDTB denotes the PDTB-styled discourse features, and TOPICSEG denotes the text segmentation feature. Compared to our own baseline, there is a relative increase of 14% in F_1, which is statistically significant when verified with the one-tailed Student's paired *t*-test ($p < 0.01$).

In addition, Do et al. (2012) have shown the value of event co-reference. Therefore we have also included this feature by making use of an automatic event co-reference system by Chen et al. (2011). The result obtained after adding this feature (denoted by COREF) is shown in Row 4. The relative increase in F_1 of about 4% from Row 3 is statistically significant ($p < 0.01$) and affirms that event co-reference is a useful feature to have, together with our proposed features. We note that our complete system in Row 4 gives a 16% improvement in F_1, relative to the reference system DO2012 in Row 1.

To get a better idea of the performance we can obtain if oracular versions of our features are available, we also show the results obtained if hand-annotated RST discourse structures, text segments, as well as event co-reference information were used. Annotations for the RST discourse structures and text segments were performed by the first author (RST annotations were made following the annotation guidelines given by Carlson and Marcu (2001)). Oracular event co-reference information was included in the dataset that we have used.

In Row 5 the prefix O denotes oracular versions of the features we had proposed. From the results we see that there is a marked increase of over 15% in F_1 relative to Row 4. Compared to Do's state-of-the-art system, there is also a relative gain of at least 35%. These oracular results further confirm the importance of non-local discourse analysis for temporal processing.

6 Discussion

Ablation tests. We performed ablation tests to assess the efficacy of the discourse features used in our earlier experiments. Starting from the full system, we dropped each discourse feature in turn to see the effect this has on overall system performance. Our test is performed over the same data set, again with 5-fold cross-validation. The results in Table 3 show a statistically significant (based on the one-tailed Student's paired *t*-test) drop in F_1 in each case, which proves that each of our proposed features is useful and required.

From the ablation tests, we also observe that the RST discourse feature contributes the most to overall system performance while the PDTB discourse feature contributes the least. However we should not conclude prematurely that the former is more useful than the latter; as the results are obtained using parses from automatic systems, and are not reflective of the full utility of ground truth discourse annotations.

Useful Relations. The ablation test results showed us that discourse relations (in particular RST dis-

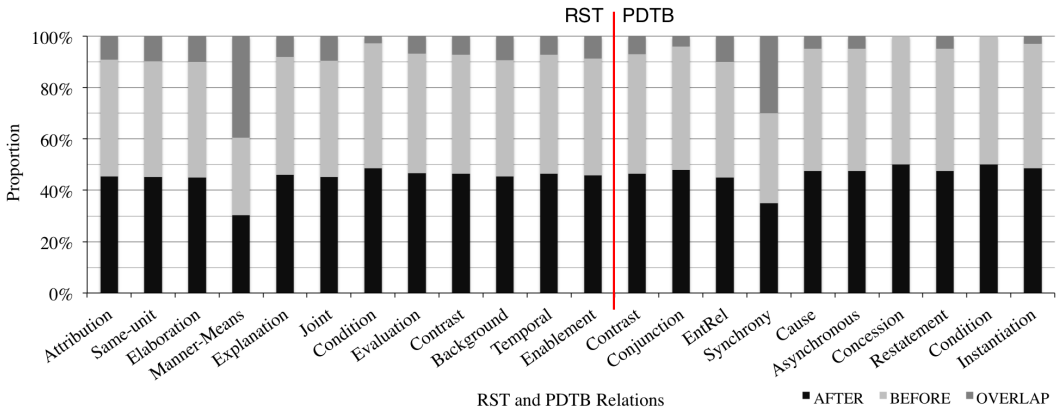

Figure 8: Proportion of occurence in temporal classes for every RST and PDTB relation.

Ablated Feature	Change in F_1	Sig
−RST	-9.03	**
−TOPICSEG	-2.98	**
−COREF	-2.18	**
−PDTB	-1.42	*

Table 3: Ablation test results. '**' and '*' denote statistically significance against the full system with $p < 0.01$ and $p < 0.05$, respectively.

course relations) are the most important in our system. We have also motivated our work earlier with the intuition that certain relations such as the RST "*Result*" and the PDTB "*Cause*" relations provide very useful temporal cues. We now offer an introspection into the use of these discourse relations.

Figure 8 illustrates the relative proportion of temporal classes in which each RST and PDTB relation appear. If the relations are randomly distributed, we should expect their distribution to follow that of the temporal classes as shown in Table 1. However we see that many of the relations do not follow this distribution. For example, we observe that several relations such as the RST "*Condition*" and PDTB "*Cause*" relations are almost exclusively found within AFTER and BEFORE event pairs only, while the RST "*Manner-means*" and PDTB "*Synchrony*" relations occur in a disproportionately large number of OVERLAP event pairs. These relations are likely useful in disambiguating between the different temporal classes.

To verify this, we examine the convolution tree fragments that lie on the support vector of our SVM classifier. The work of Pighin and Moschitti (2010) in linearizing kernel functions allows us to take a look at these tree fragments. Applying the linearization process leads to a different classifier from the one we have used. The identified tree fragments are therefore just an approximation to those actually employed by our classifier. However, this analysis still offers an introspection as to what relations are most influential for classification.

BEFORE		OVERLAP	
B1	(Temporal ...	O1	(Manner-means ...
B2	(Temporal (Elaboration ...		
B3	(Condition (Explanation ...		
B4	(Condition (Attribution ...		
B5	(Elaboration (Bckgrnd ...		

Table 4: Subset of top RST discourse fragments on support vectors identified by linearizing kernel function.

Table 4 shows a subset of the top RST discourse fragments identified for the BEFORE and OVERLAP one-vs-all classifiers. The list is in line with what we expect from Figure 8. The former consists of fragments containing relations such as "*Temporal*" and "*Condition*", while the latter has a sole fragment containing "*Manner-Means*".

To illustrate what these fragments may mean, we show several example sentences from our data set in Example 4. Sentence A consists of the tree fragment B1, *i.e.* "*(Temporal...*". Its corresponding discourse structure is illustrated in the top half of Figure 9. This fragment indicates to us (correctly) that the event "**wielded**" happened BEFORE Milosevic was "**swept out**" of power. Sentence B is made up of tree fragment O1, *i.e.* "*(Manner-means...*",

and its discourse structure is shown in the bottom half of Figure 9. As with the previous example, the fragment suggests (correctly) that there should be a OVERLAP relationship for the "**requested** – **said**" event pair.

> [A] Milosevic and his wife **wielded** enormous power in Yugoslavia for more than a decade before he was **swept out** of power after a popular revolt in October 2000.
>
> [B] The court order was **requested** by Jack Welch's attorney, Daniel K. Webb, who **said** Welch would likely be asked about his business dealings, his health and entries in his personal diary. (4)

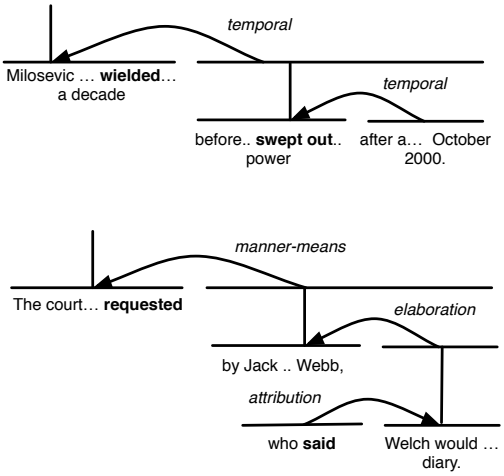

Figure 9: RST discourse structures for sentences A (top half) and B (bottom half) in Example 4.

Segment Numbers. From the ablation test results, text segmentation is the next most important feature after the RST discourse feature. This is interesting given that the defined feature structure for topical text segmentation is not the most intuitive. By using actual segment numbers, the structure may not generalize well for articles of different lengths for example, as each article may have vastly different number of segments. The transition across segments may also not carry the same semantic significance for different articles.

Our experiments have however shown that this feature design is useful in improving performance. This is likely because:

1. The default settings of the text segmentation system we had used are such that precision is favoured over recall (Kazantseva and Szpakowicz, 2011, p. 292). As such there is just an average of between two to three identified segments per article. This makes the feature more generalizable despite making use of actual segment numbers.

2. The style of writing in newswire articles which we are experimenting on generally follows common journalistic guidelines. The semantics behind the transitions across the coarse-grained segments that were identified are thus likely to be of a similar nature across many different articles.

We leave for future work an investigation into whether more fine-grained topic segments can lead to further performance gains. In particular, it will be interesting to study if work on argumentative zoning (Teufel and Kan, 2011) can be applied to newswire articles, and whether the subsequent learnt document structures can be used to delineate topic segments more accurately.

Error Analysis. Besides examining the features we had used, we also want to get a better idea of the errors made by our classifier. Recall that we are using separate one-vs-all classifiers for each of the temporal classes, so each of the three classifiers generates a column in the aggregate confusion matrix shown in Table 5. In cases where none of the SVM classifiers return a positive confidence value, we do not assign a temporal class (captured as column **N**). The high number of event pairs which are not assigned to any temporal class explains the lower recall scores obtained by our system, as observed in Table 2.

	Predicted			
	O	**B**	**A**	**N**
O	119 (14.7%)	114 (14.1%)	104 (12.8%)	474 (58.5%)
B	19 (0.5%)	2067 (57.9%)	554 (15.5%)	928 (26.0%)
A	16 (0.5%)	559 (15.7%)	2046 (57.3%)	947 (26.5%)

Table 5: Confusion matrix obtained for the full system, classifying into (**O**)VERLAP, (**B**)EFORE, (**A**)FTER, and (**N**)o result.

Additionally, an interesting observation is the low percentage of OVERLAP instances that our classifier managed to predict correctly. About 57% of BEFORE and AFTER instances are classified cor-

rectly, however only about 15% of OVERLAP instances are correct.

Figure 10 offers more evidence to suggest that our classifier works better for the BEFORE and AFTER classes than the OVERLAP class. We see that as sentence gap increases, we achieve a fairly consistent performance for both BEFORE and AFTER instances. OVERLAP instances are concentrated where the sentence gap is less than 7, with the best accuracy figure coming in below 30%.

Although not definitive, this may be because our data set consists of much fewer OVERLAP instances than the other two classes. This bias may have led to insufficient training data for accurate OVERLAP classification. It will be useful to investigate if using a more balanced data set for training can help overcome this problem.

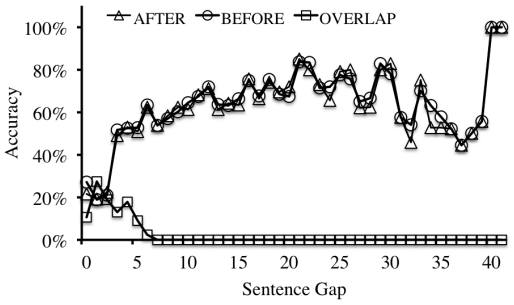

Figure 10: Accuracy of the classifer for each temporal class, plotted against the sentence gap of each *E-E* pair.

7 Conclusion

We believe that discourse features play an important role in the temporal ordering of events in text. We have proposed the use of different discourse analysis frameworks and shown that they are effective for classifying the temporal relationships of article-wide *E-E* pairs. Our proposed discourse-based features are robust and work well even though automatic discourse analysis is noisy. Experiments further show that improvements to these underlying discourse analysis systems will benefit system performance.

In future work, we will like to explore how to better exploit the various discourse analysis frameworks for temporal classification. For instance, RST relations are either *hypotactic* or *paratactic*. Marcu (1997) made use of this to generate automatic summaries by considering EDUs which are nuclei to be more salient. We believe it is interesting to examine how such information can help. We are also interested to apply discourse features in the context of a global inferencing system (Yoshikawa et al., 2009; Do et al., 2012), as we think such analyses will also benefit these systems as well.

Acknowledgments

We like to express our gratitude to Quang Xuan Do, Wei Lu, and Dan Roth for generously making available the data set they have used for their work in EMNLP 2012. We would also like to thank the anonymous reviewers who reviewed this paper for their valuable feedback.

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

References

ACE. 2005. The ACE 2005 (ACE05) Evaluation Plan. October.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research (JAIR)*, 17:35–55.

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 (SemEval)*, pages 129–132, June.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging manual. Technical Report ISI-TR-545, Information Sciences Institute, University of Southern California, July.

Bin Chen, Jian Su, Sinno Jialin Pan, and Chew Lim Tan. 2011. A Unified Event Coreference Resolution by Integrating Multiple Resolvers. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 102–110, November.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33–40, July.

Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. In *Proceedings of NIPS*.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pages 677–689, July.

Jennifer D'Souza and Vincent Ng. 2013. Classifying Temporal Relations with Rich Linguistic Knowledge. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 918–927, June.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistics Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*, pages 60–68, July.

Eun Young Ha, Alok Baikadi, Carlyle Licata, and James C. Lester. 2010. NCSU: Modeling Temporal Relations with Markov Logic and Lexical Ontology. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 341–344, July.

Marti A. Hearst. 1994. Multi-Paragraph Segmentation of Expository Text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9–16, June.

Anna Kazantseva and Stan Szpakowicz. 2011. Linear Text Segmentation Using Affinity Propagation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 284–293, July.

Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. 2010. JU_CSE_TEMP: A First Step Towards Evaluating Events, Time Expressions and Temporal Relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 345–350, July.

Alex Lascarides and Nicholas Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense Entailment. *Linguistics and Philosophy*, 16(5):437–493.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2013. A PDTB-styled End-to-End Discourse Parser. *Natural Language Engineering*, FirstView:1–34, February.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine Learning of Temporal Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 753–760, July.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. From Discourse Structures to Text Summaries. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 97, pages 82–88, July.

Alessandro Moschitti. 2006. Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In *Proceedings of the 17th European Conference on Machine Learning (ECML)*, September.

Jun-Ping Ng and Min-Yen Kan. 2012. Improved Temporal Relation Classification using Dependency Parses and Selective Crowdsourced Annotations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 2109–2124, December.

Daniele Pighin and Alessandro Moschitti. 2010. On Reverse Feature Engineering of Syntactic Tree Kernels. In *Proceedings of the 14th Conference on Natural Language Learning (CoNLL)*, August.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, May.

Andrea Setzer, Robert Gaizauskas, and Mark Hepple. 2003. Using Semantic Inferences for Temporal Annotation Comparison. In *Proceedings of the 4th International Workshop on Inference in Computational Semantics (ICoS)*, September.

Eduard F. Skorochod'Ko. 1972. Adaptive Method of Automatic Abstracting and Indexing. In *Proceedings of the IFIP Congress*, pages 1179–1182.

Carlota S. Smith. 2010. Temporal Structures in Discourse. *Text, Time, and Context*, 87:285–302.

Simone Teufel and Min-Yen Kan. 2011. Robust Argumentative Zoning for Sensemaking in Scholarly Documents. In *Advanced Language Technologies for Digital Libraries*, pages 154–170. Springer.

Naushad Uzzaman and James F. Allen. 2010. TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, pages 276–283, July.

Naushad Uzzaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *Computing Research Repository (CoRR)*, abs/1206.5333.

Vladimir N. Vapnik, 1999. *The Nature of Statistical Learning Theory*, chapter 5. Springer.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval Challenge: Identifying

Temporal Relations in Text. *Language Resources and Evaluation*, 43(2):161–179.

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 (SemEval)*, pages 57–62, July.

Marc Verhagen. 2005. Temporal Closure in an Annotation Environment. *Language Resources and Evaluation*, 39(2-3):211–241.

Bonnie Webber. 2004. D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science*, 28(5):751–779.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly Identifying Temporal Relations with Markov Logic. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL) and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (AFNLP)*, pages 405–413, August.