Inferring Activity Time in News through Event Modeling

Vladimir Eidelman
Department of Computer Science
Columbia University
New York, NY 10027
vae2101@columbia.edu

Abstract

Many applications in NLP, such as question-answering and summarization, either require or would greatly benefit from the knowledge of when an event occurred. Creating an effective algorithm for identifying the activity time of an event in news is difficult in part because of the sparsity of explicit temporal expressions. This paper describes a domain-independent machine-learning based approach to assign activity times to events in news. We demonstrate that by applying topic models to text, we are able to cluster sentences that describe the same event, and utilize the temporal information within these event clusters to infer activity times for all sentences. Experimental evidence suggests that this is a promising approach, given evaluations performed on three distinct news article sets against the baseline of assigning the publication date. Our approach achieves 90%, 88.7%, and 68.7% accuracy, respectively, outperforming the baseline twice.

1 Introduction

Many practical applications in NLP either require or would greatly benefit from the use of temporal information. For instance, question-answering and summarization systems demand accurate processing of temporal information in order to be useful for answering 'when' questions and creating coherent summaries by temporally ordering information. Proper processing is especially relevant in news, where multiple disparate events may be described within one news article, and it is necessary to identify the separate timepoints of each event.

Event descriptions may be confined to one sentence, which we establish as our text unit, or be spread over many, thus forcing us to assign all sentences an activity time. However, only 20%-30% of sentences contain an explicit temporal expression, thus leaving the vast majority of sentences without temporal information. A similar proportion is reported in Mani et al. (2003), with only 25% of clauses containing explicit temporal expressions. The sparsity of these expressions poses a real challenge. Therefore, a method for efficiently and accurately utilizing temporal expressions to infer activity times for the remaining 70%-80% of sentences with no temporal information is necessary.

This paper proposes a domain-independent machine-learning based approach to assign activity times to events in news without deferring to the publication date. Posing the problem in an information retrieval framework, we model events by applying topic models to news, providing a way to automatically distribute temporal information to all sentences. The result is prototype system which achieves promising results.

In the following section, we discuss related work in temporal information processing. Next we motivate the use of topic models for our task, and present our methods for distributing temporal information. We conclude by presenting and discussing our results.

2 Related Work

Mani and Wilson (2000) worked on news and introduced an annotation scheme for temporal expressions, and a method for using explicit tempo-

Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 13–18,
Columbus, June 2008.

Sentence Order	Event	Temporal Expression
1	Event X	None
2	Event Y	January 10, 2007
3	Event X	None
4	Event X	November 16, 1967
5	Event Y	None
6	Event Y	January 10, 2007
7	Event X	None

Table 1: Problematic Example

ral expressions to assign activity times to the entirety of an article. Their preliminary work on inferring activity times suggested a baseline method which spread time values of temporal expressions to neighboring events based on proximity. Filatova and Hovy (2001) also process explicit temporal expressions within a text and apply this information throughout the whole article, assigning activity times to all clauses.

More recent work has tried to temporally anchor and order events in news by looking at clauses (Mani et al., 2003). Due to the sparsity of temporal expressions, they computed a reference time for each clause. The reference time is inferred using a number of linguistic features if no explicit reference is present, but the algorithm defaults to assigning the most recent time when all else fails.

A severe limitation of previous work is the dependence on article structure. Mani and Wilson (2000) attribute over half the errors of their baseline method to propagation of an incorrect event time to neighboring events. Filatova and Hovy (2001) infer time values based on the most recently assigned date or the date of the article. The previous approaches will all perform unfavorably in the example presented in Table 1, where a second historical event is referred to between references to a current event. This kind of example is quite common.

3 Modeling News

To address the aforementioned issues of sparsity while relieving dependence on article structure, we treat event discovery as a clustering problem. Clustering methods have previously been used for event identification (Hatzivassiloglou et al., 2000; Siddharthan et al., 2004). After a topic model of news text is created, sentences are clustered into topics - where each topic represents a specific event. This allows us to utilize all available temporal information in each cluster to distribute to all the sentences within that cluster, thus allowing for assigning of activity times to sentences without explicit temporal expressions. Our key assumption is that similar sentences describe the same event.

Our approach is based on information retrieval techniques, so we subsequently use the standard language of text collections. We may refer to sentences, or clusters of sentences created from a topic model as 'documents', and a collection of sentences, or collection of clusters of sentences from one or more news articles as a 'corpus'. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative model for describing collections of text corpora, which represents each document as a mixture over a set of topics, where each topic has associated with it a distribution over words. Topics are shared by all documents in the corpus, but the topic distribution is assumed to come from a Dirichlet distribution. LDA allows documents to be composed of multiple topics with varying proportions, thus capturing multiple latent patterns.

Depending on the words present in each document, we associate it with one of N topics, where N is the number of latent topics in the model. We assign each document to the topic which has the highest probability of having generated that document. We expect document similarity in a cluster to be fairly high, as evidenced by document modeling performance in Blei et al. (2003). Since each cluster is a collection of similar documents, with our assumption that similar documents describe the same event, we conclude that each cluster represents a specific event. Thus, if at least one sentence in an event cluster contains an explicit temporal expression, we can distribute that activity time to other sentences in the cluster using an inference algorithm we explain in the next section. More than one event cluster may represent the same event, as in Table 3, where both topics describe a different perspective on the same event: the administrative reaction to the incident at Duke.

Creating a cluster of similar documents which represent an event can be powerful. First, we are no longer restricted by article structure. To refer back to

Table 1, our approach will assign the correct activity time for all event X sentences, even though they are separated in the article and only one contains an explicit temporal expression, by utilizing an event cluster which contains the four sentences describing event X to distribute the temporal information[1].

Second, we are not restricted to using only one article to assign activity times to sentences. In fact, one of the major strengths of this approach is the ability to take a collection of articles and treat them all as one corpus, allowing the model to use all explicit temporal expressions on event X present throughout all of the articles to distribute activity times. This is especially helpful in multidocument summarization, where we have multiple articles on the same event.

Additionally, using LDA as a method for event identification may be advantageous over other clustering methods. For one, Siddharthan et al. (2004) reported that removing relative clauses and appositives, which provide background or discourse related information, improves clustering. LDA allows us to discover the presence of multiple events within a sentence, and future work will focus on exploiting this to improve clustering.

3.1 Corpus

We obtained 22 news articles, which can be divided into three distinct sets: Duke Rape Case (DR), Terrorist Bombings in Mumbai (MB), Israeli-Lebanese conflict (IC) (Table 2). All articles come from English Newswire text, and each sentence was manually annotated with an activity time by people outside of the project. The Mumbai Bombing articles all occur within a several day span, as do the Israeli-Conflict articles. The Duke Rape case articles are an exception, since they are comprised of multiple events which happened over the course of several months: Thus these articles contain many cases such as *"The report said...on March 14..."*, where the report is actually in May, yet speaks of events in March. For the purposes of this experiment we took the union of the possible dates mentioned in a sentence as acceptable activity times, thus both the report statement date and the date mentioned in the report are correct activity times for the sentence. Future work will investigate whether we can discriminate between these two dates.

[1] Analogously, our approach will assign correct activity time to all event Y sentences

Article Set	# of Articles	# of Sentences
Duke Rape Case	5	151
Mumbai Bombing	8	284
Israeli Conflict	9	300

Table 2: Article and Sentence distribution

Our approach relies on prior automatic linguistic processing of the articles by the Proteus system (Grishman et al., 2005). The articles are annotated with time expression tags, which assign values to both absolute *"July 16, 2006"* and relative *"now"* temporal expressions. Although present, our approach does not currently use activity time ranges, such as "past 2 weeks" or "recent days". The articles are also given entity identification tags, which assigns a unique intra-article id to entities of the types specified in the ACE 2005 evaluation. For example, both "they" - an anaphoric reference - and "police officers" are recognized as referring to the same real-world entity.

3.2 Feature Extraction

From this point on unless otherwise noted, reference to news articles indicates one of the three sets of news articles, not the complete set. We begin by breaking news articles into their constituent sentences, which are our 'documents', the collection of them being our 'corpus', and indexing the documents.

We use the bag-of-words assumption to represent each document as an unordered collection of words. This allows the representation of each document as a word vector. Additionally, we add any entity identification information and explicit temporal expressions present in the document to the feature vector representation of each document.

3.3 Intra-Article Event Representation

To represent events within one news article, we construct a topic model for each article separately. The Intra-Article (IAA) model constructed for an article allows us to group sentences within that article together according to event. This allows the formation of new 'documents', which consist not of single

The administrators did not know of the racial dimension until March 24, the report said. The report did say that Brodhead was hampered by the administration's lack of diversity. He said administrators would be reviewed on their performance on the normal schedule and he had no immediate plans to make personnel changes. Administrators allowed the team to keep practicing; Athletics Director Joe Alleva called the players "wonderful young men."
Yet even Duke faculty members, many of them from the '60s and '70s generations that pushed college administrators to ease their controlling ways, now are urging the university to require greater social as well as scholastic discipline from students. Duke professors, in fact, are offering to help draft new behavior codes for the school. With years of experience and academic success to their credit, faculty members ought to be listened to. For the moment, five study committees appointed by Brodhead seem to mean business, which is encouraging.

Table 3: Two topics representing a different perspective on the same event

sentences, but a cluster of sentences representing an event. Accordingly, we combine the feature vector representations of the single sentences in an event cluster into one feature vector, forming an aggregate of all their features. Although at this stage we have everything we need to infer activity times, our approach allows incorporating information from multiple articles.

3.4 Inter-Article Event Representation

To represent events over multiple articles, we suggest two methods for Inter-Article (IRA) topic modeling. The first, IRA.1, is to combine the articles and treat them as one large article. This allows processing as described in IAA, with the exception that event clusters may contain sentences from multiple articles. The second, IRA.2, builds on IAA models of single articles and uses them to construct an IRA model. The IRA.2 model is constructed over a corpus of documents containing event clusters, allowing a grouping of event clusters from multiple articles. Event clusters may now be composed of sentences describing the same event from multiple articles, thus increasing our pool of explicit temporal expressions available for inference.

3.5 Activity Time Assignment

To accurately infer activity times of all sentences, it is crucial to properly utilize the available temporal expressions in the event clusters formed in the IRA or IAA models. Our proposed inference algorithm is a starting point for further work. We use the most frequent activity time present in an event cluster as the value to assign all the sentences in that event cluster. In phase one of the algorithm we process each event cluster separately. If the majority of sentences with temporal expressions have the same activity time, then this activity time is distributed to the other sentences. If there is a tie between the number of occurrences of two activity times, both these times are distributed as the activity time to the other sentences. If there is no majority time and no tie in the event cluster, then each of the sentences with a temporal expression retains its activity time, but no information is distributed to the other sentences. Phase two of the inference algorithm reassembles the sentences back into their original articles, with most sentences now having activity times tags assigned from phase one. Sentences that remain unmarked, indicating that they were in event clusters with no majority and no tie, are assigned the majority activity time appearing in their reassembled article.

4 Empirical Evaluation

In evaluating our approach, we wanted to compare different methods of modeling events prior to performing inference.

- Method (1) IAA then IRA.2 - Creating IAA models with 20 topics for each news article, and IRA.2 models for each of the three sets of IAA models with 20, 50, and 100 topic.
- Method (2) IAA only - Creating an IAA model with 20 topics for each article
- Method (3) IRA.1 only - Creating IRA.1 model with 20 and 50 topics for each of the three sets of articles.

4.1 Results

Table 4 presents results for the three sets of articles on the six different experiments performed. Since our approach assigns activity times to all sentences, overall accuracy is measured as the total number of correct activity time assignments made out of the total number of sentences. The baseline accuracy is computed by assigning each sentence the article publication date, and because news generally describes current events, this achieves remarkably high performance.

The overall accuracy measures performance of the complete inference algorithm, while the rest of the metrics measure the performance of phase one only, where we process each event cluster separately. Assessing the performance of phase one allows us to indirectly evaluate the event clusters which we create using LDA. M1 accuracy represents the number of sentences that were assigned the correct activity time in phase one out of the total number of activity time inferences made in phase one. Thus, this does not take into account any assignments made by phase two, and allows us to examine our assumptions about event representation expressed earlier. A large denominator in M1 indicates that many sentences were assigned in phase one, while a low one indicates the presence of event clusters which were unable to distribute temporal information.

M2 looks at how well the algorithm performs on the difficult cases where the activity time is not the same as the publication date. M3 looks at how well the algorithm performs on the majority of sentences which have no temporal expressions.

For the IC and DR sets, results show that Method (1), where IAA is performed prior to IRA.2 achieves the best performance, with accuracy of 88.7% and 90%, respectively, giving credence to the claim that representing events within an article before combining multiple articles improves inference.

The MB set somewhat counteracts this claim, as the best performance was achieved by Method (3), where IRA.1 is performed. This may be due to the fact that MB differs from DR and IC sets in that it contains several regurgitated news articles. Regurgitated news articles are comprised almost entirely of statements made at a previous time in other news articles. Method (3) combines similar sentences from all the articles right away, placing sentences from regurgitated articles in an event cluster with the original sentences. This allows our approach to outperform the baseline system by 4.3%, with and accuracy of 68.7%.

5 Discussion

There are limitations to our approach which need to be addressed. Foremost, evidence suggests that event clusters are not perfect, as error analysis has shown event clusters which represent two or more

Set	Setup	Accur.	M1	M2	M3
DR	Base	**135/151**			
		89.4%			
DR	(1) 20	121/151	55/83	5/12	27/43
		80.1%	66.2%	41.6%	62.7%
DR	(1) 50	**136/151**	91/105	4/13	60/66
		90.0%	86.6%	30.7%	90.9%
DR	(1)100	128/151	87/109	4/13	58/70
		84.7%	79.8%	30.7%	82.8%
DR	(2) 20	106/151	45/68	4/11	20/33
		70.2%	66.2%	36.4%	60.6%
DR	(3) 20	111/151	82/110	8/14	49/71
		73.5%	74.7%	57.1%	69.0%
DR	(3) 50	99/151	92/135	6/14	63/95
		65.5%	68.1%	42.9%	66.3%
Set	Setup	Accur.	M1	M2	M3
MB	Base	**183/284**			
		64.4%			
MB	(1) 20	166/284	116/187	41/68	60/104
		58.5%	62.0%	60.2%	57.7%
MB	(1) 50	152/284	121/206	41/72	66/120
		53.5%	58.7%	56.9%	55.0%
MB	(1)100	139/284	112/204	41/81	60/124
		48.9%	54.9%	50.6%	48.4%
MB	(2) 20	143/284	103/161	40/63	49/85
		50.3%	63.9%	63.5%	57.3%
MB	(3) 20	146/284	99/160	45/64	47/81
		51.4%	61.9%	70.3%	58.0%
MB	(3) 50	**195/284**	123/184	32/67	74/103
		68.7%	66.8%	47.8%	71.8%
Set	Setup	Accur.	M1	M2	M3
IC	Base	**272/300**			
		90.7%			
IC	(1) 20	250/300	158/205	12/22	118/151
		83.3%	77.1%	54.5%	78.1%
IC	(1) 50	263/300	168/192	12/19	127/139
		87.7%	87.5%	63.2%	91.4%
IC	(1)100	**266/300**	173/202	11/20	130/149
		88.7%	85.6%	55.0%	87.2%
IC	(2) 20	250/300	156/181	11/18	117/130
		83.3%	86.2%	61.1%	90.0%
IC	(3) 20	225/300	112/145	14/21	75/95
		75.0%	77.2%	66.7%	78.9%
IC	(3) 50	134/300	115/262	14/25	76/206
		44.7%	43.9%	56.0%	36.9%

Table 4: Results : Sentence Breakdown

events. Event clusters which contain sentences describing several events pose a real challenge, as they are primarily responsible for inhibiting performance. This limitation is not endemic to our approach for event discovery, as Xu et al. (2006) stated that event extraction is still considered as one of the most challenging tasks, because an event mention can be expressed by several sentences and different linguistic expressions.

One of the major strengths of our approach is the ability to combine all temporal information on an event from multiple articles. However, due the imperfect event clusters, combining temporal information from different articles within an event cluster has not yet yielded satisfactory results.

Although sentences from the same article in IRA event clusters usually represent the same event, other sentences from different articles may not. We modified the inference algorithm to reflect this, and only consider sentences from the same news article when distributing temporal information, even though sentences from other articles may be present in the event cluster. Therefore, further work to construct event clusters which more closely represent events is expected to yield improvements in performance. Future work will explore a richer feature set, including such features as cross-document entity identification information, linguistic features, and outside semantic knowledge to increase robustness of the feature vectors. Finally, the optimal model parameters are currently selected by an oracle, however, we hope to further evaluate our approach on a larger dataset in order to determine how to automatically select the optimal parameters.

6 Conclusion

This paper presented a novel approach for inferring activity times for all sentences in a text. We demonstrate we can produce reasonable event representations in an unsupervised fashion using LDA, posing event discovery as a clustering problem, and that event clusters can further be used to distribute temporal information to the sentences which lack explicit temporal expressions. Our approach achieves 90%, 88.7%, and 68.7% accuracy, outperforming the baseline set forth in two cases. Although differences prevent a direct comparison, Mani and Wilson (2000) achieved an accuracy of 59.4% on 694 verb occurrences using their baseline method, Filatova and Hovy (2001) achieved 82% accuracy on time-stamping clauses for a single type of event on 172 clauses, and Mani et al. (2003) achieved 59% accuracy in their algorithm for computing a reference time for 2069 clauses. Future work will improve upon the majority criteria used in the inference algorithm, on creating more accurate event representations, and on determining optimal model parameters automatically.

Acknowledgements

We wish to thank Kathleen McKeown and Barry Schiffman for invaluable discussions and comments.

References

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp.993–1022

Elena Filatova and Eduard Hovy. 2001. Assigning Time-Stamps to Event-Clauses. *Workshop on Temporal and Spatial Information Processing, ACL'2001* 88-95.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. NYU's English ACE 2005 system description. *In ACE 05 Evaluation Workshop*.

Vasileios Hatzivassiloglou, Luis Gravano, and Ankineedu Maganti. 2000. An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering. *In Proceedings of the 23rd ACM SIGIR*, pages 224-231.

Inderjeet Mani, Barry Schiffman and Jianping Zhang. 2003. Inferring Temporal Ordering of Events in News. *Proceedings of the Human Language Technology Conference*.

Inderjeet Mani and George Wilson. 2000. Robust Temporal Processing of News. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 69-76. Hong Kong.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. *In 20th International Conference on Computational Linguistics* .

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2006. Automatic event and relation detection with seeds of varying complexity. *In Proceedings of the AAAI Workshop Event Extraction and Synthesis*, pages 1217, Boston.