Contents

Marking Atomic Events in Sets of Related Texts

ELENA FILATOVA & VASILEIOS HATZIVASSILOGLOU
Columbia University

Abstract

The notion of an *event* has been widely used in the computational linguistics literature as well as in information retrieval and various NLP applications, although with significant variance in what exactly an event is. We describe an empirical study aimed at developing an operational definition of an event at the *atomic* (sentence or predicate) level, and use our observations to create a system for detecting and prioritizing the atomic events described in a collection of documents. We report results from testing our system on several sets of related texts, including human assessments of the system's output and a comparison with information extraction techniques.

1 Introduction

Events have received a lot of attention in the theoretical linguistics literature (e.g., Chung & Timberlake 1985, Bach 1986, Pustejovsky 2000). At the same time, several natural language applications deal with information about events extracted from a text or collection of texts, for example, information retrieval systems participating in the Topic Detection and Tracking initiative (Yang et al. 1999; Allan 2002), and information extraction systems participating in the Message Understanding Conferences (Marsh & Perzanowski 1997). But however intuitive the notion of 'event' might seem, exact definitions of what an event is are usually not supplied, and the implicit definitions seem to conflict. Not only is the exact meaning of an event in dispute, but also the extent of an event's realization in text.

Linguists have been looking at semantic constraints in sentences to distinguish between events, extended events, and states; see for example (Chung & Timberlake 1985, Bach 1986, Pustejovsky 2000). Often in such research the event analysis is based on properties of the verb, and verbs are classified according to their relationships with event classes (Levin 1993).

From a computational perspective, there have been attempts to classify verbs into those denoting events and those denoting processes (Siegel & McKeown 2000) and to investigate what text structure can be considered to be a minimal unit for event description (McCoy & Strube 1999, Filatova & Hovy 2001). The work most commonly referred to as event detection is that originating from the Topic Detection and Tracking (TDT) research effort sponsored by DARPA. An important contribution of that research program is the recognition of the distinction between an event and a topic.

Systems participating in the Scenario Template task of the Message Understanding Conference (Marsh & Perzanowski 1997) competitions use information extracted and inferred from a text to fill in the appropriate fields in predefined templates corresponding to the domain of the text. However, the MUC systems suffer from two drawbacks: First, the fixed templates preclude detecting multiple events of different types, or of types not anticipated during system design. Second, they are heavily dependent on the domain, which requires a lot of time to create accurate templates defining possible events for that particular domain, and even more effort in adapting the system to the sublanguage and knowledge model of that domain.

In this paper, we draw on a synthesis of the above three competing approaches to events (linguistics, information retrieval, and information extraction) to obtain a method for constructing a representation of the *atomic events* in multiple related documents. We aim at small text pieces and multiple low-level events rather than the most generic events targeted by IR, but we incorporate information about the similarity of texts to find topic-specific relationships. We do not rely on predefined templates and slots, as IE does, but we discover relationships in a domain-independent fashion and label them with appropriate verbs and nouns. Our approach is informed by linguistic theory, but remains operational for arbitrary texts.

2 A study of event annotation

As it has been noted in the introduction there is huge diversity in both the structure and length of events. Thus, before going any further we describe a two-stage evaluation experiment the major goal of which is to obtain a definition of events that can be used in a system for the automatic detection and extraction of events from a document.

We conducted our first study of text annotation for event information by asking a number of computer science graduate students (mostly in computational linguistics) to mark text passages that describe important events in news stories. The annotators were given 13 news articles randomly selected from the DUC-2001 (Document Understanding Conference) corpus. The texts varied in length from 15 to 60 sentences. Five of the thirteen texts were each annotated by two participants in the study.

We deliberately provided no definition of *event* for this study, to see if the respondents would naturally converge to an operational definition (as evidenced by high agreement on what they marked). Our study had two further aims: to determine what text range in the absence of instructions on the length of what they should mark, people tend to favor as the appropriate text parts describing an event; and to gather evidence of features that frequently occur in the marked passages and could be automatically extracted by a system simulating the human

annotators.

We noticed substantial disagreement between annotators on what text passages should be marked as events. Since our annotation instructions left unspecified the length of event descriptions, a basic text unit that could be marked or unmarked is not defined either and therefore it is hard to quantitatively measure annotators' agreement.

While the annotators disagreed on what text pieces to select as event descriptions, they exhibited more agreement on how long these pieces should be. Out of 190 text regions marked as events, 46 (24%) consisted of one clause within a longer sentence, 22 (11%) of one sentence minus one short prepositional phrase, 95 (50%) of exactly one sentence, and 27 (14%) of multiple sentences.[1]

We analyzed the passages marked as event descriptions looking for text features that could be included in an automated event detection system. Naturally, the verb itself often provides important information (via tense, aspect, and lexical properties) about the event status of a clause or sentence. In addition, the following features are correlated with the presence of events: *Proper nouns* occur more often within event regions, possibly because they denote the participants in events. In contrast, *pronouns* are less likely to occur in event regions than in non-events. As expected, the presence of *time phrases* increases the likelihood of a text region being marked as an event.

We thus came up with a procedural definition of atomic events which we used for detecting, extracting and labeling of atomic events in our system. The details of this definition are presented in the next section.

3 Detecting and labeling events

Drawing from our event annotation study, we decided on an algorithm for detecting, extracting, and labeling events that is based on the features that seemed more strongly correlated with event regions. Event regions are contained within a sentence. Thus, we anchor events on their major constituent parts (Named Entities for people and organizations, locations, and time information)[2] and expect at least two such major elements in a sentence to consider extracting an event. The procedure for extracting atomic events is the following:

- We analyze a collection of documents clustered on a specific topic.
- We take the sentence as the scope of an event. Our algorithm ignores sentences that contain one named entity or none.

[1] Although words like *war* or *earthquake* can denote events, single nouns were never marked as events by our annotators.

[2] All these major elements can be retrieved with a named entity tagger; we use BBN's IdentiFinder (Bikel et al. 1999).

- We extract all the possible pairs of named entities (preserving the order). Such pairs of named entities are called *relations*.
- For each relation we extract all the words that occur in-between the elements of the relation. These are extracted together with their part of speech tags which we get with the help of Collins' (1996) parser.
- Out of all the words that occur in-between elements of relations we are now interested only in those which are either non-auxiliary verbs or nouns which are hyponyms of *event* or *activity* in WordNet (Miller et al. 1999). We call these words *connectors*.
- For each relation we calculate how many times it occurs, irrespective of the connectors.
- For each connector we calculate how many times this connector is used in a particular relation.

Our hypothesis is that if named entities are often mentioned together, these named entities are strongly related to each other within the topic from which the relation was extracted. Although our method can be applied to a single text (which by itself assures some topical coherence), we have found it benefi cial to extract events from sets of related articles. Such sets can be created by clustering texts according to topical similarity, or as the output of an information retrieval search on a given topic. Following the above procedure we create a list of relations together with their connectors for a set of documents.

Out of all the relations we leave only the top n events that have the highest frequency out of all the relations and we also eliminate those relations that are not supported by high frequency connectors (both of these parameters are adjustable and are determined empirically).

We then examine the graph of connections induced by the surviving pairs. For each two edges in that graph with a common endpoint (e.g., (A, B) and (A, C)), we examine if their list of connectors is substantially similar. We consider two such lists substantially similar if one contains at least 75% of the elements in the other. When that condition applies, we merge the two candidate events into one link between A and a new element $\{B, C\}$ (i.e., we consider B and C identical for the purpose of their relationship to A), and add the scores of the two original events to obtain the score of the composite event.

4 System output

We ran our system on a subset of the topics provided by the Topic Detection and Tracking Phase 2 research effort (Fiscus et al. 1999). The topics consist of articles or transcripts from newswire, television, and radio. We used 70 of the 100 topics, those containing more than 5 but less than 500 texts. Since human annotators created these topical clusters in a NIST-sponsored effort, we can be assured of a certain level of coherence in each topic.

Relation Frequency	First Element	Second Element
0.0212	China Airlines	Taiwan
0.0191	China Airlines	Taipei
0.0170	China Airlines	Monday
0.0170	Taiwan	Monday
0.0170	Bali	Taipei
0.0148	Taipei	Taiwan

Table 1: *Top 6 named entity pairs for the 'China Airlines crash' topic*

Relation	Connector	Connector Frequency
China Airlines – Taiwan	crashed/VBD	0.0312
	trying/VBG	0.0312
	burst/VBP	0.0267
	land/VB	0.0267
China Airlines – Taipei	burst/VBP	0.0331
	crashed/VBD	0.0331
	crashed/VBN	0.0198
Taipei – Taiwan	–	–

Table 2: *Top connectors for three of the relations in Table 1*

TDT provides descriptions of each topic that annotators use to select appropriate documents by issuing and modifying IR queries. Here is the official explication of one topic ("China Airlines crash"):

> *The flight was from Bali to Taipei. It crashed several yards short of the runway and all 196 on board were believed dead. China Airlines had an already sketchy safety record. This crash also killed many people who lived in the residential neighborhood where the plane hit the ground. Stories on topic include any investigation into the accident, stories about the victims/their families/the survivors. Also on topic are stories about the ramifications for the airline.*

Table 1 shows the top 6 pairs of named entities extracted from the topic at the first stage of our algorithm (before considering connectors). The normalized relation frequency is calculated by dividing the score of the current relation (how many times we see the relation within a sentence in the topic) by the overall frequency of all relations within this topic.

It is clear from the table that the top relations mention the airline company whose plane crashed (*China Airlines*), where the crash happened (*Taiwan*, *Taipei*), where the plane was flying from (*Bali*), and when the crash happened (*Monday*). Interestingly, we obtain a clique for the three elements *China Airlines*, *Taiwan* and *Taipei*. Let us analyze the connectors for the three pairs among these three elements (Table 2). The normalized connector frequency is

calculated by dividing the frequency of the current connector (how many times we see this connector for the current relation) by the overall frequency of all connectors for the current relation. According to this table the relation *Taiwan – Taipei* does not have any connectors linking these two named entities. For us it means that the named entities in the relation *Taiwan – Taipei* are linked to each other not through an event but through some other type of static relation (indeed, Taipei is the capital of Taiwan).

Finally, we factor in topic specificity for the extracted events. We calculate the ratio of the relation frequency for a specific topic over the relation frequency of that same pair for all the topics under analysis. This ratio is equal to *1.0* for the relations *China Airlines – Monday*, *Bali – Taipei* and it is equal to *0.0850* for *CNN – New York*. The specificity feature is helpful in deciding what relations are important for a given topic and what are not.

We close this section with a comment on the anchor points used by our algorithm. Such anchor points (by default named entities) are necessary in order to limit the amount of relations considered. We chose named entities on the basis of our analysis of events marked by people (Section 2). However, the system is adaptable and the user can specify additional words or phrases that should be used as anchor points. In this example, if the word *passengers* is submitted to the system, then the third most important event extracted will refer to the deaths of the passengers. The problem of how to find the best candidate nouns to add to the list of nouns which anchor events (by default this list consists only of named entities) is very similar to the problem of creating a right query within Information Retrieval tasks.

5 Comparison with Information Extraction

We compare our system's output to ideal output for one of the most well-known information extraction competitions, the Message Understanding Conferences (Marsh & Perzanowski 1997) organized by NIST between 1992 and 1998. In MUC's Scenario Template task events are extracted for several pre-specified domains. For each domain a list of templates is created in advance and event extraction is equated to filling these templates, a typical approach in information extraction. Events are extracted from one text at a time and not a collection of texts. [3]. Each text can contain one, several, or no events.

MUC systems produce output in the form of predefined well-structured templates. Classical IE systems have developed a repository of powerful tools intended for extracting information within specific domains. Our system based on the presented definition of atomic events is domain-independent; though it does

[3] There are IE systems which try to fill predefined templates from several texts but during the MUC competition systems analyzed and extracted events for one text at a time.

not assign slots to the constituent parts of the extracted events, our system goes beyond the limits of predefined templates and extracts all possible relations it can find. For example, MUC systems are not supposed to extract any events for the following MUC-7 text:

> *Paris (France), 5 April 90 (AFR) – Colombian leader Virgilio Barco briefed French president François Mitterand here Wednesday on the efforts made by Bogota to fight the country's powerful cocaine traffickers. Mr. Barco told reporters after the meeting at the Elysée Palace that the French leader, who visited Bogota in October 1989, had said once again that he was "very interested" in the drug problem.*

This text is from the terrorism domain collection. And though really no terrorist attacks are described in this text it does not mean that there are no events described. These events include the meeting between François Mitterand and Virgilio Barco, Mitterand's earlier visit to Bogota, and Barco's speaking to reporters. Thus, following the described procedural definition, our system extracts information about this meeting by pointing out that the named entities in the relations *François Mitterand – Virgilio Barco*, *François Mitterand – Wednesday* and *Virgilio Barco – Wednesday* are all linked to each other through the connector *briefed/VB*.

In fairness to the MUC systems we note that they perform additional tasks such as the semantic classification of the information (deciding which slot to select for a given piece of extracted text). Our approach does not assign labels such as *perpetrator* or *target* to named entities. It provides for a more superficial "understanding" of the elements of the event and the roles they play in it, in exchange for portability, generality and robustness.

6 System evaluation

6.1 *Methodology*

To evaluate our system we chose randomly 9 topics out of the 70 TDT-2 topics containing more than 5 and less than 500 texts. For each of these topics we randomly chose 10 texts, ran our system on these 10 texts only, and produced a ranked list of events with verb and noun labels, as described above. Each set of ten texts, and the top ten events extracted by our system from it, were given to one evaluator. The evaluators were asked to first read the texts[4] and then provide a numerical score for the system in the following areas:

- Whether the named entities in the events extracted by our system are really related to each other in the texts. A separate score between 0 and 1 was given for each extracted event.

[4] Which was the reason we limited the number of texts per topic to 10.

- Whether the extracted relations between named entities, if valid, are also important. Again a 0 to 1 score was assigned to each extracted event.
- Whether the label(s) provided for a (valid) event adequately describe the relationship between the named entities.

For these three questions, the evaluators gave a separate score for each extracted event. They were free to use a scale of their own choosing between 0 (utter failure) and 1 (complete success).

6.2 *Results*

Table 3 shows the scores obtained during the evaluation. We report the average rating our system obtained on each of the three questions, across both the ten extracted events in each set and the nine evaluators/topics. We also report the percentage of extracted events that received a non-zero score and a score above 0.5.

Question	Avg. rating	% non-zero	% above 0.5
Link quality	0.7506	92.22%	74.44%
Importance	0.6793	95.00%	62.87%
Label quality	0.6178	90.91%	51.09%

Table 3: *Evaluation scores for our system*

We note that the easiest task for the system is to find valid relationships between named entities, where we obtain about 75% precision by either the average score or the number of scores above 0.5. Next comes the task of selecting important links, with precision of 63–68%.[5] The hardest task is to provide meaningful labels for the events; we succeed in this task slightly in more than half of the valid extracted events, or approximately 40% of the total extracted events. Our system is getting lower scores for topics where the event's major constituent parts are not represented by a named entity (i.e., an earthquake topic) or for very disperse topics, (i.e., the topic about the Israeli-Palestinian peace negotiations.) Regardless, our system overall extracted at least somewhat useful information, as manifested by the fact that over 90% of the reported events received non-zero scores.

7 Conclusions

We have reported on an empirical study of event annotation and a new approach for automatic event detection in text. We have implemented a robust, statistical

[5] Importance and label quality are measured only on extracted relations of reasonable quality (with link quality score above 0.5, 74.44% of the total extracted events).

system that detects, extracts, and labels atomic events at the sentence level without using any prior world or lexical knowledge. Our system uses text collections vs. individual texts (cf. IR approach). It extracts relations between named entities and links that relate these named entities (cf. IE approach), though, in contrast to the classical IE task our relations and links are domain-independent and not defined beforehand but built on the fly. To assign labels to the extracted relations our system uses both verbs and nouns (cf. computational linguistics). The system is immediately portable to new domains, and utilizes information present in similar documents to automatically prioritize events that are specific (and therefore likely more interesting) to a given set of documents. Our examination of results and a first small-scale evaluation indicate that the approach is promising as a means for obtaining a shallow interpretation of event participants and their relationships.

The extracted event information can be used for indexing, visualization and question-answering.

Acknowledgments. We wish to thank Kathy McKeown and Becky Passonneau for numerous comments and suggestions on earlier versios of our system, and John Chen for providing tools for preprocessing and assigning parts of speech to the text. We also thank the members of the Natural Language Group and other graduate students at Columbia University who participated in our evaluation experiments. This work was supported by ARDA under Advanced Question Answering for Intelligence (AQUAINT) project MDA908-02-C-0008. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect ARDA's views.

REFERENCES

Allan, James, ed. 2002. *Topic Detection and Tracking: Event-Based Information Organization*. Boston, Mass.: Kluwer Academic.

Bach, Emmon. 1986. "The Algebra of Events". *Linguistics and Philosophy* 9:1.5-16.

Bikel, Daniel M., Richard Schwartz & Ralph Weischedel. 1999. "An Algorithm that Learns What's in a Name". *Machine Learning* 34:1/3.211-231.

Chung, Sandra & Timberlake, Alan. 1985. "Tense, Aspect and Mood". *Language Typology and Syntactic Description* ed. by Timothy Shopen, volume 3, 202-248 (chapter 4) Cambridge, U.K.: Cambridge University Press.

Collins, Michael. 1996. "A New Statistical Parser Based on Bigram Lexical Dependencies". *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics* (*ACL'96*), 184-191. Santa Cruz, Calif.

Filatova, Elena & Eduard Hovy. 2001. "Assigning Time–Stamps to Event–Clauses". *Proceedings of the Workshop on Temporal and Spatial Information Processing at ACL'01*, 445-482. Toulouse, France.

Fiscus, Jon, George Doddington, John Garofolo & Alvin Martin. 1999. "NIST's 1998 Topic Detection and Tracking Evaluation (TDT2)". *Proceedings of the 1999 DARPA Broadcast News Workshop*, 19-24. Herndon, Virginia, U.S.A.

Harman, Donna & Daniel Marcu, eds. 2001. *Proceedings of the Document Understanding Conference* (*DUC-2001*). NIST, New Orleans, U.S.A.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, Ill.: University of Chicago Press.

Marsh, Elaine & Dennis Perzanowski. 1997. "MUC-7 Evaluation of IE technology: Overview of Results". *Proceedings of the 7th Message Understanding Conference* (*MUC-7*). Fairfax, Virginia, U.S.A. — `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/proceedings_index.html` [Source checked in May 2004]

McCoy, Kathleen F. & Michael Strube. 1999. "Taking Time to Structure Discourse: Pronoun Generation Beyond Accessibility". *Proceedings of the 1999 Meeting of the Cognitive Science Society*, 378-383. Vancouver, British Columbia, Canada.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. 1990. "Introduction to WordNet: An Online Lexical Database". *International Journal of Lexicography* 3:4.235-312.

Pustejovsky, James. 2000. "Events and the Semantics of Opposition". *Events as Grammatical Objects* ed. by C. Tenny & J. Pustejovsky, 445-482. Stanford, Calif.: CSLI Publications.

Siegel, Eric V. & Kathleen McKeown. 2000. "Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights". *Computational Linguistics* 26:4.595-628.

Yang, Yiming, Jaime Carbonell, Ralf Brown, Thomas Price, Brian Archibald & Xin Liu. 1999. "Learning Approaches for Detecting and Tracking News Events". *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval* 14:4.32-43.

Index