Cross-document Temporal and Spatial Person Tracking System Demonstration

Heng Ji

Queens College and the Graduate Center

Zheng Chen

The Graduate Center

The City University of New York

New York, NY, 11367

hengji@cs.qc.cuny.edu

zchen1@gc.cuny.edu

Abstract

Traditional Information Extraction (IE) systems identify many *unconnected* facts. The objective of this paper is to define a new cross-document information extraction task and demonstrate a system which can extract, rank and track events in two dimensions: temporal and spatial. The system can automatically label the person entities involved in significant events as '*centroid arguments*', and then present the events involving the same centroid on a time line and on a geographical map.

1 Introduction

Information Extraction (IE) systems can identify 'facts' (entities, relations and events) of particular types within individual documents, and so can unleash the knowledge embedded in texts for many domains, such as military monitoring, daily news, financial analysis and biomedical reports. However, most current IE systems focus on processing single documents and, except for coreference resolution, operate a sentence at a time. The result are large databases containing many *unconnected, unranked, redundant* (and some *erroneous*) facts.

McNamara (2001) proved that a high-coherence text has fewer conceptual gaps and thus requires fewer inferences and less prior knowledge, rendering the text easier to understand. In our task text coherence is the extent to which the relationships between events in a text can be made explicit. We noted that linking all events in temporal and spatial directions for the entire corpus was not feasible because of the large number of event arguments. Grosz et al. (1995) claimed that certain entities are

more central than others and that this property imposed constraints on discourse coherence. Therefore we have developed a system which can extract globally salient and novel arguments as *centroid arguments*, and link all events involving each centroid argument on a time line and on a geographical map.

Beyond extracting isolated facts from individual sentences, we provide coherent event chains so that the users can save time in connecting relevant events and conducting reasoning, such as tracking a person's movement activities and an organization's personnel changes. This will provide a richer set of views than is possible with document clustering for summarization or with topic tracking. In addition, such cross-document extraction results are indexed and allow a fast entity searching mechanism. Beyond traditional search, the system can correlate and organize information across different time series by temporal tracking, and deliver to users in different geographies by spatial tracking.

The rest of this paper is structured as follows. Section 2 presents the overall system architecture including the baseline system and the detailed approaches to extract event chains. Section 3 then presents the experimental results compared to traditional IE. Section 4 demonstrates the system output. Section 5 compares our approach with related work and Section 6 then concludes the paper and sketches our future work.

2 System Overview

In this section we will present the overall procedure of our system.

2.1 Within-document IE

We first apply a state-of-the-art English IE system (Ji and Grishman, 2008) to extract events from each single document. The IE system includes entity extraction, time expression extraction and normalization, relation extraction and event extraction. Entities include persons, locations, organizations, facilities, vehicles and weapons; Events include the 33 distinct event types defined in Automatic Content Extraction (ACE05)[1].

The event extraction system combines pattern matching with statistical models. For every event instance in the ACE training corpus, patterns are constructed based on the sequences of constituent heads separating the trigger and arguments. In addition, a set of Maximum Entropy classifiers are trained: to distinguish events from non-events; to classify events by type and subtype; to distinguish arguments from non-arguments; to classify arguments by argument role; and given a trigger, an event type, and a set of arguments, to determine whether there is a reportable event mention. In addition, the global evidence from related documents is combined with local decisions to conduct cross-document inference for improving the extraction performance as described in (Ji and Grishman, 2008).

2.2 Centroid Argument Detection

After we harvest a large repository of events we can label those important person entities which are involved frequently in events as *'centroid arguments'*. Not only are such arguments central to the information in a collection (high-frequency), they also should have higher accuracy (high-confidence). In this project we exploit global confidence metrics to reach both of these two goals.

For an event mention, the within-document event classifiers produce the following local confidences values:

- *LConf(trigger,etype)*: The probability of a string *trigger* indicating an event mention with type *etype*.
- *LConf(arg, etype)*: The probability that a mention *arg* is an argument of some particular event type *etype*.

- *LConf(arg, etype, role)*: If *arg* is an argument with event type *etype*, the probability of *arg* having some particular *role*.

We use the INDRI information retrieval system (Strohman et al., 2005) to obtain the top N related documents for each test document to form a *topically-related cluster*. The intuition is that if an argument appears frequently as well as with high extraction confidence in a cluster, it is more salient. For each argument *arg* we also added other person names coreferential with or bearing some ACE relation to the argument as *argset*.

In addition we developed a cross-document person name disambiguation component based on heuristic rules to resolve ambiguities among centroid arguments. Then we define the following global metric weighted with the local confidence values to measure *salience*, and generate the top-ranked entities as *centroid arguments*.

- *Global-Confidence(arg):* The frequency of *argset* appearing as an event argument in a cluster, weighted by local confidence values: *LConf(trigger,etype)*LConf(arg, etype)* LConf(arg, etype, role).*

2.3 Cross-document Event Aggregation and Global Time Discovery

If two events involve the same centroid argument, we order them along a time line according to their time arguments and group them into specific geographical locations based on their place arguments. When ordering a pair of entity arguments, we replace pronouns with their coreferential names or nominals, and replace nominals with their coreferential names, if applicable. If the normalized dates are the same for two events, we further compare them based on their time roles (e.g. 'time-end' should be ordered after 'time-beginning').

We start from aggregating events by merging coreferential event mentions using the within-document coreference resolution component in the IE system. However, the degree of similarity among events contained in a group of topically-related documents is much higher than within a document, as each document is apt to describe the main point as well as necessary shared background.

---

[1] http://www.nist.gov/speech/tests/ace/

Relation	Event_i Arguments	Event_j Arguments	Centroid	Event Type	Event Time
Coreference	Entity[Ariel Sharon] Place [Jerusalem]	Entity[Sharon] Place[Jerusalem]	Powell	Contact-Meet	2003-06-20
Subset	Entity[Bush]	Entity[Bush] Place[Camp David]	Blair	Contact-Meet	2003-03-27
Subsumption	Destination[Mideast]	Destination[Egypt]	Bush	Movement-Transport	2003-06-02
Complement	Sentence [nine-year jail] Crime[corruption]	Adjudicator[court] Place[Malaysia] Sentence [nine-year prison]	Anwar Ibrahim	Justice-Sentence	2003-04-18

Table 1. Cross-document Event Aggregation Examples

Therefore in order to maximize *diversity*, we merge any pair of events that have the same event type and involve the same centroid argument, via one of the operations in Table 1.

3 Experimental Results

We used 10 newswire texts from ACE 2005 training corpora as our test. For each test text we retrieved 25 related texts from English Topic Detection and Tracking (TDT-5)[2] corpus which in total consists of 278,108 texts. The IE system extracted 179 event mentions including 140 Name arguments. We define an argument is correctly extracted if its event type, offsets, and role match any of the reference argument mentions.

We found that after ranking with the global confidence metrics, the top-ranked event arguments are substantially more accurate than the arguments as a whole: the overall accuracy without ranking is about 53%; but after ranking the top 85 arguments (61% of total) get accuracy above 70% and the top 116 arguments (83% of total) are above 60% accuracy. It suggests that aggregating and ranking events according to global evidence can enable users to access salient and accurate information rapidly.

4 Demonstration

In this section we will demonstrate the results on all the documents in the English TDT5 corpus. In total 7962 person entities are identified as centroid arguments. The offline processing takes about three hours on a single PC. The real time browsing only takes one second in a standard web browser.

Figure 1 and Figure 2 present the temporal and spatial event chains involving the top 5 centroid arguments: "Bush", "Arafat", "Taylor", "Saddam" and "Abbas". The events involving each centroid are ordered on a time line (Figure 1) and associated with their corresponding geographical codes in a map (Figure 2).

The users can drag the timeline and map to browse the events. In addition, the aggregated event arguments are indexed and allow fast centroid searching. Each argument is also labeled by its global confidence, language sources, and linked to its context sentences and other event chains it is involved. We omit these details in these screenshots.
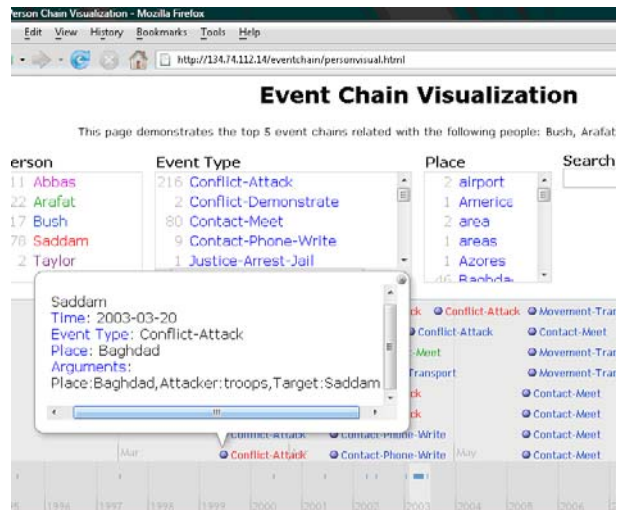

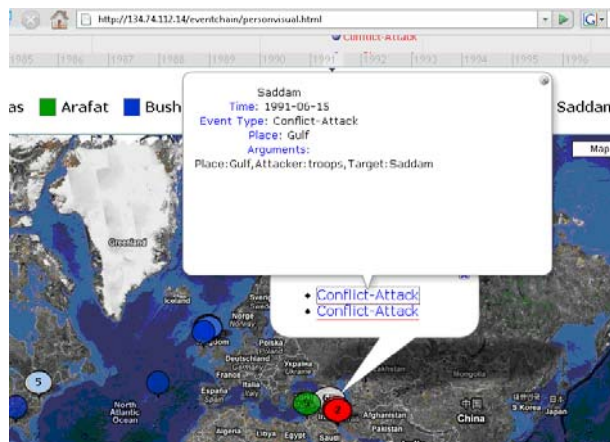Figure 1. Temporal Person Tracking

[2] http://projects.ldc.upenn.edu/TDT5/

Figure 2. Spatial Person Tracking

5   Related Work

Recently there has been heightened interest in discovering temporal event chains. For example, Bethard and Martin (2008) applied supervised learning to classify temporal and causal relations simultaneously. Chambers and Jurafsky (2008) extracted narrative event chains based on common protagonists. In this paper we import these ideas into IE while take into account some major differences. Following the original idea of centering (Grosz et al., 1995) and the approach of centering events involving protagonists (Chambers and Jurafsky, 2008), we introduce a new concept of 'centroid arguments' to represent those entities which are involved in all kinds of salient events frequently. We operate cross-document instead of within-document, which requires us to resolve more conflicts and ambiguities. In addition, we study the temporal and spatial linking task on top of IE results. In this way we extend the representation of each node in the chains to a structured aggregated event including fine-grained information such as event types, arguments and their roles.

6   Conclusion and Future Work

In this paper we described several new modes for browsing and searching a large collection of news articles, and demonstrated a system implementing these modes. We introduced ranking methods into IE, so that the extracted events are connected into temporal and spatial chains and presented to the user in an order of *salience*. We believe these new forms of presentation are likely to be highly beneficial, especially to users whose native language is not English, by distilling the information landscape contained in the large collection of daily news articles – making more information sources accessible and useful to them.

On the other hand, for the users searching news about particular person entities, our system can suggest a list of centroid event arguments as key words, and provide a brief story by presenting all connected events. We believe this will significantly speed up text comprehension. In this paper we only demonstrated the results for person entities, but this system can be naturally extended to other entity types, such as company names to track their start/end/acquire/merge activities. In addition, we plan to automatically adjust cross-document event aggregation operations according to specific compression ratios provided by the users.

References

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. *Proc. ACL-HLT 2008*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised Learning of Narrative Event Chains. *Proc. ACL 2008*.

Barbara Grosz, Aravind Joshi, and Scott Weinstein. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics*, 2(21), 1995.

Heng Ji and Ralph Grishman. 2008. Refining Event Extraction Through Unsupervised Cross-document Inference. *Proc. ACL 2008*.

Danielle S McNamara. 2001. Reading both High-coherence and Low-coherence Texts: Effects of Text Sequence and Prior Knowledge. *Canadian Journal of Experimental Psychology*.

Trevor Strohman, Donald Metzler, Howard Turtle and W. Bruce Croft. 2005. Indri: A Language-model based Search Engine for Complex Queries (extended version). *Technical Report IR-407, CIIR, Umass Amherst, US*.