CAREER: Cross-document Cross-lingual Event Extraction and Tracking

Heng Ji (hengji@cs.qc.cuny.edu)

1 Research Objectives

Information Extraction (IE) is a task of identifying 'facts' (entities, relations and events) within unstructured documents, and converting them into structured representations (e.g. databases). Most current IE systems focus on processing single documents and the result are many *unconnected, unranked, redundant* (and some *erroneous*) facts. The research objective of this proposal is to define several new extensions to the state-of-the-art IE paradigm beyond 'slot filling', and set the following specific aims for processing a large collection of multi-lingual documents:

- **Aim 1: More Accurate IE by Cross-document Inference and Correction**

Event extraction – 'classical' information extraction – remains a very challenging task, because it's situated at the end of an IE pipeline and thus suffers from the errors propagated from upstream processing such as name tagging and entity coreference resolution. Recognizing the different forms in which an event may be expressed, distinguishing events of different types, and finding the arguments of an event are all difficult tasks. Message Understanding Conference (MUC) IE systems in the 1990's rarely broke the 60% 'performance ceiling' (Hirschman 1998), and the NIST Automatic Content Extraction (ACE) systems barely exceeded 50% F-score on argument labeling. When a typical IE system processes one document in a large collection, it uses primarily *prior* knowledge in the form of extraction patterns, classifiers trained on annotated corpora, ontologies, etc. Such knowledge is relatively static -- it is not updated during the extraction process. The system makes only limited use of 'facts' already extracted in the current document, such as names, noun phrases and time expressions. Achieving really high performance for event processing requires that we take a broader view, one that looks outside a single document in order to exploit *posteriori* knowledge. We intend to aggregate similar events, and apply statistical global inference methods to favor consistency of interpretation across documents to enhance the extraction performance. Such methods allow us to glean *dynamic background knowledge* as required to interpret a document and can compensate for the limited annotated training data.

- **Aim 2: More Salient/Complete/Concise/Coherent IE by Cross-document Ranking and Tracking**

Consider a user monitoring or browsing a multi-source multi-lingual news feed, with assistance from an IE system. A stream of such news documents may contain a temporal or locative dimension, typical in stories about an unfolding event. Various events are evolving, updated, repeated and corrected in different documents; later information may override earlier more tentative or incomplete facts. In this environment, traditional single-document IE would be of little value. For example, the Topic Detection and Tracking (Allan, 2002) news corpus contains several hundred articles a day (from several sources). Instead of examining these articles, an IE user would be confronted by thousands of *unconnected* events with tens of thousands of arguments. Add to this the fact that the extracted results contain *unranked, redundant* and *erroneous* facts and some crucial facts are *missing*, and it's not clear whether these IE results are really beneficial. How can we take proper advantage of the power of extraction to aid news analysis? We aim to provide a more coherent presentation by linking events based on shared arguments. In the news from a certain period some entities are more central than others; we propose to identify these *centroid entities*, and then link the events involving the same centroid entity on a time line. In this way we will provide coherent event chains so that users can more efficiently review and analyze events. For instance, business or international affairs analysts review many news reports to track people, companies, and government activities and trends. To aggregate information from wider sources, besides English we also intend to extract event chains from Chinese, then translate and merge the event chains into English.

The remaining of this proposal is organized as follows:

- Formulate a tractable but challenging task of cross-document IE (section 3);
- Present a first cut at a prototype system performing this task and its evaluation results (section 4);
- Lay out the potential main research challenges and the concrete plans to address them (section 5-9);
- Demonstrate the effectiveness of the proposed research through a utility evaluation (section 10);
- Effectively integrate the proposed research into graduate and undergraduate education (section 11).

2 Prior Work

Several recent studies have stressed the benefits of using information redundancy on estimating the correctness of the IE output (Downey et al., 2005), improving disease event extraction (Yangarber et al. 2005, 2006, 2007) and MUC event extraction (Mann, 2007; Patwardhan and Riloff, 2007, 2009).

Text summarization progressed from single-document to multi-document processing by centroid based sentence linking and ranking (e.g. Goldstein et al., 2000; McKeown et a., 2001; Barzilay et al., 2002; Lin and Hovy, 2002; Radev et al., 2004; Mihalcea, 2004; Nenkova, 2005; Nastase, 2008; Wan, 2008). There has been heightened interest in discovering temporal event chains (e.g. Girju, 2003; Lloyd et al., 2005; Chambers and Jurafsky, 2008, 2009). Most of the recent work (e.g. Bouguraev and Ando, 2005; Bramsen, 2006; Lapata and Lascarides, 2006; Bethard et al., 2007, 2008) has been developed around the TempEval task (Verhagen et al., 2007) using TimeBank (Pustejovsky et al., 2003). Various methods have been exploited to identify or infer implicit time arguments (e.g. Filatova and Hovy, 2001; Mani et al., 2003; Lapata and Lascarides, 2006; Mann, 2007; Eidelman, 2008). Our research is also similar to the task of topic detection and tracking (TDT) (Allan, 2002). The Europe Media Monitor (European Commission, 2009) also provides related functions for centroid entity extraction and temporal trend analysis. Yahoo! News service ranks the news articles according to their salience. We will import these ideas of ranking and linking into IE while taking into account some major differences. Following the centering theory (Grosz et al., 1995) and centering events involving protagonists (Chambers and Jurafsky, 2008, 2009), we will propose a new concept of 'centroid entities' to aggregate events and extract event chains across documents. Compared to the multi-document summarization task, we extend the definition of "centroid" from a word to an entity, incorporate quality measures based on confidence estimation into the ranking metric. We also extend the representation of each "node" in the linking task from a document (information retrieval), a story (e.g. TDT and Europe Media Monitor) and a sentence (multi-document summarization) to a structured aggregated event including fine-grained information such as event types, arguments and their roles.

In addition, in our task it's important to disambiguate entities across documents before centroid extraction. Gooi and Allan (2004) compared various clustering algorithms. Fleischman and Hovy (2004) described a two-step approach. The recent research has been mainly promoted in the web people search task (Artiles et al., 2007) such as (Balog et al., 2008), NIST ACE (NIST, 2008) such as (Baron and Freedman, 2008) and TAC KBP (NIST, 2009) evaluations.

One key issue in cross-document IE is to remove redundancy using event coreference resolution. Earlier work on event coreference resolution (e.g. Humphreys et al., 1997; Bagga and Baldwin, 1999) was limited to several MUC scenarios. We will focus on much wider coverage of event types. Our research on event coreference resolution is also related to the information fusion of relations across multiple documents (e.g. Mann, 2007; Downey et al., 2005; Sutton and McCallum, 2004; Finkel et al., 2005).

We intend to automatically detect novel even types based on word clustering. Hirchman et al. (1975) described an automatic word clustering method using the syntactic relations for a sublanguage. Later extensive techniques have been used to cluster words from large unlabeled corpora (e.g. Brown et al., 1990; Pereira et al., 1993; Lee and Pereira, 1999; Lin and Wu, 2009), mono-lingual parallel corpora (e.g. Barzilay and McKeown, 2001; Lin and Pantel, 2001; Ibrahim et al., 2003; Pang et al., 2003), bi-lingual parallel corpora (e.g. Callison-Burch et al., 2005, 2008), and WordNet (Fellbaum, 1998) (Green et al., 2004).

3 A New Cross-document Cross-lingual IE Task

3.1 A Reflection: Traditional Single-document IE

We shall start by illustrating, through the ACE event extraction task, the limitations of traditional single-document IE. ACE defines the following terminology:

entity: a set of objects in one of the semantic categories of interest, e.g. persons, locations, organizations.
relation: one of a specified set of relationships between a pair of entities.
event: a specific occurrence involving participants, including 8 types of events, with 33 subtypes.
event mention: a phrase or sentence within which an event is described.
event trigger: the main word which most clearly expresses an event occurrence.
event argument: an entity involved in an event with some specific role.

For example, for a sentence "*Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment, the entertainment unit of French giant Vivendi Universal*", a single-document ACE IE system should detect the following information:
entity: person: {Barry Diller, chief}; …**relation**: "the entertainment unit" is part of "French Giant"
event trigger: quit; **event argument**: position: "chief"; person: "Barry Diller"; time: Wednesday…

The events extracted from a large corpus by traditional single-document IE are often:
Unconnected and Unordered: For example, the events about the topic "*Tony Blair's visiting other countries*" are presented in 49 isolated event mentions in the extraction output of the TDT-5 corpus.
Unranked: Event mentions are considered equally important.
Redundant: More critically, many events are frequently repeated in different documents. For example, "*Tony Blair met with Bush in Washington on March 27, 2003*" appears five times in the TDT-5 output.
Erroneous and Incomplete: Some extraction errors came from limitations on the use of facts already extracted from other documents. For example, about 50% of events don't include explicit time arguments.

3.2 A Vision: Cross-document Event Extraction and Tracking

A high-coherence text has fewer conceptual gaps and thus requires fewer inferences and less prior knowledge, rendering the text easier to understand (Beck et al., 1991; Britton and Gulgoz, 1991; McNamara, 2001). In our proposed research, text coherence is the extent to which the relationships between events in a text can be made explicit. We aim to explore various new methods of identifying the logical connections among events to reveal this coherence. We noted that linking all events for the entire corpus was not feasible because of the large number of event arguments, and therefore will introduce a new concept of *centroid entity* to address this problem. We define the following new terminology:

- **centroid entities**: N entities most frequently appearing as arguments of events.
- **temporal event chain**: a list of temporally-ordered events involving the same centroid entity.

Given a test set of English and Chinese documents, our cross-document IE task is to identify a set of centroid entities; and then for each centroid entity, link and order the events centered around it on a time line; and then translate the Chinese event information into English. What might such event chains look like? For example, from the following documents about "Stig Toefting" written in Chinese and English:

<DOC1>***<S1-1>周日晚上，丹麦国家队的主力球员托夫丁在丹麦首都哥本哈根市中心的一家餐馆中殴打两名工作人员…<S1-1>***<S1-2>托夫丁因为觉得有一位顾客…</S1-2> …</DOC1>
<DOC2><S2-1>Stig Toefting dropped his appeal of an assault conviction... </S2-1> ***<S2-2>He was sentenced to four months in prison at the end of March.</S2-2>*** <S2-3>The hearing was scheduled for April 10, quoting Toefting's lawyer Anders Nemeth…</S2-3>…</DOC2>
<DOC3>***<S3-1>The Danish international was convicted of assault by Copenhagen City Court this Tuesday after being charged with attacking a restaurant manager at a post-World Cup party. </S3-1>*** <S3-2>Tofting, 33, flew back to England last night…</S3-2> …</DOC3>

For a question "What happened related to Stig Toefting's conviction in 2002?", the possible extracted sentences from a multi-document multi-lingual summarization system are marked in bold, while our cross-document IE approach may produce a temporal event chain as shown in Figure 1.

Time	*2002-01-01*
Event	Attack
Person	***Toefting***
Place	Copenhagen
Target	workers
Context	S1-1
Language	Chinese

Time	*2002-10-15*
Event	Convict
Defendant	***Toefting***
Crime	assault
Context	S3-1, S2-1
Language	English

Time	*2003-03-31*
Event	Sentence
Defendant	***Toefting***
Sentence	four months in prison
Context	S2-2
Language	English

Figure 1. Example of Ranked Cross-document Cross-lingual Temporal Event Chains

Each event is linked to its context sentences and language sources. Each argument will also be labeled by its global confidence, coreferred entity mentions, document IDs, and linked to other event chains it is involved in (we omit these details in this example due to space limit). In addition, such cross-document extraction results will be indexed and allow a fast entity searching mechanism.

3.3 Evaluation Metrics of Research Progress

We propose the following new measures to gauge our research progress on different approaches and assess the impact of the different components in our pipeline.

(1) Centroid Entity Detection

To measure how well a system performs in *selecting the correct centroid entities* in a set of documents, we will compute the Precision, Recall and F-measure of the top N centroid entities identified by the system as a function of N (the value of N can be considered as reflecting the 'compression ratio' in a summarization task):

- *A centroid entity is correctly detected* if its substring and the ID of any document mentioning this entity match a reference centroid. Document IDs are used to distinguish distinct entities with the same name.

In the reference the centroids are the top N entities ranked by the number of events in which that entity appears as an argument. For those correctly identified centroid entities, we will use a standard ranking metric, normalized Kendall tau distance (Kendall, 1938), to evaluate how a system performs in ranking:

- *Normalized Kendall tau distance (Centroid Entities)* = the fraction of correct system centroid entity pairs out of salience order.
- *Centroid Entity Ranking Accuracy* = 1- *Normalized Kendall tau distance (Centroid Entities).*

(2) Browsing Cost: Incorporate Novelty/Diversity into F-Measure

It's important to measure how accurately a system performs at presenting the events involving the centroid entities. The easiest solution is to borrow the argument based F-Measure in the traditional IE task. However, as we pointed out in section 3.1, many events are reported redundantly across multiple documents, so we should incorporate novelty and diversity into the metric and assign penalties to redundancy. We define an evaluation metric *Browsing Cost* which is similar to the *Search Length i* metric (Cooper, 1968) for this purpose:

- *An argument is correctly extracted* in an event chain if its event type, string, role and any document ID match any of the reference argument mentions.
- *Two arguments in an event chain are redundant* if their event types, event time, string (the full or partial name) and roles overlap.
- *Browsing Cost (i)* = the number of incorrect or redundant event arguments that a user must examine before finding i correct event arguments.

We will examine the centroid entities in rank order and, for each argument, the events in temporal order, inspecting the arguments of each event.

(3) Temporal Correlation: Measure Coherence

Since the traditional IE task doesn't evaluate temporal ordering, we will introduce a correlation metric to evaluate how well a system performs at presenting the events in proper temporal order. Assume the event chain *ec* includes a set of correct arguments *args*, then the *temporal correlation* is measured by:

- *Temporal Correlation (ec)* = the correlation of the temporal order of *args* in the system output and reference

In assessing temporal correlation, we should also take into account the number of argument pairs over which temporal order is measured:

- *Argument recall* = number of unique and correct arguments in response / number of unique arguments in key

The general idea follows the event ordering metric in TempEval (Verhagen et al., 2007), but we will evaluate over event arguments instead of triggers because a node in the chain is an aggregated event including fine-grained argument information. Also similar to TempEval we will focus more on the overall temporal order instead of the exact date associated with each individual event. This is different from other time identification and normalization tasks such as TERN (Ferro et al., 2005). In some cases the events can be inserted into the correct positions in the chains even by rough date periods (e.g. “a few weeks ago”). Our proposed temporal correlation metric is able to assign appropriate credit to these cases.

4 A Starting Point: This P.I.’s Earlier Related Research

This P.I. has published 4 invited book chapters and 38 papers at NLP conferences and journals and has experience with the management of cross-site projects of similar scale. This P.I. created an NLP research lab at CUNY, which during the past academic year has published 15 papers including an ACL main conference paper (oral presentation) by an undergraduate student. This P.I. is a recipient of Google Faculty Research Award in 2009. The research topics in this proposal will be systematically built on this P.I.’s prior research. This P.I. has also developed a pilot system to verify the proposed research.

4.1 Information Extraction

Since 2003 this P.I. has been developing an English and Chinese IE system, which achieved about 88%-91% F-measure for name tagging (Ji and Grishman, 2006), 76%-84% F-measure for coreference resolution (Ji et al., 2005b), and 64% for relation extraction. In the ACE05 evaluation, the Chinese system was ranked top 1 on mention detection and top 2 on entity extraction. The event extraction system combines pattern matching with a set of Maximum Entropy classifiers for trigger labeling and argument labeling (Grishman et al., 2005; Ji et al., 2005a; Ji and Grishman, 2008; Chen and Ji, 2009a). Recently substantial improvements in both languages were achieved by cross-lingual trigger clustering (Ji, 2009a) and cross-lingual co-training (Chen and Ji, 2009c). With the system generated entities as input, the English system can achieve 64% F-measure on event trigger labeling and 41% on argument labeling; the Chinese system can achieve 60% on trigger labeling and 44% on argument labeling. Both systems are competitive with the best performing systems on ACE event extraction. This P.I. has also conducted research on event coreference resolution (Chen and Ji, 2009b), which achieved 53% ECM F-measure (Luo, 2005) on system generated event mentions and 87% on perfect event mentions.

4.2 Information Translation

From 2006 to 2009 this P.I. coordinated a cross-site team in developing a Chinese to English entity extraction and translation system for the DARPA GALE program (Ji et al., 2009a). This system provided a 29.3% relative name translation error reduction over a state-of-the-art phrase-based MT system (Zens et al., 2005); and was ranked top 2 in the NIST ACE07 entity translation evaluation. Recently this P.I. proposed a new method to extract and align information networks from comparable corpora and acquired highly accurate name translation pairs (Ji, 2009b). After we extract Chinese event chains we will translate triggers by cross-lingual clustering (Ji, 2009a) and arguments by entity translation (Ji et al., 2009a).

4.3 A Pilot Study Done for the Research Topics in this Proposal

More importantly, this P.I. has developed a pilot system (Ji et al., 2009b) to verify the proposed research (except section 9). A monolingual demo is at: http://nlp.cs.qc.cuny.edu/demo/personvisual.html.

In order to evaluate this pilot system, 106 newswire texts from ACE 2005 training corpora were constructed as a *pilot test set*. Then we extracted the top 40 ranked person names as centroid entities, and manually created temporal event chains as answer keys. We used 278,108 texts from English TDT-5 corpus and 148 million sentences from Wikipedia as our background data. In these event chains there are 140 events with 368 arguments (257 are unique). During the pilot study we have investigated various challenging aspects and recognized the following research topics crucial to this new task.

- **More Accurate and Complete IE**: Exploit knowledge derived from the background data to correct and enrich event argument labeling (section 5) and predict implicit time arguments (section 6).
- **More Salient IE**: Conduct cross-document entity coreference using semantic feedback and centroid entity detection by some salience criteria (section 7);
- **More Concise and Diverse IE**: Conduct cross-document event coreference resolution to remove redundancy (section 8).
- **More Open IE**: Identify novel event types based on word clustering and annotate corpora for new events using active learning (Section 9).

5 Research Topic 1: Cross-document Event Inference and Refinement [Risk: Low]

In this project we intend to investigate the use of cross-document inference to enhance event extraction performance, by favoring interpretation consistency across sentences and documents. This matches the situation of human annotation as well: we may need to frequently consult wider discourse, additional similar web pages or even Wikipedia databases to label events and their arguments correctly.

5.1 Hypotheses

We generally follow the idea of "One Sense Per Discourse"-- the idea of sense consistency introduced in (Gale et al., 1992), extending its scope from a single document to operate across related documents. We have proved the following two hypotheses in (Ji and Grishman, 2008).

- **One Trigger Sense Per Cluster**

Across a heterogeneous document corpus, a particular verb can sometimes be an event trigger and sometimes not, and can represent different event types. However, for a collection of topically-related documents, the distribution may be much more convergent. We found that the likelihood of a candidate word being an event trigger in the test document is much closer to its distribution in the collection of its related documents than in the uniform training corpora. For example, the word "fire" appears 81 times in the training corpora and only 7% of them indicate "End-Position" events; while all of the "fire" in a test document and its related documents are "End-Position" events. So if we can determine the sense (event type) of a word in the related documents, this will allow us to infer its sense in the test document.

- **One Argument Role Per Cluster**

We propose a similar hypothesis for event arguments – one argument role per cluster for event arguments. In other words, each entity plays the same argument role, or no role, for events with the same type in a collection of related documents. For example,

> [***Test Sentence***] ***Vivendi*** *earlier this week confirmed months of press speculation that it planned to* ***shed*** *its entertainment assets by the end of the year.*
> [***Sentences from Related Documents***]***Vivendi*** *has been trying to* ***sell*** *assets to pay off huge debt, estimated at more than $13 billion. Blackstone Group would* ***buy Vivendi****'s theme park division…*

The above test sentence doesn't include an explicit trigger to indicate "Vivendi" as a "seller" of a "Transfer-Ownership" event mention, but "Vivendi" is correctly identified as "seller" in many other related sentences (by matching patterns "[Seller] sell" and "buy [Seller]'s"). So we can incorporate such additional information to enhance the confidence of "Vivendi" as a "seller" in the test sentence.

5.2 Preliminary Experiments

In (Ji and Grishman, 2008) we used the INDRI information retrieval system (Strohman et al., 2005) to obtain a cluster of top N related documents for each test document. For each entity e_i, we collected a set of related entities *argset*: *argset* = {n_j | n_j *is a name,* n_j and e_i are coreferential or linked by a relation; and n_j is involved in an event mention}. Then we computed the global confidence *gc* of e_i based on the sum of local extraction confidence of each member in *argset*, and automatically learned weights for various inference rules to remove triggers and arguments with low confidence, and to adjust trigger and argument labeling to achieve document-wide or cluster-wide consistency. This approach obtained 7.6% higher F-Measure in trigger labeling and 6% higher F-Measure in argument labeling. By using cross-document inference, we can reduce the browsing cost (defined in section 3.3) from 103 to 52 incorrect/redundant arguments before seeing 71 correct arguments in the event chains (Ji et al., 2009b).

5.3 Markov Logic Networks to Model Cross-document Event Inference

The results from the pilot study are promising, but we noticed that heuristic inferences are highly dependent on the order of applying rules, and the performance may have been limited by the thresholds which may overfit a small development corpus. In this project we will attempt to use Markov Logic Networks (Richardson and Domingos, 2006), a statistical relational learning language, to model these global inference rules more declaratively. Markov Logic can be viewed as a formalism that extends first order logic to allow formulae that can be violated with some penalty. It has been proved effective in other NLP tasks such as entity coreference resolution (Poon and Domingos, 2008), temporal relation identification (Yoshikawa et al., 2009) and semantic role labeling (Meza-Ruiz and Riedel, 2009). In Markov Logic we can model our event inference task by introducing a set of logical predicates such as *eventType(trigger, etype)* and *role(arg, role)*. Then we will specify a set of weighted first order formulae that define the inference rules. For example a simple inference rule of adjusting trigger identification to achieve cluster-wide consistency can be given by the formula:

$$isEvent(trigger_i) \wedge stringMatch(trigger_i, trigger_j) \wedge sameCluster(trigger_i, trigger_j) \Rightarrow isEvent(trigger_j)$$

Then our remaining uncertainty with regard to this formula will be captured by a weight associated with it. Markov Logic will make it possible to compactly specify probability distributions over these complex relational inferences. Given a set of weights, candidate event mentions and a given sentence, Markov Logic will infer the event types and argument roles with a maximal posteriori probability. This type of algorithm can also be realized by an Integer Linear Programming based approach (e.g. Lapata and Lascarides, 2006). However Markov Logic allows us to easily capture non-deterministic (soft) rules that tend to hold among event triggers and arguments but do not have to. Exploiting this approach will also provide greater flexibility to incorporate additional linguistic and world knowledge into inference.

6 Research Topic 2: Global Time Reasoning and Prediction [Risk: Moderate]

6.1 Motivation: Half of the events don't include explicit time arguments

After extracting the centroid entities, we intend to link events centered around the same centroid entities on a time line. The text order by itself is a poor predictor of chronological order (only 3% temporal correlation with the true order). Single-document IE technique can identify and normalize event time arguments from the texts, which results in a much better correlation score of 44%. But this is still far from the ideal performance for real applications. Temporal ordering is a challenging task in particular because about half of the event mentions don't include explicit time arguments. In order to alleviate this bottleneck, we intend to exploit global knowledge from the related documents and Wikipedia to recover and predict some implicit time arguments. We are also interested in those more challenging cases in which an event mention and all of its coreferential event mentions do not include any explicit or implicit time expressions; and therefore its time argument can only be predicted based on other related events.

6.2 Background Knowledge Reasoning

For each test document, we will analyze its related documents and Wikipedia and store the extracted events, their time arguments and global confidence into an offline knowledge base. Then if any event mention in the test collection is missing its time argument, we can search for this event type and arguments in the knowledge base (using cross-document name disambiguation technique described in the next section), seeking the time argument with the highest global confidence. For some biographical facts for famous persons, hardly any time arguments can be found from the news articles. However, we can infer them from the knowledge base extracted from Wikipedia. For example, we can find the time argument for the *start-position* event involving "*Diller*" in the following test sentence as "1966":

> [***Test Sentence***] ***<person>Diller</person> started*** *his entertainment career at* ***<entity>ABC</entity>****, where he is credited with creating the ``movie of the week'' concept.*
> [***Sentence from Wikipedia***] *<person>Diller</person> was hired by <entity> ABC</entity> in <time>1966</time> and was soon placed in charge of negotiating broadcast rights to feature films.*

6.3 Cross-Event Time Propagation

Different events with particular types tend to occur together frequently and the news writers rarely provide time arguments for all of these events. Therefore, it's possible to predict the unknown time argument of an event from its related events, especially if they are involved in a precursor/consequence, subevent or causal relation. For example, in the following we can propagate the time "Sunday (normalized into "2003-04-06")" from a "Conflict-Attack" EM_i to a "Life-Die" EM_j :

> [***Sentence including*** $\boldsymbol{EM_i}$] Injured Russian diplomats and a convoy of America's *Kurdish* comrades in arms were among unintended victims caught in **crossfire** and friendly fire ***Sunday***.
> [***Sentence including*** $\boldsymbol{EM_j}$] *Kurds* said 18 of their own **died** in the mistaken U.S. air **strike**.

The useful propagation evidence includes the event types of EM_i and EM_j, whether they are located in the same sentence, if so the number of time expressions in the sentence; whether they share coreferential arguments, if so the roles of the arguments; the distance between them, etc. Some argument types may be more informative to indicate the event time than others, thus we will attempt to identify such "Time-Cue" argument roles. For example, in a "Conflict-Attack" event, "Attacker" and "Target" are more important than "Person" to indicate the event time. The general idea is similar to extracting the cue phrases for text summarization (Edmundson, 1969).

6.4 Preliminary Experiments

In this P.I.'s recent work (Gupta and Ji, 2009), three simple propagation rules were able to correctly predict 74% of the unknown event time arguments with 70% precision. We were able to improve the temporal correlation score from 55.7% (with argument recall 30.7%) to 70.1% (with argument recall 33.1%) (Ji et al., 2009b). So we can generally conclude that our global time prediction methods can deliver significantly better temporal order than single-document IE, and thus more coherent extraction results.

7 Research Topic 3: Centroid Entity Detection [Risk: Low]

7.1 Motivation of Centroid Entity Detection

In this section we will propose the methods to identify centroid entities. Not only are such entities central to the information in a collection (high-frequency), they also should have higher accuracy (high-confidence). Figure 2 and 3 show the argument accuracy results on the pilot test set using the global confidence metric in section 5.2 to measure *salience.*

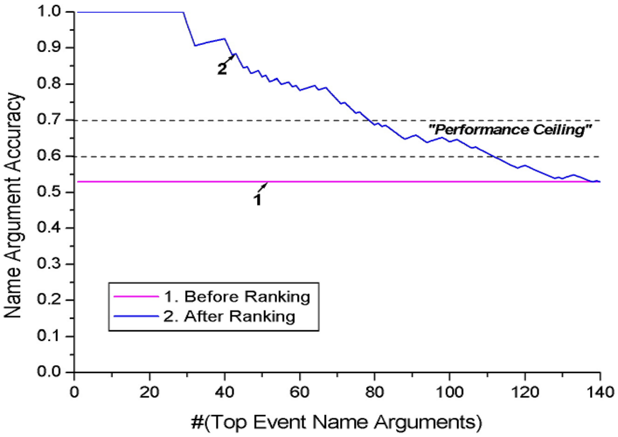

Figure 2. Name Argument Accuracy

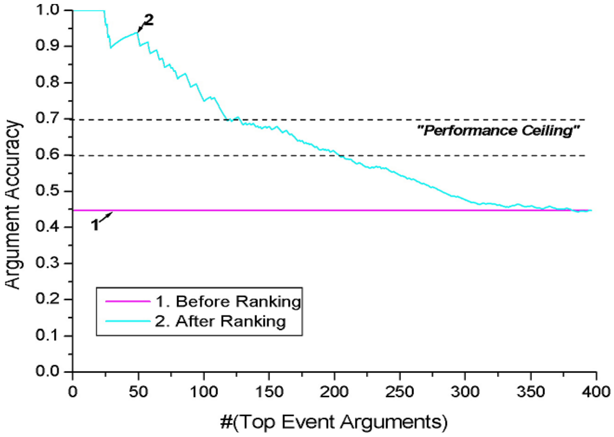

Figure 3. Overall Argument Accuracy

We can clearly see the top-ranked event arguments are substantially more accurate than the arguments as a whole. For the 140 name arguments, the overall accuracy without ranking is about 53%; but after ranking the accuracy of the top 78 is above 70% (a more aggressive 'performance target') and the top 112 arguments are above 60% accuracy (a less aggressive 'performance target'). For the 396 event arguments, the overall argument accuracy is about 45%; however, after ranking, the top 140 arguments achieve accuracy above 70%, and the top 208 arguments achieve accuracy above 60%.

Top Unique Name Arguments			Bottom Unique Arguments		
Argument	Event Type/Role	Salience	Argument	Event Type/Role	Salience
Iraq	Attack/Place	220.78	*them*	Demonstrate/Place	0.117
Baghdad	Attack/Place	70.17	*Amnesty*	Die/Agent	0.115
US	Attack/Attacker	67.81	*convicted*	Life-Die/Victim	0.097
Israel	Attack/Place	11.97	*Al-Sheikh*	Arrest/Place	0.077
Gaza	Die/Place	8.33	*smugglers*	Sentence/Adjudicator	0.075

Table 1. Examples of Centroid Entities and Low-ranked Arguments

Table 1 shows the arguments ranked at the top and bottom after ranking, along with their ranking scores and roles. It clearly indicates that most of the low-ranked arguments are neither particularly important nor meaningful. A lot of them are nominal and pronoun arguments with role annotation errors. It suggests that aggregating and ranking events can enable users to access salient and accurate information rapidly. Therefore the challenge is developing algorithms that can effectively separate the "wheat" from the "chaff" when extracting gazillions of events from a corpus of documents. We will first extract the candidates through cross-document entity coreference and then rank these candidates.

7.2 Cross-document Entity Coreference Resolution Using Global Feedback Knowledge

Centroid detection will benefit from precise clustering of names into correct entities. There are two principal challenges: the same entity can be referred to by more than one name string and the same name string can refer to more than one entity. In this P.I.'s prior work (Ji et al., 2005b), we demonstrated the effectiveness of using semantic relations to improve single-document entity coreference resolution. We aim to extend this work to resolve cross-document entity coreference, and incorporate more semantic knowledge from the feedback of event chains. For example, if we can identify the following overlapped event chain then we can conclude that "Mahmoud Abbas" and "Abu Mazen" are likely to be coreferential.

Time	*1935-03-26*
Event	Life-Born
Person	***Mahmoud Abbas/ Abe Mazen***

Time	*2004-11-11*
Event	Start-Position
Person	***Mahmoud Abbas/ Abe Mazen***
Organization	PLO
Position	Chairman

Time	*2005-01-15*
Event	Start-Position
Defendant	***Mahmoud Abbas/ Abe Mazen***
Organization	PA
Position	President

On the other hand we can filter some spurious cross-document coreference links if an event chain includes conflicting temporal order or arguments. For example, a simple substring-matching approach may fail to disambiguate "Michael Jordan" as an athlete or a politician and generate an erroneous event chain:

Time	*1927-09*
Event	Elected
Person	***Michael Jordan***
Organization	*Dáil éireann*

Time	*1963-02-17*
Event	Life-Born
Person	***Michael Jordan***

Time	*1984*
Event	Start-Position
Defendant	***Michael Jordan***
Organization	Chicago Bulls

But if we can identify the temporal ordering confliction – a "Life-Born" cannot occur after "Elected" for the same entity, this coreference error is easily detected. We plan to apply a statistical cross-document entity coreference resolver to generate multiple coreference hypotheses, and then use the feedback from relation extraction and temporal event extraction to rescore the coreference hypotheses.

7.3 Global Entity Ranking

We will adapt the node centrality problem in graph theory to our centroid detection research. If an entity is involved in events frequently as well as with high extraction confidence, it is more salient. Our basic underlying hypothesis is that the salience of an entity should be calculated by taking into consideration both its confidence and the confidence of other entities connected to it, which is inspired by PageRank (Page et al., 1998) and LexRank (Erkan and Radev, 2004). We propose to compute the salience of and entity e_i based on:

$$salience(e_i) = \sum_j \sum_k salience(n_j, event_k) / nl(n_j),$$

where *salience* will be initialized as the local extraction confidence value by the baseline single-document event extraction, n_j is a name in the related entity set *argset* as defined in the section 5.2, and *nl* indicates the number of entities connected to n_j. In this way we intend to explore more than each individual coreference or relation link, and also analyze the entities that cast the 'vote'.

In our pilot study, we adopted a simple substring matching based cross-document name coreference approach and single-document coreference techniques in (Ji et al., 2005b), and then ranked the entities using the simple metric in section 5.2. The F-measure of detecting the 40 centroid entities was 67.5%

with a ranking accuracy 73.0%, which is much higher than random ranking (42%) and position based ranking (47.3%).

8 Research Topic 4: Cross-document Event Coreference Resolution [Risk: Moderate]

Once the collection grows beyond a certain size, an issue of critical importance is how a human can monitor new event mentions without having to (re) read a large number of earlier event mentions. Two relations are central for event aggregation: *contradiction* – part of one event mention contradicts (is inconsistent with) part of another, and *redundancy* – part of one event mention conveys the same content as (or is entailed by) part of another. Once these central relations are identified they will provide a basis for identifying more complex relations such as elaboration, presupposition or consequence. It is important to note that redundancy and contradiction among event mentions are *logical* relations that are not captured by traditional topic-based techniques for similarity detection (e.g. Allan, 2002; Hatzivassiloglou et al., 1999; Brants and Stolle, 2002). Event coreference resolution is challenging because each linking decision needs to be made based upon the overall similarity of the trigger and multiple arguments. We propose to attempt the following two approaches to enhance event coreference resolution.

8.1 Event Attribute Labeling

We intend to exploit the following event attributes: Modality, Polarity, Genericity and Tense (Sauri et al., 2006). These event attributes will play an important role in event coreference resolution because two event mentions cannot be coreferential if any of the attributes conflict with each other. We will experiment with automatic approaches to label these attributes and study the impact of each individual attribute. Such attempts have been largely neglected in the prior research due to the low weights of attribute labeling in the ACE scoring metric (NIST, 2005). In our pilot study, we demonstrated that simple automatic event attribute labeling can significantly improve event coreference resolution - 6.1% absolute improvement in ECM score over an 80.4% baseline on perfect event mentions. Among these four attributes, it’s not surprising that the tense attribute provided the most significant gain. The performance of our current attribute labeling approaches is still not satisfying: 78.4% F-measure for Polarity, 79.5% for Modality, 88.5% for Genericity and 78.3% for Tense, due to the limited amount and quality of these attributes in ACE data. We plan to exploit more linguistic corpora such as FactBank (Sauri and Pustejovsky, 2009).

8.2 Comparing Agglomerative Clustering and Graph-cut based Clustering

Some very recent work including (Ng, 2009) found that graph-cut based clustering can improve entity coreference resolution. We intend to explore a similar graph based clustering algorithm for event coreference resolution, and compare it with the traditional agglomerative clustering algorithm. We will view the event coreference space as an undirected weighted graph in which the nodes represent all the event mentions in a document and the edge weights indicate the coreference confidence between two event mentions. We will initially construct different graphs for separate event types, such that, in each graph, all the event mentions have the same event type. The problem of event coreference resolution in this framework corresponds to a graph partitioning that optimizes the normalized-cut criterion (Shi and Malik, 2000). Such optimization can be achieved by computing the second generalized eigenvector (“spectral”). In our research we will focus on studying how to compute the coreference matrix (equivalently, the affinity matrix in Shi and Malik’s algorithm). Event mentions include very rich linguistic structures such as triggers, arguments and roles, and thus there will be an excellent opportunity to explore various elements for the coreference matrix. This P.I.’s recent work (Chen and Ji, 2009b) proposed to measure the coreference matrix based on the number of matched trigger pairs and argument pairs, and achieved comparable performance (ECM score 84.5%) with the agglomerative clustering method (ECM score 84.2%).

Besides using cross-document entity coreference results in section 7.2 to measure the similarity between a pair of arguments, we will adopt relation extraction techniques, e.g. using “PART-WHOLE” relation between “Egypt” and “Mideast” to determine that “Destination[Egypt] Bush[Centroid] Time [2003-06-02]” and “Destination[Mideast] Bush[Centroid] Time [2003-06-02]” are likely to be coreferential. We also plan to use the UMD semantic annotation corpus (Dorr and Onyshkevych, 2008) to expand the training data. Finally because we are aggregating events from two languages, we expect that the translation errors will bring us great challenges in determining the relations among arguments. Chen et al. (2003)

showed that translation after event aggregation is better than translation before aggregation for multi-document summarization. We intend to experiment with both pipelines in our new task.

9 Research Topic 5: Novel Event Discovery and Corpus Annotation [Risk: High]

9.1 Portability Issue in IE and Our General Solution

The central goal of our proposal is advancing the *performance* of IE. ACE05 identifies 33 common types of events in the news and provides some fine-grained annotations, and thus provides a good starting point for our research. However, like other researchers in the IE community (e.g. Riloff, 1996; Day et al., 1997; Yangarber et al., 2000; Grishman, 2001; Sudo et al., 2003), we have been aware of the limitation of pre-defined event types in the ACE program. Another central track of IE research is the issue of *portability* – how can an IE system rapidly and automatically (semi-automatically) move to new event types. This goal would have led to another five-year project; in this proposal we will only attempt the following simple approach to semi-automatically discover some novel and salient event types and their annotations, which in principle could be extended to more complete event types:

- Automatically acquire verb clusters and merge with manually constructed verb clusters (section 9.2);
- Rank the verb clusters by their salience and novelty, pre-process their context sentences by our multi-lingual SRL system (Meyers et al., 2009a, 2009b), and then use active learning (e.g. Jones et al., 2003; Riccardi et al., 2004) to annotate argument roles (section 9.3).

9.2 Open-domain Automatic Cross-lingual Verb Clustering

We aim to extensively exploit the manually constructed verb clusters such as the VerbNet (Kipper-Schuler 2006) and FrameNet (Baker et al., 1998). In addition, we will explore open-domain automatic verb clustering methods to increase coverage. For each Chinese verb in the semantic corpora such as PropBank (Palmer et al., 2005, Xue and Palmer, 2009), we will search its aligned English words from the parallel corpora to construct a cluster including frequent English verbs. Then we will acquire Chinese verbs from the other direction and continue the iterations. For example, "announce" is not an ACE-type event, but we can get its cluster from small parallel corpora:

{宣布, 通告，断言，宣告，声明，预报，传达，阐明，提出，显露，显示，陈述，提出…}
→ *{announce, declare, herald, proclaim, set forth, set out, state, unveil, convey, affirm, assert… }*

Using this approach our recent work (Ji, 2009a) extracted 438 English verb clusters and 543 Chinese verb clusters from 50,000 sentence pairs. We will filter the word alignment errors using verb lists, Part-of-Speech tagging, and the following active learning method.

9.3 Active Learning for Novel Event Annotation

After we discover new event types, it will be beneficial to annotate novel events on top of semantic role labeling (SRL) results. If we consider SRL as a simplified 'event extraction' task (considering each verb as a single event type), we may extract an event chain involving "Bush" from the section 2 of PropBank:

Time	*June*
Event	veto
ARG0	***Bush***
ARG1	*measure*

Time	*yesterday*
Event	announce
ARG0	***Bush***
ARG1	*meeting*

Time	*July*
Event	send
ARG0	***Bush***
ARG1	*proposal*
ARG2	*Gorbachev*

Time	*December*
Event	meet
ARG0	***Bush***
ARG0	*Gorbachev*

For each verb cluster *C*, we will gather it together with the entities in ACE corpora *E* as a query, and then use information retrieval methods to obtain *related sentences*. For any verb $v\in C$, if an entity *e* is identified by SRL as an argument of *v* in a sentence *s*, and *v* is not tagged as an ACE trigger, we consider *s* as a *novel-event related sentence*. Then we will compute the salience of C as follows and rank the clusters based on their salience:

$$salience(C,E)=\sum_{v\in C\wedge e\in E}\#novel-event-relatedsentence(v,e)\Big/\sum_{v\in C\wedge e\in E}\#relatedsentence(v,e)$$

For any combination <*v*, *e*, *s*> extracted for a top-ranked cluster, we will apply active learning to annotate the event argument role of *e*. If the SRL confidence for <*v*, *e*> is lower than a threshold, we will manually fix the annotation. Then we will automatically map all SRL roles into event argument roles based on the frame descriptions in PropBank and FrameNet (e.g. map ARG0 of "announce" to "announciator" and ARG1 to "message"). This learning procedure may also involve defining new argument roles and fixing some verb clustering errors. Compared to the prior active learning work for IE, we take into account the novelty of a cluster, and target at open-domain clusters instead of a restricted domain.

The ultimate goal of corpora preparation is to obtain answer-key event chains of about 1,000 centroid entities in each language, from about 2,000 documents, for the extended event types. Each step in the pipeline will be done by two annotators independently and adjudicated for the final answer-key. The inter-annotator agreement for the pilot test corpus is around 90% for event aggregation and 82% for time prediction. In addition we will exploit the corpora in TDT, GALE, and Wikipedia as our background data, in total about 2,000 million tokens in each language.

10 Does Cross-document Cross-lingual IE Help? A Utility Evaluation

10.1 Study Execution

A significant question remains: will the events extracted by cross-document IE actually help end-users to make better use of the large volumes of news? We propose to perform regular extrinsic utility (i.e., usefulness) and usability evaluations on our proposed research. We have performed a preliminary summary writing evaluation on the pilot system. Schmettow (2008) claimed that a sample size larger than five is required to detect a satisfying amount of usability problems, thus we asked 11 students to write summaries for the top 11 centroid entities in our pilot test set by:

- Level (I): Only by reading the news articles, with assistance of keyword based sentence search;
- Level (II): (I) + with assistance from single-document IE results, presented in tabular form and linked back to the context sentences;
- Level (III): (I) + with assistance from cross-document IE results, presented in a graphical interface displaying event chains linked back to the context sentences.

Each of these has to be done in 10 minutes. In order to evaluate these three levels independently, each student was asked to write at most one summary, using one of the three levels, for any single centroid entity. To avoid the impact of diverse text comprehension abilities, each student was involved in all of these three levels for different centroid entities.

10.2 Preliminary Results

- **Observer-based Quantity**

Centroid	(I)	(II)	(III)	Centroid	(I)	(II)	(III)	Centroid	(I)	(II)	(III)
Bush	3/4	5/8	6/6	Al-douri	4/10	4/6	6/7	Ba'asyir	3/4	3/3	5/5
Ibrahim	4/5	5/5	8/8	Giuliani	2/2	3/5	5/5	Erdogan	1/2	4/4	4/4
Toefting	0/0	7/8	4/4	Blair	2/3	3/3	5/5	Diller	3/3	4/5	3/3
Putin	2/3	4/9	7/9	Pasko	3/3	3/3	2/2	**Overall**	**27/39**	**44/59**	**55/58**

Table 2. # (uniquely correct sentences)/#(total extracted sentences in a summary)

Table 2 presents the quantified results for each centroid separately and the overall score. It clearly shows that overall Level (II) contained 17 more correct sentences than the baseline (I), while (III) achieved 11 further correct sentences. (I) obtained significantly fewer sentences due to the ten-minute time limit. We can also see that for some centroid entities such as "Putin", "Giuliani" and "Erdogan", (II) generated more sentences but also introduced more redundant information. The user feedback has indicated that they were not able to get enough time to remove redundancy. In contrast, (III) contained much less redundant information. In fact, the average time the students spent using (III) was only about 7.2 minutes.

- **Observer-based Quality**

The evaluation also showed that (III) produced summaries with better quality: (1) Better pronoun resolution; (2) More complete and accurate temporal order; (3) Can generate abstractive summaries. For example, a sentence "Bush and Blair met at Camp David and the UK three times in March 2003" was derived from three different "Contact-Meeting" events in the event chains. (4) Can connect related events into more concise summaries. For example, two events were connected to generate the sentence “Pasko was appealed for treason crime on April 16, 2003 *and then* released on June 15, 2003”.

- **User-based Usability**

The user feedback also showed that (II) and (III) results were trusted almost equally, and (III) was claimed to provide the most useful functions. The positive comments about (III) include “Temporal Linking allows logical reasoning and generalization”, “Centroid search helps to focus immediately”, “Locative Linking allows to browse all the places which a person has visited” and “Name disambiguation helps to filter irrelevant information”; and the negative comments include “Sometimes IE errors mislead locating the sentences” and “Should have supported name pair search for meeting events”.

10.3 Summary and Plan: Setup an Online News Event Extraction and Tracking System

Our results show that, in comparison to source documents only, the quality of summary reports assembled using cross-document IE was significantly better. Also, as extraction quality increases from no IE at all to single-document IE and then to cross-document IE, user satisfaction increases. In the future we aim to set up an online news article analysis system (including newswire, broadcast transcripts and web blogs) and perform larger and regular utility evaluations. Such measures of the benefits to the eventual end user will provide feedback on what works well and will identify further research.

11 Integrating Cross-document Cross-lingual IE Research into Educational Activities

This P.I.’s educational objectives are to attract undergraduate students to careers in computer science, to inform them about current work in computer science, and to encourage and train computer science graduate and undergraduate students to address research issues in Cross-document Cross-lingual IE.

11.1 CUNY’s Diverse Student Body

The City University of New York (CUNY) is the largest urban public university in the U.S. with almost 500,000 students. CUNY has a long tradition of providing an affordable education to a diverse group of the city’s students; its founding mission is “the provision of equal access and opportunity for students, faculty, and staff from all ethnic and racial groups and from both sexes.” Many CUNY students are from groups the NSF has identified as underrepresented in science, technology, engineering, and mathematics (STEM): (e.g. 62% women, 28.8% African-American, 27.4% Hispanic Americans, 16.4% Asian). They speak 131 native languages in addition to English (e.g. 25% Chinese, 14% Spanish, %10 Hindi, 6% Korean, 5% Urdu, 4.5% Hebrew, 4% Bengali and 4% Russian) and represent 172 countries. This P.I. holds a joint appointment in the Computer Science department at Queens College and the PhD program in Computer Science at the Graduate Center. While the research proposed by this P.I. can increase the accessibility of information, the research process itself will open educational opportunities for students from groups underrepresented in STEM. For example, woman students pursue computer science careers at a quite low rate; they represent 12% of Bachelor’s degree recipients (Computing Research Association, 2009).

11.2 Hierarchical Curriculum Development

This P.I. plans to develop a new sequence of hierarchical courses in Table 3. This P.I. is officially scheduled to teach Level 1 and/or Level 3 during each Spring semester and Level 2 during each Fall semester. This P.I. created a Level 1 course as a CS master thesis requirement in Fall 2008 and as a qualifying exam in Spring 2009. This P.I. led about 20 semester-long student term projects. Five students joined the P.I.’s lab afterwards to continue their research on IE. The creation of these new courses targets a broad undergraduate and graduate audience, and will address a wide range of interesting and practical issues concerning computing, information processing, the Internet, etc., and provide students hands-on NLP tutorials and implementation practice. Developing courses such as Level 3 to contextualize computer science that explore interdisciplinary issues was part of Carnegie Mellon University's successful approach for attracting and retaining more undergraduate women in computing (Margolis & Fisher, 2002); this approach may benefit other underrepresented groups. Beyond that, these courses will expose students to both basic principles and the latest in extraction tools, such as those to be developed and enhanced under the proposed research. As part of the growing computational linguistics program at the Graduate Center, the P.I. will also give guest lectures in the Linguistics department on topics described in this proposal.

Course	Goals	Topics to Cover
Level 1: Introduction to NLP	Stimulate students' interest in NLP and awareness of possible computing careers	Fundamental concepts and methods in NLP
Level 2: Statistical NLP	Enrich students' knowledge of statistical methods and how to use them in NLP research, including IE as case studies	Statistical machine learning methods and their applications in NLP
Level 3: IE Seminar	Stimulate students' interest in IE research; and get them up to speed with state-of-the-art	IE research including the proposed research, the interdisciplinary issue of IE and IR

Table 3. A Hierarchical NLP Curriculum Development Plan

11.3 Research Opportunities for PhD and Undergraduate Students

More directly, in each of the five years this project will support two PhD students in their dissertation studies and two undergraduate students planning to pursue graduate school education in the NLP field. They will be divided into two teams (one PhD and one undergraduate in each team). PhD students will focus on designing and experimenting with various research ideas. Undergraduate students will study the fundamental NLP techniques, get familiar with some basic machine learning methods, and assist in data annotation, algorithm implementations and result analysis.

11.4 Outreach to Non-CS major Undergraduate Students by Utility Evaluation

This P.I. plans to recruit eight undergraduate students who are not in the computer science major to participate in the utility evaluation of our project. The research questions addressed by this project will enable the development of an automatic news article extraction system, which will especially benefit non-native speakers by distilling facts from daily news – giving them a voice as to what features they want from information processing software. Further, this project's graphical interface on time line and geographical map is visually appealing for discussions with a non-technical audience. On the other hand, an effective "Hallway testing" user-study method (Nielsen, 1994) will require a diversity of users but not pre-existing computer skills, and it is a central activity which will make valuable contributions to this research project by pointing out the weaknesses of our methods from the usability point of view and suggesting additional research topics to purse.

Furthermore, each of these non-CS undergraduate students will be asked to take one of the NLP courses this P.I. is teaching. This project will cover their tuition. To make the training most effective, students will be full members of this P.I.'s research lab, will be given project responsibility, and will be able to see what they have accomplished by the end of the project. After finishing the utility evaluations, this P.I. intends to provide more research opportunities to the students who want to continue research in the IE area. In this way the undergraduate students will have research experiences on the project during their junior and senior academic years – at a time when they are deciding whether to pursue graduate school education and what field to study. Undergraduate research can increase retention in STEM of underrepresented groups and raise the likelihood that students pursue graduate education. People-focused computer science research promotes the involvement of underrepresented groups in the field (Margolis and Fisher, 2002). Specifically, research topics with a real-world impact can increase female students' interest (Nagda et al., 1999; Hathaway et al., 2002). Thus this experience will also encourage the various underrepresented groups to pursue careers in computing, and establish an infrastructure for collaborative research between computer science and other departments. The P.I. will actively advertise these positions through student organizations, courses/seminars, undergraduate student mailing lists, personal contacts so that students who are underrepresented in STEM would be likely to be aware of these opportunities.

These undergraduate students will complete a Survey of Undergraduate Research Experiences (SURE II) (Lopatto, 2004), a standard metric that is used to evaluate the impact on students of undergraduate research experiences in the sciences. Pre- and post-questionnaires will assess changes in the views of participants. To document how their research progresses, papers describing this project with participation counts and survey responses will be submitted to the ACM SIGCSE Technical Symposium on Computer

Science Education and ACL workshops on teaching for NLP. These undergraduate students will also give presentations about their work to the other students in their departments.

12 Conclusion

One of the initial goals for IE was to produce a unified database for an entire collection of documents, and allow further logical reasoning on the database. The artificial constraint that extraction should be done independently for each document was introduced in part to simplify the task and its evaluation. We feel the time is now ripe to explore some novel methods to break down the document boundaries and raise IE to a higher level of performance. By a thorough pilot study, we have demonstrated that these new modes of event inference, ranking and tracking can lay the groundwork for enhancing the accuracy and usefulness of the state-of-the-art in IE.

Recently some researchers have explored different aspects similar to our proposed tasks, but there are no standard task definitions, annotated corpora and scoring metrics for a fair comparison. Therefore our research on one coherent project can serve as a useful platform for the IE community. We plan to organize a workshop including a shared task on cross-document cross-lingual IE at ACL conferences, to make our programs and resources freely available. The proposed research would save anybody concerned with staying informed about events an enormous amount of time. The recent research award this P.I. received from Google will further promote this research to process web-scale data. The work will serve as an example in both graduate and undergraduate courses and research offered to a diverse student body. Our research would also have a potential profound benefit in E-Science and E-Learning (Jankowski, 2009). For example, our research can suggest methods to accurately extract and track the related knowledge from the scientific literature. Our techniques can also be potentially used in elementary schools (e.g. providing coherent answers for what/who/when/where/how/why questions) and high schools (e.g. providing people-centered event chains for history courses). These are certainly ambitious applications and require well-developed domain adaptation methods. But our research will bring us closer to the goal.

13 Timeline of Proposed Research and Educational Activities

This work will be led by Prof. Heng Ji at CUNY. Overall coordination of this project will be done through weekly staff meetings. In addition, the P.I. will meet with each student individually twice every week. The research will be conducted over five years. Table 4 shows a timeline of the proposed project with research and educational activities of the P.I. (P), the PhD students (A and B) and the undergraduate students (C and D). Eight non-CS major undergraduate students (U) will perform the utility evaluation.

Task		Y1	S1	Y2	S2	Y3	S3	Y4	S4	Y5	S5
Task Definition and Annotation Guideline		P									
Corpus Annotation	Annotate centroid entities	AC	AC			BC	BC				
	Annotate implicit time	AD	AD			BD					
	Annotate event chains	AD	AD				BD				
Topic 1	Cross-document Inference	PA	PA								
Topic 2	Global Time Discovery	PC	PC								
Topic 3	Cross-doc Entity Coreference			PA	PA						
	Centroid Entity Detection					PAC	PAC				
Topic 4	Event Attribute Labeling				PB	PB	PB				
	Graph-cut based Algorithms							PB	PB	PB	
Topic 5	Verb Clustering			PA	PA				PC	PC	
	Active Learning			PC	PC					PBD	PBD
Online Daily News Processing			PD		PD		PD		PD		PD
Utility Evaluation		U		U		U		U		U	
Curriculum Development Work		P		P		P		P		P	

Table 4. Tentative Timeline of Proposed Tasks for the Five Academic Years (Y#) and Summers (S#)