# Exploiting Background Information Networks to Enhance Bilingual Event Extraction Through Topic Modeling

Hao Li, Heng Ji
*Computer Science Department*
*Queens College and Graduate Center, CUNY*
*New York, USA*
{*haoli.qc,hengjicuny*}@*gmail.com*

Hongbo Deng, Jiawei Han
*Computer Science Department*
*University of Illinois at Urbana-Champaign*
*Urbana-Champaign, USA*
{*hbdeng,hanj*}@*uiuc.edu*

***Abstract*—In this paper, we describe a novel approach of biased propagation based topic modeling to exploit global background knowledge for enhancing both the quality and portability of event extraction on unstructured data. The distributions of event triggers and arguments in topically-related documents are much more focused than those in a heterogeneous corpus. Based on this intuition, we apply topic modeling to automatically select training documents for annotation, and demonstrate it can significantly reduce annotation cost in order to achieve comparable performance for two different languages and two different genres. In addition, we conduct cross-document inference within each topic cluster and show that our approach advances state-of-the-art.**

***Keywords*-Event Extraction; Background Information Network; Biased Propagation based Topic Modeling.**

## I. Introduction

Event extraction is the task of identifying events of a particular type and their participants (arguments) from documents. It is a complex task which suffers from two major problems: (1) quality: challenges in disambiguating event types indicated by trigger words and roles played by arguments; (2) portability: high-cost of manual annotation in obtaining training data. With the rapid growth of new genres, such as web blogs and Twitter which is far more informal and noisy, these challenges become more critical. We found that event extraction performed notably worse on web blogs than on newswire texts. While labeled newswire documents are widely available, labeled informal texts are often expensive to obtain, and are generally scarcely available.

Most of the previous event extraction methods focused on improving the performance for one single document in isolation. When a typical event extraction system processes one document in a large collection, it makes only limited use of local 'facts' already extracted in the current document, such as names, noun phrases and time expressions. However, if we take one step back by looking at human learning, we often see that students study in groups of two or more, mutually searching for the best understanding, solution or meaning; researchers gather together as a "committee" or "panel" to select the best paper/project proposal. Such activities are formalized as "collaborative learning" [26]. Similarly, when dealing with large amounts of data, the event extraction task is naturally embedded in rich contexts. Events no longer exist on their own; they are connected to other topically-related documents, associated with authors (e.g., posters for blogs, reporters for news, speakers for conversation transcripts, authors for papers) and the publication venues (e.g., forums for blogs, agencies for news, conferences for papers), and linked to the geographical places where the documents are published. We call such heterogeneous contexts as the background "information networks" for each candidate event in a test document, as depicted in Figure 1. However, it is not trivial to encode such contextual clues directly into the event extraction system because they are ubiquitously interrelated in various network structures.

In this paper, we propose to directly incorporate multi-dimensional heterogeneous background information networks, through a new and uniformed biased propagation based topic modeling framework as described in our recent work [11].

The underlying intuition is that multi-typed contextual information should be integrated but treated differently in the topic model. This method is designed to imitate human collaborative learning to seek topically-related events as "collaborators" and enhance both training (portability) and testing (quality) an event extractor:

- *training*: automatically select topically-related documents as for training data annotation; we shall demonstrate that this method can significantly reduce annotation cost.
- *testing*: conduct statistical cross-document inference within each topic cluster to favor consistency of interpretation across documents and achieve higher extraction quality; we shall demonstrate topic modeling provides a more effective way than information retrieval (IR)-based document clustering.

We extensively evaluate the proposed approach and compare to state-of-the-art techniques on different data genres (newswire and web blogs) and different languages (English and Chinese). Experimental results demonstrate that the improvement in our proposed approach is language-independent, genre-independent, consistent and promising.

ISBN: 978-1-61208-162-5

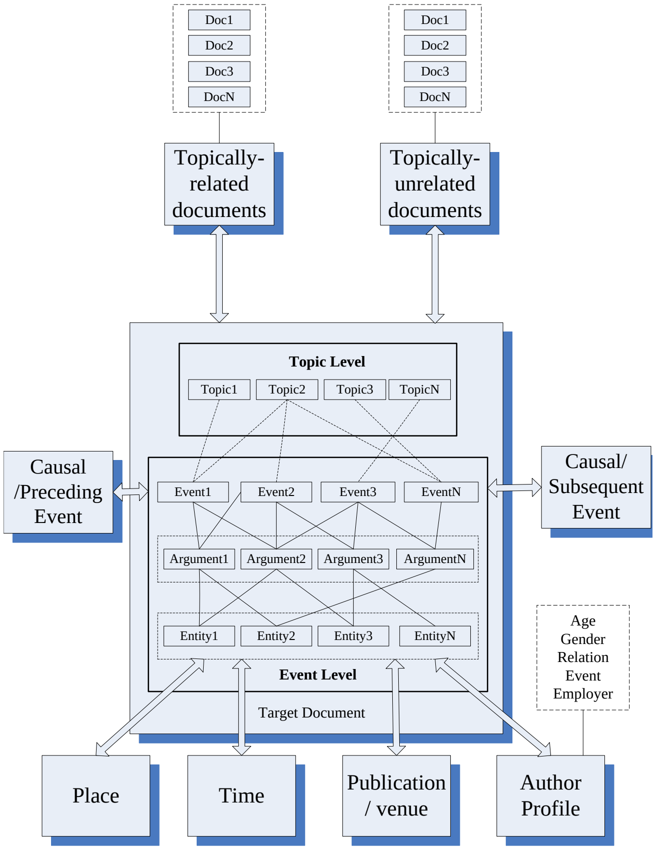

Figure 1. Background Information Networks for Event Extraction

The novel contributions of this paper are two-fold: (1) the first attempt to integrate background knowledge into topic modeling for general news and web blog domains; (2) the first work on exploiting topic modeling to improve both portability and quality of event extraction.

The paper is structured as follows. We briefly review related work in section 2. Section 3 introduces task definition and baseline systems. Then, we propose a novel topic modeling in section 4. In section 5, we apply the topic modeling on event extraction task. The experimental results are presented in section 6. The conclusion and summary is presented in section 7.

## II. Related Work

Some recent work exploited global background knowledge to enhance information extraction tasks, such as entity coreference resolution [25], entity linking [9][13] and relation extraction [7]. Most of these methods incorporated background knowledge from external resources (e.g., Wikipedia). Several recent IE studies have stressed the benefits of using information redundancy on estimating the correctness of the information extraction output [12][18][24][31] or conducting cross-event reasoning [14][21]. We apply topic modeling to locate specific background documents more accurately.

Recently topic models have been successfully applied to various fields of natural language processing, such as Information Retrieval (e.g., [22][29]), Word Sense Disambiguation(WSD) [5], Person Name Disambiguation [27], Text Categorization [34] and Temporal Event Tracking [15]. When reading on-topic stories to understand the events that happened, people tend to segment such stories into various activities (or topics) [32]. Previous research also recognized the benefits of organizing information by events, such as topic detection and tracking [2]. However, very little work has used topic information as feedback to improve event extraction.

In addition, almost all of these previous applications utilized topic models during the test phase, while we demonstrate that topic models can also be used as an effective way to select training data for event extraction, and thus predict the extraction performance before annotating the whole training set. Agichtein and Cucerzan [1] described a language modeling approach to quantify the difficulty of entity extraction and relation extraction. Active learning methods have been applied to reduce annotation cost for information extraction (e.g., [19]). Patwardhan and Riloff [23] also demonstrated that selectively applying event patterns to relevant regions can improve MUC event extraction. Our experiments suggested that topical relatedness can serve as a potential metric to be integrated into other standard data selection criteria in active learning.

## III. Event Extraction Task and Baseline System

### A. Task Definition

The event extraction task we are addressing is that of the Automatic Content Extraction (ACE) [20].

ACE defines the following terminology:

- Event type: a particular event class
- Event trigger: the main word which most clearly expresses an event occurrence
- Event arguments: the mentions that are involved in an event (participants) with particular roles

The 2005 ACE evaluation had 8 types of events, with 33 subtypes; for the purpose of this paper, we will treat these simply as 33 distinct event types. For example, the sentence "*the US-led coalition troops are reportedly thrusting into the second Iraqi city of Basra.*" includes a "*Movement_Transport*" event that is indicated by a trigger word ('*thrusting*"), and a set of event arguments: the Artifact ("*troops*") and the destination ("*Basra*").

We define the following standards to determine the correctness of an event mention:

- A trigger is correctly labeled if its event type/subtype and offsets match a reference trigger.

- An argument is correctly labeled if its event type/subtype, offsets, and role match any of the reference argument mentions.

### B. Baseline Bilingual Event Extraction

We use two state-of-the-art event extraction systems ([8][18]) as our baseline, one for English and the other for Chinese. The system combines pattern matching with a set of Maximum Entropy classifiers incorporating diverse lexical, syntactic, semantic and ontological knowledge. It takes raw documents as input and conducts some pre-processing steps. The texts are automatically annotated with word segmentation, part-of-speech tags, parsing structures, entities, time expressions, and relations. The annotated documents are then sent to the following classifiers: 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 Chinese system incorporates some language-specific features to address the problem of word segmentation and special noun phrase structures. Each component can produce reliable confidence values.

## IV. Capture Background Knowledge Through Topic Modeling

In this section, we will describe a novel topic modeling approach to integrate background knowledge.

### A. Probabilistic Latent Semantic Analysis

Many topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [16] and Latent Dirichlet Allocation (LDA) [4], have been proposed and shown to be useful for document analysis. The basic idea of these approaches to modeling document content is that the probability distribution over words in a document can be expressed as a mixture model of $K$ topics, where each topic is a probability distribution over words. In this paper, we use PLSA as the first topic modeling approach. In PLSA, an unobserved topic variable $z_k \in \{z_1, ..., z_K\}$ is associated with the occurrence of a word $w_i$ in a particular document $d_j$. By summing out the latent variable $z$, the joint probability of an observed pair $(d, w)$ is defined as

$$P(w_i, d_j) = P(d_j) \sum_{k=1}^{K} P(w_i|z_k) P(z_k|d_j), \tag{1}$$

where $P(w_i|z_k)$ is the probability of word $w_i$ according to the topic model $z_k$, and $P(z_k|d_j)$ is the probability of topic $z_k$ for document $d_j$. Following the likelihood principle, these parameters can be determined by maximizing the log-likelihood of a collection $C$ as follows:

$$\mathcal{L}(C) = \sum_i \sum_j N_{ij} \log \sum_{k=1}^{K} P(w_i|z_k) P(z_k|d_j), \tag{2}$$

where $N_{ij}$ denotes the occurrences of word $w_i$ in $d_j$. The model parameters $\{P(w_i|z_k)\}$ and $\{P(z_k|d_j)\}$ are estimated by using a standard Expectation-maximization (EM) algorithm [10]. The estimated conditional probability (e.g., $P(z_k|d_j)$) is used to infer the cluster label for each document.

We use $\{P(z_k|d_j)\}$ as the weights of topics for document $d_j$, and the hidden topics can be regarded as clusters.

### B. Biased Propagation based Topic Modeling

In the meanwhile, in order to emphasis more on event-related entities, we apply an entity-driven topic modeling approach described in our recent work [11], which is more suitable for the event extraction task because each event is associated with a set of entity arguments. For each document and its associated background metadata, we extract the named entities, such as persons and organizations, which may not only be highly correlated with the events but also cover the authors, publication venues and geographical places information of the documents.

We use a state-of-the-art bi-lingual entity extraction system [17] as our baseline to identify entities from English and Chinese. The system is trained on several years of ACE corpora, and can identify entities and classify them as persons, organizations, geo-political entities, locations, facilities, weapons and vehicles. For Chinese data, we applied the Tsinghua word segmenter [28] for pre-processing. The entity extraction system consists of a Hidden Markov Model (HMM) tagger augmented with a set of post-processing rules. The HMM tagger generally follows the Nymble model [3].

In general, the interactions among multi-typed entities play a key role at disclosing the rich semantics of the documents, and it is reasonable to build a ‘virtual document’ for each entity (e.g., person and organization) by aggregating their associated documents. Then, we obtain the term-person matrix $U$ and the term-organization matrix $V$. In this way, documents and their associated entities are composed of words, so each of them can be decomposed by topic models, such as PLSA [16], respectively.

However, this method only considers the textual information while ignores the background network structures between documents and multi-typed entities. Here we apply the topic model with biased propagation (TMBP) [11] between documents and multi-typed entities to directly incorporate the heterogeneous information network with topic modeling in a unified way. The underlying intuition is that multi-typed entities should be treated differently along with their inherent textual information and the rich semantics of the relationships. For example, the topic distribution of an entity without explicit text information (e.g., person $u_l$) depends on the topic distribution of the documents that mention $u_l$. On the other hand, the topic of a document $d_j$ is also correlated with its mentioned entities to some extend, but, most importantly, its topic should be principally determined by its inherent content of the text. Thus, we define a regularization term as

$$
\begin{aligned}
R_U &= \frac{1}{2}\sum_{i=1}^{|\mathcal{D}|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l\in\mathcal{U}_{d_i}}\frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 \\
&+ \frac{\tau}{2}\sum_{l=1}^{|\mathcal{U}|}\sum_{k=1}^{K}\left(P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|}\right)^2 . \qquad (3)
\end{aligned}
$$

A natural explanation of minimizing $R_U$ is that entities should have similar topic distribution with their articles, and vice versa. Note that $\tau$ is the biased parameter. When $\tau \rightarrow \infty$, minimizing $R_U$ will ensure the hypothesis that objects without explicit textual information are completely dependent on the estimated topic distributions of connected documents. Then the objective function $R_U$ can be rewritten as

$$
\begin{aligned}
R_U &= \frac{1}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_l\in U_{d_i}}\frac{P(z_k|u_l)}{|U_{d_i}|}\right)^2 \qquad (4)\\
s.t. &\quad P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|} = 0. \qquad (5)
\end{aligned}
$$

Similarly, we could obtain the regularization term for other entities, e.g., organization $v$.

To incorporate both the textual information and the relationships between documents and multi-typed entities, we define a biased regularization framework by adding the regularization terms to the log-likelihood along with their constraints:

$$
\begin{aligned}
\mathcal{L} &= \sum_i\sum_j N_{ij}\log\sum_{k=1}^{K}P(w_i|z_k)P(z_k|d_j) \\
&\quad -\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{u_j\in\mathcal{U}_{d_i}}\frac{P(z_k|u_j)}{|\mathcal{U}_{d_i}|}\right)^2 \\
&\quad -\frac{\lambda}{2}\sum_{i=1}^{|D|}\sum_{k=1}^{K}\left(P(z_k|d_i) - \sum_{v_j\in\mathcal{V}_{d_i}}\frac{P(z_k|v_j)}{|\mathcal{V}_{d_i}|}\right)^2 \qquad (6)\\
s.t. &\quad P(z_k|u_l) - \sum_{d_i\in\mathcal{D}_{u_l}}\frac{P(z_k|d_i)}{|\mathcal{D}_{u_l}|} = 0, \qquad (7)\\
&\quad P(z_k|v_m) - \sum_{d_i\in\mathcal{D}_{v_m}}\frac{P(z_k|d_i)}{|\mathcal{D}_{v_m}|} = 0. \qquad (8)
\end{aligned}
$$

where $\lambda$ is the regularization parameter which is used to control the balance between the data likelihood and the smoothness of topic distributions. We empirically set $\lambda$ to 1000, and use generalized EM [6] for model fitting.

### C. Performance Comparison

This new TMBP framework has proven much more effective than PLSA on scientific paper (DBLP and NSF award) domain. We believe that it is important to verify its effectiveness on the more general news domain before we apply it to enhance event extraction.

| | NMI(%) | Accuracy(%) |
|---|---|---|
| PLSA | 85.77 | 72.01 |
| TMBP | **90.30** | **84.80** |

Table I
TOPIC MODELING PERFORMANCE

We evaluate the topic modeling approaches on the Topic Detection and Tracking (TDT5) English corpus, which consists of data collected during April to September 2003, and taken from 7 sources, including Agence France Press, Associated Press, CNN, LA Times, New York Times, Ummah and Xinhua. It consists of 10,002 on-topic documents which are classified into 250 semantic topics. In our experiment, those documents appearing in two or more categories were removed, and only the largest 30 categories were kept, thus leaving us with 4,966 documents in total. There are 2,597 unique person entities, 2,161 unique organization entities and 1,199 unique geo-political entities embedded in these documents. There are in total 103,201 links among these entities and documents.

We adopted the following two standard scoring metrics, accuracy (AC) and normalized mutual information (NMI) [30] to measure the topic clustering performance:

$$
AC = \frac{\sum_{i=1}^{n}\delta(a_i, map(l_i))}{n} \qquad (9)
$$

where $n$ denotes the total number of objects; $\delta(x, y)$ equals 1 if $x = y$ otherwise 0; $map(l_i)$ is the mapping function [6] that maps each cluster label $l_i$ to the equivalent label in the corpus.

Given two sets of document clusters $C$ and $C'$, the mutual information metric $MI(C, C')$ is defined as:

$$
MI(C, C') = \sum_{c_i\in C, c'_j\in C'} p(c_i, c'_j)\cdot log_2\frac{p(c_i, c'_j)}{p(c_i\cdot c'_j)} \qquad (10)
$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from $c_i$ and $c'_j$, and $p(c_i, c'_j)$ denotes the joint probability that a arbitrarily selected belongs to both $c_i$ and $c'_j$ at the same time. Let $H(C)$ denote the entropy of $C$, we use the normalized mutual information $NMI$ as the $MI(C, C')$ normalized by $max(H(C), H(C'))$ which reaches from 0 to 1.

Table I shows the performance of PLSA and TMBP on document clustering for TDT5. We can clearly see that TMBP achieved much better performance than PLSA with both scoring metrics.

## V. Applying Topic Modeling to Enhance Event Extraction

### A. Data and Motivations

We use the 109 English newswire documents, 119 English web blogs and 238 Chinese newswire documents from ACE2005 training corpora to evaluate our approach. To simplify the analysis and experiment, we assign the most probable single topic cluster to each document. However, it is worth noting that although we discriminate the most relevant topic cluster and all other clusters, our documents are general news and blog articles and therefore each document may include more than one central topic. Therefore our method is not restricted to documents including single topics.

The most representative words in the resulting 5 (The value of 5 was arbitrarily chosen; variations in this number of clusters produce only small changes in performance) topics from our new topic model are presented in Table II and Table III, which coalesce around reasonable themes. For example, one can easily assemble possible topics correlating with certain types of events (e.g., "*attack*" events involving "*Israel*" in Cluster 1, "*meeting*" events involving "*Korean*" and "*nuclear weapons*" in Cluster 4, and "*Transaction*" and "*Justice*" events in Cluster 5).

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Palestinian | Iraq | Iraqi | north | court |
| Israel | war | forces | nuclear | dollars |
| police | United | Baghdad | Korea | year |
| Israeli | States | Iraq | weapons | appeal |
| people | Bush | troops | Korean | million |
| bank | Nations | city | talks | years |
| year | Iraqi | Saddam | officials | government |
| Monday | minister | military | Washington | convicted |
| killed | council | British | Putin | billion |
| west | resolution | American | south | sentence |
| security | security | officials | China | AFP |
| peace | country | regime | president | group |
| attack | president | army | United | Friday |
| city | role | Iraqis | Russian | April |
| university | Russia | Kurdish | States | life |
| officials | told | control | official | company |
| world | Tuesday | fighting | Pyongyang | case |
| attacks | France | northern | Russia | media |
| military | Washington | force | foreign | charges |
| house | government | Hussein | program | York |

Table II
The Most Probable Words in 5 English Clusters

Across a heterogeneous document corpus, a particular verb can sometimes be an event trigger and sometimes not, and can represent different event types. However, within a cluster of topically-related documents, the distribution is much more focused. The word "*fire*" appears 81 times in the English training corpora and only 7% of them indicate "*End-Position*" events (a person stops working for an organization); while all of the "*fire*" in a topic cluster are "*End-Position*" events. The word "*Da*" appears 58 times in the Chinese training corpora and most of them indicate "*Attack*" events; while all of the "*Da*" in a topic cluster are "*Phone-write*" events (contact by phones or mails). Similarly, each entity tends to play the same argument role, or no role, for events with the same type in a topic cluster.

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Palestine | U.S. | company | team | China |
| Asia | president | court | game | Beijing |
| Israel | alliance | airline | Olympic | development |
| special | Bush | airplane | China | meeting |
| meeting | Relation | personnel | match | progress |
| conflict | State | three | world | national |
| Iraq | Europe | this year | coach | international |
| country | Germany | defendant | sports | country |
| army | problem | police | athletes | city |
| government | Yugoslavia | case | reporter | construction |

Table III
The Most Probable Words in 5 Chinese Clusters (Translated)

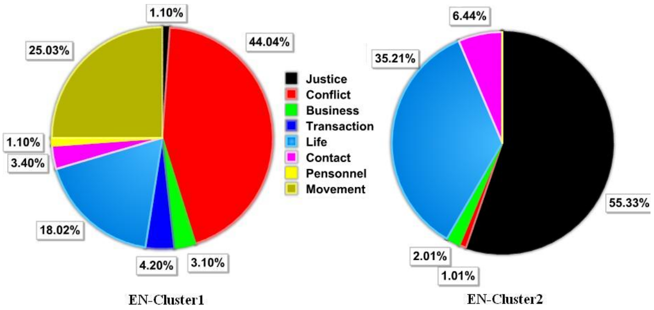

Figure 2. Event Distribution in Two English Clusters

Figure 2 and Figure 3 present the distributions of various event types in different topically-related document clusters. Although both EN-Cluster1 and EN-Cluster2 include certain amount of "*Life*" events (mostly "*Die*" subtypes), we can see that such "*Die*" events in EN-Cluster1 may have been caused by many "*Conflict*" events (44%), while EN-Cluster2 includes very few "*Conflict*" events. Instead, EN-Cluster2 includes many more "*Justice*" events (55%) than EN-Cluster 1(1.1%). Similarly, CH-Cluster1 includes a lot more "*Conflict*" events and fewer "*Movement*" events than CH-Cluster2.

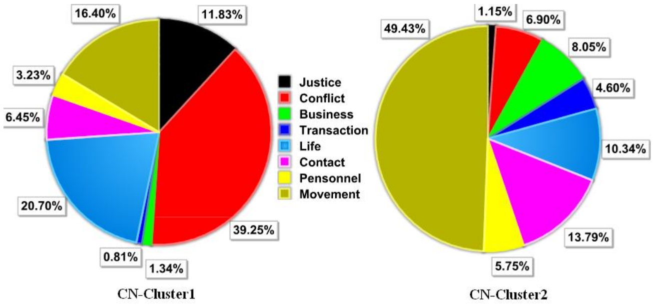

Figure 3. Event Distribution in Two Chinese Clusters

### *B. Topically Related Data is Better Data: Training Data Selection*

Based on the intuition that the likelihood of a candidate word being an event trigger or an entity mention being an event argument in the test document is closer to its distribution in the collection of topically related documents than the uniform training corpora, we design the following active learning approach at selecting training data.

1. Apply topic modeling to the merged set of test documents and training documents to form various topic clusters.

2. For each test document $d_i^j$, which also denotes that $d_i^j$ is the $j^{th}$ document in the $i^{th}$ topic cluster, the procedure can be formalized as follows.

(1) Add all topically-related training documents $\{d_i^k\}$ for training.

(2) Add all topically-unrelated training documents $\{d_l^m | l \neq i\}$ for training.

### *C. Cross-document Inference*

We, generally, follow the hypotheses of "One Trigger Sense Per Cluster" and "One Argument Role Per Cluster" proposed by [18] to conduct cross-document inferences within each topic cluster. If we can determine the event type of a word or the role of an entity within a cluster of topically-related documents, this will allow us to infer its label in the test document. This method can fix event annotation errors produced by single-document extraction. Within each cluster, we conduct two types of inferences to favor interpretation consistency across documents:

- to remove triggers and arguments with low local and cluster-wide confidence;
- to adjust trigger and argument labeling to achieve cluster-wide consistency.

Ji and Grishman [18] required a large external collection of documents which were presumably topically related with the test set. In contrast, we found that the quality of extraction can be improved by partitioning the test set itself using topic models. In addition, they sent the candidate events produced by their baseline system as a query to an IR system to obtain the cluster for each test document. Therefore their inference performance may be limited by the quality of baseline extraction. In our topic modeling approach, we are able to take into account both candidate events and informative context words.

## VI. Experimental Results

In this section, we present the results of applying this new topic modeling method to improve event extraction through extensive experiments.

### *A. Training Data Selection Results*

For the active learning experiments, we setup a baseline passive learning approach by randomly selecting the same number of training documents for each test document.

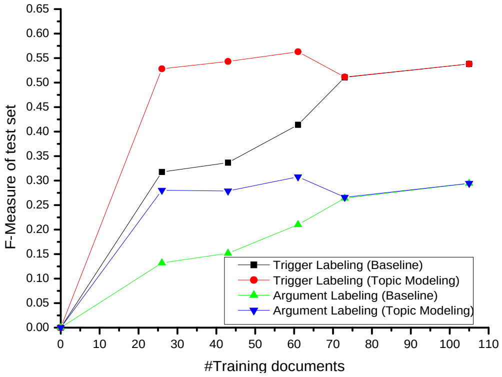

Figure 4. English Newswire Event Extraction

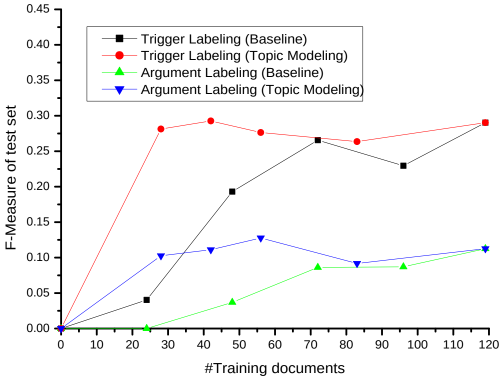

Figure 5. English Web Blog Event Extraction

Because of the data scarcity, leave-one-document-out cross-validation was used to train and test the event extraction systems. Figure 4, Figure 5 and Figure 6 present the F-measure results for both trigger labeling and argument labeling in two languages. The *x* axis in each figure shows the average number of training documents. The first point on each topic modeling curve indicates using all of the topically-related documents at once.

As expected, the baseline approach based on random

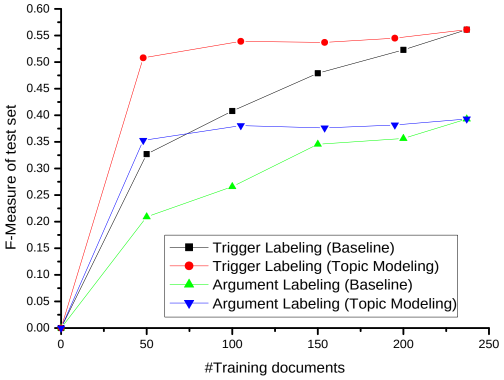

Figure 6. Chinese Newswire Event Extraction

| Performance | | | Trigger Labeling | | | Argument Labeling | | |
|---|---|---|---|---|---|---|---|---|
| System | | | P | R | F | P | R | F |
| English newswire | Baseline | | 74.1 | 49.6 | 59.4 | 50.4 | 28.7 | 36.6 |
| | Cross-doc Inference | IR | 66.5 | 67.4 | 66.9 | 60.8 | 32.2 | 42.1 |
| | | Topic Modeling | **73.3** | **66.3** | **69.6** | **59.4** | **36.5** | **45.2** |
| English web blog | Baseline | | 43.2 | 29.4 | 35.0 | 20.9 | 15.6 | 17.9 |
| | Cross-doc Inference | IR | 38.5 | 42.6 | 40.4 | 30.2 | 21.4 | 25.0 |
| | | Topic Modeling | **41.9** | **43.8** | **42.8** | **32.3** | **23.8** | **27.4** |
| Chinese newswire | Baseline | | 78.8 | 48.3 | 60.0 | 60.6 | 34.3 | 43.8 |
| | Cross-doc Inference | IR | 69.9 | 62.3 | 65.9 | 67.5 | 38.3 | 48.9 |
| | | Topic Modeling | **76.5** | **61.9** | **68.4** | **66.4** | **42.4** | **51.8** |

Table IV
CROSS-DOCUMENT INFERENCE RESULTS

selection produced almost linear increase as we add more and more training documents. In contrast, using only the topically-related documents, we can achieve comparable results as using the whole data sets. This indicates that our topic modeling based approach can dramatically speed up training data selection. Using the same amount of training data at the first point of each curve, topic modeling-based selection performs much better than random selection (18.1%-24.3% higher F-measure on trigger labeling and 10.3%-14.8% higher F-measure on argument labeling). For example, the named entity "*Putin*" appeared as different roles in various types of events in the English newswire data set, including "*meeting/entity*", "*movement/person*", "*transaction/recipient*" and "*election/person*". But the topic model was able to successfully divide them into different clusters, for example, "*Putin*" only played as an "*election/person*" in one cluster. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on a document basis. The results show that the improvement using topically-related data over random selection is significant at more than 99.9% confidence levels for both trigger labeling and argument labeling in any language and genre. In fact, comparing the results using three clusters and all five clusters, we can see that adding topically unrelated documents can hurt performance for English.

### B. *Cross-document Inference Results*

In order to conduct a fair comparison, we duplicated the IR-based clustering approach described in [18], and selected a similar size of data set for blind test (55 documents) for each setting (English newswire, English webblog and Chinese newswire). We then use all other documents in ACE2005 training corpora to train baseline event extraction systems.

Table IV shows the overall Precision (P), Recall (R) and F-Measure (F) scores for the blind test sets. Cross-document inference within topic clusters provided significant improvement over the baselines in both trigger labeling and argument labeling. We can also see that topic modeling-based approach achieved further improvement over the IR-based clustering approach. We conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on a document basis. The results show that the improvement using topic modeling over IR-based clustering is significant at more than 96% confidence levels for both trigger labeling and argument labeling for any language and genre.

## VII. CONCLUSIONS AND FUTURE WORK

Most previous event extraction methods did not explore semantic links across multiple documents and background knowledge. In this paper, we described a novel genre-independent and language-independent topic modeling approach which structurally integrates interconnected entities and events across many documents. The resulting topic models were then used to effectively select training data and conduct global inference for event extraction. We expect this new framework will be also beneficial for other information extraction tasks. In the future we will aim to measure and reduce the impact of topic modeling errors on event extraction. We are also interested in extending our method to jointly enhance cross-lingual topic modeling [33] and event extraction.

## ACKNOWLEDGMENTS

This work was supported in part by the U.S. NSF IIS-09-05215, the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. EAGER grant No. 1144111 and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

[1] E. Agichtein and S. Cucerzan. Predicting accuracy of extracting information from unstructured text collections. In *CIKM*, pages 413–420, 2005.

[2] J. Allan. Topic detection and tracking: Event-based information organization. In *Springer*, 2002.

[3] D. M. Bikel, S. Miller, R. M. Schwartz, and R. M. Weischedel. Nymble: a high-performance learning namefinder. In *ANLP*, pages 194–201, 1997.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. L. Boyd-Graber, D. M. Blei, and X. Zhu. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033, 2007.

[6] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *CIKM*, pages 911–920, 2008.

[7] Y. S. Chan and D. Roth. Exploiting background knowledge for relation extraction. In *COLING*, pages 152–160, 2010.

[8] Z. Chen and H. Ji. Can one language bootstrap the other: A case study on event extraction. In *HLT-NAACL 2009 Workshop on Semi-supervised Learning for Natural Language Processing*, 2009.

[9] Z. Chen and H. Ji. Collaborative ranking: A case study on entity linking. In *EMNLP*, 2011.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38*, 1977.

[11] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*, pages 1271–1279, 2011.

[12] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *IJCAI*, pages 1034–1041, 2005.

[13] N. Fernandez, J. A. Fisteus, L. Sanchez, and E. Martin. Webtlab: A cooccurencebased approach to kbp 2010 entity-linking task. In *TAC 2010 Workshop*, 2010.

[14] P. Gupta and H. Ji. Predicting unknown time arguments based on cross-event propagation. In *ACL/AFNLP (Short Papers)*, pages 369–372, 2009.

[15] V. Ha-Thuc, Y. Mejova, C. Harris, and P. Srinivasan. A relevance-based topic model for news event tracking. In *SIGIR*, pages 764–765, 2009.

[16] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[17] H. Ji and R. Grishman. Improving name tagging by reference resolution and relation detection. In *ACL*, 2005.

[18] H. Ji and R. Grishman. Refining event extraction through cross-document inference. In *ACL*, pages 254–262, 2008.

[19] R. Jones, R. Ghani, T. Mitchell, and E. Riloff. Active learning for information extraction with multiple view feature sets. In *ECML Workshop on Adaptive Text Extraction and Mining*, 2003.

[20] LDC. Automatic content extraction (ace): http://www.itl.nist.gov/iad/mig/tests/ace/. 2005.

[21] S. Liao and R. Grishman. Using document level cross-event inference to improve event extraction. In *ACL*, pages 789–797, 2010.

[22] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *SIGIR*, pages 186–193, 2004.

[23] S. Patwardhan and E. Riloff. Effective information extraction with semantic affinity patterns and relevant regions. In *EMNLP-CoNLL*, pages 717–727, 2007.

[24] S. Patwardhan and E. Riloff. A unified model of phrasal and sentential evidence for information extraction. In *EMNLP*, pages 151–160, 2009.

[25] A. Rahman and V. Ng. Coreference resolution with world knowledge. In *ACL*, pages 814–824, 2011.

[26] B. L. Smith and J. T. MacGregor. What is collaborative learning? In *Collaborative Learning: A Sourcebook for Higher Education*, 1992.

[27] J. Sun, T. Wang, L. Li, and X. Wu. Person name disambiguation based on topic model. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2010.

[28] M. Wan and Z. Luo. Study on topic segmentation method in automatic abstracting system. In *Natural Language Processing and Knowledge Engineering*, 2003.

[29] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR*, pages 178–185, 2006.

[30] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR*, pages 267–273, 2003.

[31] R. Yangarber. Verification of facts across document boundaries. In *International Workshop on Intelligent Information Access*, 2006.

[32] J. M. Zacks and B. Tversky. Event structure in perception and conception. In *Psychological Bulletin, Vol. 127 (2001), pp. 3-21*, 2001.

[33] D. Zhang, Q. Mei, and C. Zhai. Cross-lingual latent topic extraction. In *ACL*, pages 1128–1137, 2010.

[34] S. Zhou, K. Li, and Y. Liu. Text categorization based on topic model. In *Lecture Notes in Computer Science, Volume 5009/2008, 572-579*, 2008.