Sentence Ordering with Event-Enriched Semantics and Two-Layered Clustering for Multi-Document News Summarization

Renxian Zhang **Wenjie Li** **Qin Lu**
Department of Computing, the Hong Kong Polytechnic University
{csrzhang,cswjli,csluqin}@comp.polyu.edu.hk

Abstract

We propose an event-enriched model to alleviate the semantic deficiency problem in the IR-style text processing and apply it to sentence ordering for multi-document news summarization. The ordering algorithm is built on event and entity coherence, both locally and globally. To accommodate the event-enriched model, a novel LSA-integrated two-layered clustering approach is adopted. The experimental result shows clear advantage of our model over event-agonistic models.

1 Introduction

One of the crucial steps in multi-document summarization (MDS) is information ordering, right after content selection and before sentence realization (Jurafsky and Martin, 2009:832–834). Problems with this step are the culprit for much of the dissatisfaction with automatic summaries. While textual order may guide the ordering in single-document summarization, no such guidance is available for MDS ordering.

A sensible solution is ordering sentences by enhancing coherence since incoherence is the source of disorder. Recent researches in this direction mostly focus on local coherence by studying lexical cohesion (Conroy et al., 2006) or entity overlap and transition (Barzilay and Lapata, 2008). But global coherence, i.e., coherence between sentence groups with the whole text in view, is largely unaccounted for and few efforts are made at levels higher than entity or word in measuring sentence coherence.

On the other hand, event as a high-level construct has proved useful in MDS content selection (Filatova and Hatzivassiloglou, 2004; Li et al., 2006). But the potential of event in summarization has not been fully gauged and few publications report using event in MDS information ordering. We will argue that event is instrumental for MDS information ordering, especially multi-document news summarization (MDNS). Ordering algorithms based on event and entity information outperform those based only on entity information.

After related works are surveyed in section 2, we will discuss in section 3 the problem of semantic deficiency in IR-based text processing, which motivates building event information into sentence representation. The details of such representation are provided in section 4. In section 5, we will explicate the ordering algorithms, including layered clustering and cluster-based ordering. The performance of the event-enriched model will be extensively evaluated in section 6. Section 7 will conclude the work with directions to future work.

2 Related Work

In MDS, information ordering is often realized on the sentence level and treated as a coherence enhancement task. A simple ordering criterion is the chronological order of the events represented in the sentences, which is often augmented with other ordering criteria such as lexical overlap (Conroy et al., 2006), lexical cohesion (Barzilay et al., 2002) or syntactic features (Lapata 2003).

A different way to capture local coherence in sentence ordering is the Centering Theory (CT, Grosz et al. 1995)-inspired entity-transition approach, advocated by Barzilay and Lapata (2005, 2008). In their entity grid model, syntactic roles played by entities and transitions between these syntactic roles underlie the coherence patterns between sentences and in the

Coling 2010: Poster Volume, pages 1489–1497,
Beijing, August 2010

whole text. An entity-parsed corpus can be used to train a model that prefers the sentence orderings that comply with the optimal entity transition patterns.

Another important clue to sentence ordering is the sentence positional information in a source document, or "precedence relation", which is utilized by Okazaki et al. (2004) in combination with topical clustering.

Those works are all relevant to the current work because we seek ordering clues from chronological order, lexical cohesion, entity transition, and sentence precedence. But we also add an important member to the panoply – event.

Despite its intuitive and conceptual appeal, event is not as extensively used in summarization as term or entity. Filatova and Hatzivassiloglou (2004) use "atomic events" as conceptual representations in MDS content selection, followed by Li et al. (2006) who treat event terms and named entities as graph nodes in their PageRank algorithm. Yoshioka and Haraguchi (2004) report an event reference-based approach to MDS content selection for Japanese articles. Although "sentence reordering" is a component of their model, it relies merely on textual and chronological order. Few published works report using event information in MDS sentence ordering.

Our work will represent text content at two levels: event vectors and sentence vectors. This is close in spirit to Bromberg's (2006) enriched LSA-coherence model, where both sentence and word vectors are used to compute a centroid as the topic of the text.

3 Semantic Deficiency in IR-Style Text Processing

As automatic summarization traces its root to Information Retrieval (IR), it inherits the vector space model (VSM) of text representation, according to which a sentence is treated as a bag of words or stoplist-filtered terms. The order or relation among the terms is ignored. For example,

1a) *The storm killed 120,000 people in Jamaica and five in the Dominican Republic before moving west to Mexico*.

1b) [*Dominican*, *Mexico*, *Jamaica*, *Republic*, *five*, *kill*, *move*, *people*, *storm*, *west*]

1c) [*Dominican Republic*, *Mexico*, *Jamaica*, *people*, *storm*]

1b) and 1c) are the term-based and entity-based representations of 1a) respectively. They only indicate what the sentence is about (i.e., some happening, probably a storm, in some place that affects people), but "aboutness" is a far cry from informativeness. For instance, no message about "people in which place, Mexico or Jamaica, are affected" or "what moves to where" can be gleaned from 1b) although such message is clearly conveyed in 1a). In other words, the IR-style text representation is semantically deficient.

We argue that a natural text, especially a news article, is not only about somebody or something. It also tells what happened to somebody or something in a temporal-spatial manner. A natural approach to meeting the "what happened" requirement is to introduce event.

4 Event-Enriched Sentence Representation

In summarization, an event is an activity or episode associated with participants, time, place, and manner. Conceptually, event bridges sentence and term/entity and partially fills the semantic gap in the sentence representation.

4.1 Event Structure and Extraction

Following (Li et al. 2006), we define an event E as a structured semantic unit consisting of one event term $Term(E)$ and a set of event entities $Entity(E)$. In the news domain, event terms are typically action verbs or deverbal nouns. Light verbs such as "take", "give", etc. (Tan et al., 2006) are removed.

Event entities include named entities and high-frequency entities. Named entities denote people, locations, organizations, dates, etc. High-frequency entities are common nouns or NPs that frequently participate in news events. Filatova and Hatzivassiloglou (2004) take the top 10 most frequent entities and Li et al. (2006) take the entities with frequency > 10. Rather than using a fixed threshold, we reformulate "high-frequency" as relative statistics based on (assumed) Gaussian distribution of the entities and consider those with z-score > 1 as candidate event entities.

Event extraction begins with shallow parsing and named entity recognition, analyzing each

sentence S into ordered lists of event terms $\{t_1, t_2, \ldots\}$. Low-frequency common entities are removed. If a noun is decided to be an event term, it cannot be (the head noun of) an entity.

The next step is to identify events with event terms and entities. Filatova and Hatzivassiloglou (2003) treat events as triplets with two event entities sandwiching one connector (event term). But the number restriction on entities is counterintuitive and is dropped in our method. We first identify $n + 1$ Seg_i segmented by n event terms t_j.

$$\underbrace{\ldots}_{Seg_0}\ t_1 \ldots \ldots\ t_{j-1}\ \underbrace{\ldots}_{Seg_j}\ t_j\ \underbrace{\ldots}_{Seg_j}\ t_{j+1}\ \ldots \ldots\ t_n\ \underbrace{\ldots}_{Seg_n}$$

Figure 1. Segments among Event Terms

For each t_j, the corresponding event E_j are extracted by taking t_j and the event entities in its nearest entity-containing Seg_p and Seg_q.

$E_j = [t_j, Entity(Seg_p) \cup Entity(Seg_q)]$ (Eq. 1)

where $p = \text{argmax}_{0 \le i \le j-1} Entity(Seg_i) \neq \emptyset$ and $q = \text{argmin}_{j+1 \le i \le n} Entity(Seg_i) \neq \emptyset$ if such p and q exist. 1d) is the event-extracted result of 1a).

1d) {[*killed*, [*storm*, *people*, *Jamaica*, *Dominican Republic*]], [*moving*, [*people*, *Jamaica*, *Dominican Republic*, *west*, *Mexico*]]}

From this representation, it is easy to identify the two events in sentence 1a) led by the event terms "killed" and "moving". Unlike the triplets (two named entities and one connector) in (Filatova and Hatzivassiloglou 2003), an event in our model can have an unlimited number of event entities, as is often the real case. Moreover, we can tell that the "killing" involves "people", "storm", "Jamaica", etc. and the "moving" involves "Jamaica", "Dominique Republic", etc.

The shallow parsing-based approach is admittedly coarse-grade (e.g., "storm" is missing from the "moving" event), but the extracted event-enriched representations help to alleviate the semantic deficiency problem in IR.

4.2 Event Relations

The relations between two events include event term relation and event entity relation. Two events are similar if their event terms are similar and/or their event entities are similar. Such similarities are in turn defined on the word level. For event terms, we first find the root verbs of deverbal nouns and then measure verb similarity by using the fine-grained relations provided by VerbOcean (Chklovski and Pantel, 2004), which has proved useful in summarization (Liu et al., 2007). But unlike (Liu et al., 2007), we count in all the verb relations except *antonymy* because considering two antonymous verbs as similar is counterintuitive. The other four relations – *similarity*, *strength*, *enablement*, *before* – are all considered in our measurement of verb similarity. If we denote the normalized score of two verbs on relation i as $VO_i(V_1, V_2)$ with i = 1, 2, 3, 4 corresponding to the above four relations, the term similarity of two events $\mu_t(E_1, E_2)$ is defined as in Eq. 2, where ε is a small number to suppress zeroes. ε = 0.01 if $VO_i(V_1, V_2) = 1$ and otherwise $\varepsilon = 0$.

$\mu_t(E_1, E_2) = \mu_t(Term(E_1), Term(E_2)) = 1 - \prod_{i=1}^{4}(1 - VO_i(Term(E_1), Term(E_2)) + \varepsilon)$ (Eq. 2)

Entity similarity is measured by the shared entities between two events. Li et al. (2006) define entity similarity as the number of shared entities, which may unfairly assign high scores to events with many entities in our model. So we decide to use the normalized result as shown in Eq. 3, where $\mu_e(E_1, E_2)$ denotes the event entity-based similarity between events E_1 and E_2.

$$\mu_e(E_1, E_2) = \frac{|Entity(E_1) \cap Entity(E_2)|}{|Entity(E_1) \cup Entity(E_2)|} \quad \text{(Eq. 3)}$$

$\mu(E_1, E_2)$, the score of event similarity, is a linear combination of $\mu_t(E_1, E_2)$ and $\mu_e(E_1, E_2)$.

$\mu(E_1, E_2) = \alpha_1 \times \mu_t(E_1, E_2) + (1 - \alpha_1) \times \mu_e(E_1, E_2)$ (Eq. 4)

4.3 Statistical Evidence for News Events

In this work, we introduce events as a middle-layer representation between words and sentences under the assumptions that 1) events are widely distributed in a text and that 2) they are natural clusters of salient information in a text. They guarantee the relevance of event to our task – summaries are condensed collections of salient information in source documents.

In order to confirm them, we scan the whole dataset in our experiment, which consists of 42 200w human extracts and 39 400w human extracts for the DUC 02 multi-document extract task. Detailed information about the dataset can be found in Section 6. Table 1 lists the statistics.

	200w	**400w**	**200w + 400w**	**Source Docs**
Entity/Sent	8.78	8.48	8.47	6.01
Entity/Word	0.34	0.33	0.33	0.30
Event/Sent	2.43	2.26	2.28	1.42

Event/Word	0.09	0.09	0.09	0.07
Sents with events/Sents	86.9%	85.1%	84.6%	71.3%

Table 1. Statistics from DUC 02 Dataset

There are on average 1.42 events per sentence in the source documents, and more than 70% of all the sentences contain events. The high event density confirms our first assumption about the distribution of events. For the 200w+400w category consisting of all the human-selected sentences, there are on average 2.28 events per sentence, a 60% increase from the same ratio in the source documents. The proportion of event-containing sentences reaches 84.6%, 13% higher than that in the source documents. Such is evidence that events count into the extract-worthiness of sentences, which confirms our second assumption about the relevance of events to summarization. The data also show higher entity density in the extracts than in the source documents. As entities are still reliable and domain-independent clues of salient content, we will consider both event and entity in the following ordering algorithm.

5 MDS Sentence Ordering with Event and Entity Coherence

In this section, we discuss how event can facilitate MDS sentence ordering with layered clustering on the event and sentence levels and then how event and entity information can be integrated in a coherence-based algorithm to order sentences based on sentence clusters.

5.1 Two-layered Clustering

After sentences are represented as collections of events, we need to vectorize events and sentences to facilitate clustering and cluster-based sentence ordering.

For a document set, event vectorization begins with aggregating all the event terms and entities in a set of **event units** (*eu*). Given m distinct event terms, n distinct named entities, and p distinct high-frequency common entities, the $m + n + p$ *eu*'s are a concatenation of the event terms and entities such that eu_i is an event term for $1 \le i \le m$, a named entity for $m + 1 \le i \le m + n$, and a high-frequency entity for $m + n + 1 \le i \le m + n + p$. The *eu*'s define the $m + n + p$ dimensions of an event vector in an eu-by-event matrix $\boldsymbol{E} = [e_{ij}]$, as shown in Figure 2.

$$\begin{array}{l} \\ eu_1 \\ \cdots \\ eu_m \\ \cdots \\ eu_{m+n} \\ \cdots \\ eu_{m+n+p} \end{array} \overbrace{\begin{bmatrix} e_{11} & \cdots & e_{1q} \\ \vdots & \ddots & \vdots \\ e_{m1} & \cdots & e_{mq} \\ \vdots & & \vdots \\ e_{m+n,1} & \cdots & e_{m+n,q} \\ \vdots & & \vdots \\ e_{m+n+p,1} & \cdots & e_{m+n+p,q} \end{bmatrix}}^{E_1, E_2, \ldots E_q}$$

Figure 2. eu-by-Event Matrix

We further define $Entity_N(E_j)$ and $Entity_H(E_j)$ to be the set of named entities and set of high-frequency entities of E_j. Then,

$$e_{ij} = \begin{cases} \mu_t(eu_i, Term(E_j)) & 1 \le i \le m \\ \dfrac{\sum_{e \in Entity_N(E_j)} v_n(eu_i, e)}{|Entity_N(E_j)|} & m + 1 \le i \le m + n \\ \dfrac{\sum_{e \in Entity_H(E_j)} v_h(eu_i, e)}{|Entity_H(E_j)|} & m + n + 1 \le i \le m + n + p \end{cases} \quad \text{(Eq. 5)}$$

$$v_n(w_1, w_2) = \begin{cases} 2 & w_1 \text{ is identical to } w_2 \\ 1 & w_1\ (w_2) \text{ is a part of } w_2\ (w_1) \text{ or they are in a hypernymy / holonymy relationship} \\ 0 & \text{otherwise} \end{cases} \quad \text{(Eq. 6)}$$

$$v_h(w_1, w_2) = \begin{cases} 1 & w_1 \text{ is identical to } w_2 \\ 0.5 & w_1 \text{ are } w_2 \text{ are synonyms} \\ 0 & \text{otherwise} \end{cases} \quad \text{(Eq. 7)}$$

In Eq. 5, $\mu_t(w_1, w_2)$ is defined as in Eq. 2. Both the entity-based $v_n(w_1, w_2)$ and $v_h(w_1, w_2)$ are measured in terms of total equivalence (identity) and partial equivalence. For named entities, partial equivalence applies to structural subsumption (e.g., "Britain" and "Great Britain") and hypernymy/holonymy (e.g., "South Africa" and "Zambia"). For common entities, it applies to synonymy (e.g., "security" and "safety"). Partial equivalence is considered because of the lexical variations frequently employed in journalist writing. The named entity scores are doubled because they represent the essential elements of a news story.

Since the events are represented as vectors, sentence vectorization based on events is not as straightforward as on entities or terms. In this work we propose a novel approach of **two-layered clustering** for the purpose. The basic idea is clustering events at the first layer and then using event clusters as a feature to vectorize and cluster sentences at the second

layer. Hard clustering of events, such as K-means, not only results in binary values in event vectors and data sparseness but also is inappropriate. For example, if EC_1 clusters events all with event terms similar to t^* and EC_2 clusters events all with event entity sets similar to e^* (a set), what about event $\{t^*, e^*\}$? Assigning it to either EC_1 or EC_2 is problematic as it is partially similar to both. So we decide to do soft clustering at the first layer.

A well-studied soft clustering technique is the Expectation-Maximization (EM) algorithm which iteratively estimates the unknown parameters in a probability mixture model. We assume a Gaussian mixture model for the q event vectors $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, …, $\boldsymbol{V}_q$, with hidden variables $\boldsymbol{H}_i$, initial means $\boldsymbol{M}_i$, priors $\boldsymbol{\pi}_i$, and covariance matrix $\boldsymbol{C}_i$. The E-step is to calculate the hidden variables $\boldsymbol{H}_i^t$ for each $\boldsymbol{V}_t$ and the M-step re-estimates the new priors $\boldsymbol{\pi}_i'$, means $\boldsymbol{M}_i'$, and covariance matrix $\boldsymbol{C}_i'$. We iterate the two steps until the log-likelihood converges within a threshold = 10^{-6}. The performance of the EM algorithm is sensitive to the initial means, which are pre-computed by a conventional K-means.

In a preliminary study, we found that the event vectors display pronounced sparseness. A solution to this problem in an effort to leverage the latent "event topics" among *eu*'s is the Latent Semantic Analysis (LSA, Landauer and Dumais, 1997) approach. We apply LSA-style dimensionality reduction to the eu-by-event matrix $\boldsymbol{E}$ by doing Singular Value Decomposition (SVD). A problem is with the number h of the largest singular values, which affects the performance of dimensionality reduction. In this work, we adopt a utility-based metric to find the best h^* by maximizing intra-cluster similarity (Φ_h) and minimizing inter-cluster similarity (Ψ_h) corresponding to the h-dimensionality reduction

$$h^* = \text{argmax}_h \ \Phi_h/\Psi_h \qquad \text{(Eq. 8)}$$

Φ_h is defined as the mean of average cluster similarities measured by cosine distance and Ψ_h is the mean of cluster centroid similarities. Because the EM clustering assigns a probability to every event vector, we also take those probabilities into account when calculating Φ_h and Ψ_h.

Based on the EM clustering of events, we vectorize a sentence by summing up the probabilities of its constituent event vectors over all event clusters (ECs) and obtaining an EC-by-sentence (S_n) matrix $\boldsymbol{S} = [s_{ij}]$.

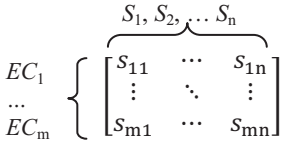

Figure 3. EC-by-Sentence Matrix

$s_{ij} = \sum_{E_r \in S_j} P(\overrightarrow{E_r}|EC_i))$ where $\overrightarrow{E_r}$ is E_r's vector.

At the sentence layer, hard clustering is sufficient because we need definitive, not probabilistic, membership information for the next step – sentence ordering. We use K-means for the purpose. The LSA-style dimensionality reduction is still in order as possible performance gain is expected from the discovery of latent EC "topics". The decision of the best dimensionality is the same as before, except that no probabilities are included.

5.2 Coherence-Based Sentence Ordering

Our ordering algorithm is based on sentence clusters, which is designed on the observation that human writers and summarizers organize sentences by blocks (paragraphs). Sentences within a block are conceptually close to each other and adjacent sentences cohere with each other. Local coherence is thus realized within blocks. On the other hand, blocks are not randomly ordered. Two blocks are put next to each other if their contents are close enough to ensure text-level coherence. So text-level, or global coherence is realized among blocks.

We believe in MDNS, the block-style organization is a sensible strategy taken by human extractors to sort sentences from different sources. Sentence clusters are simulations of such blocks and our ordering algorithm will be based on local coherence and global coherence described above.

First we have to pinpoint the leading sentence for an extract. Using the heuristic of time and textual precedence, we first generate a set of possible leading sentences $L = \{L_i\}$ as the intersection of the document-leading extract sentence set L_{Doc} and the time-leading sentence set L_{Time}. Note that $|L_{Doc}|$ = the number of documents, L_{Time} is in fact a sentence collection of time-leading documents, and $L_{Doc} \cap L_{Time} \neq \emptyset$.

If L is a singleton, finding the leading sentence S_L is trivial. If not, S_L is decided to be the sentence in L most similar to all the other sentences in the extract sentence set P so that it qualifies as a good topic sentence.

$$S_L = \text{argmax}_{L_i \in L} \sum_{L' \in P \setminus \{L_i\}} Sim_{\mu+\nu}(L_i, L') \quad \text{(Eq. 9)}$$

where $Sim_{\mu+\nu}(S_1, S_2)$ is the similarity between S_1 and S_2 in terms of their event similarity $\mu(S_1, S_2)$ and entity similarity $\nu(S_1, S_2)$. $\mu(S_1, S_2)$ is an extended version of $\mu(E_1, E_2)$ (Eq. 4) by averaging the $\mu_t(E_i, E_j)$ and $\mu_e(E_i, E_j)$ for all (E_i, E_j) pairs in $S_1 \times S_2$.

$$\mu(S_1, S_2) = \alpha_2 \times \frac{\sum_{E_i \in S_1, E_j \in S_2} \mu_t(E_i, E_j)}{|Event(S_1) \times Event(S_2)|} + (1 - \alpha_2) \times \frac{\sum_{E_i \in S_1, E_j \in S_2} \mu_e(E_i, E_j)}{|Event(S_1) \times Event(S_2)|} \quad \text{(Eq. 10)}$$

where $Event(S)$ is the set of all events in S. Next, $\nu(S_1, S_2)$ is the cosine similarity between their entity vectors $\overrightarrow{S_1}$ and $\overrightarrow{S_2}$ with entity weights constructed according to Eq. 6 and 7. Then,

$$Sim_{\mu+\nu}(S_1, S_2) = \alpha_3 \times \mu(S_1, S_2) + (1 - \alpha_3) \times \nu(S_1, S_2) \quad \text{(Eq. 11)}$$

After the leading sentence is determined, we identify the leading cluster it belongs to and our local coherence-based ordering starts with this cluster. We adopt a greedy algorithm, which selects each time from the unordered sentence set a sentence that best coheres with the sentence just selected, called **anchor sentence**.

Matching each candidate sentence with the anchor sentence only in terms of $Sim_{\mu+\nu}$ would assume that the sentences are isolated and decontextualized. But the anchor sentence did not come from nowhere and in order to find its best successor, we should also seek clues from its source context, which is inspired by the "sentence precedence" by Okazaki et al. (2004).

More formally, given an anchor sentence S_i at the end of the ordered sentence list, we select the next best sentence S_{i+1} according to their **associative similarity** and **substitutive similarity**, two crucial measures invented by us.

Associative similarity $Sim_{ASS}(S_i, S_j)$ measures how S_i and S_j associate with each other in terms of their event and entity coherence, which almost is $Sim_{\mu+\nu}(S_i, S_j)$. But to better capture the transition between entities and the flow of topic, we also consider a topic-continuity score $tc(S_i, S_j)$ according to the Centering Theory. If the topic continuity is measured in terms of entity change, local coherence can be captured by the centering transitions (*CB* and *CP*) in adjacent sentences. Based on (Taboada and Wiesemann, 2009), we assign 0.2 to the *Establish* and *Continue* transitions, 0.1 to *Smooth Shift* and *Retain*, and 0 to other centering transitions.

Since $tc(S_i, S_j)$ only applies to entities, it is treated as a bonus affiliated to $\nu(S_i, S_j)$.

$$Sim_{ASS}(S_i, S_j) = \alpha_4 \times \mu(S_i, S_j) + (1 - \alpha_4) \times \nu(S_i, S_j) \times (1 + tc(S_i, S_j)) \quad \text{(Eq. 12)}$$

Substitutive similarity accommodates what we earlier emphasized about the "source context" of the extracted sentences by measuring to what degree S_i and S_j resemble each other's relevant source context. More formally, let $LC(S_i)$ and $RC(S_i)$ be the left and right source contexts of S_i respectively, and the substitutive similarity $Sim_{SUB}(S_i, S_j)$ is defined as follows.

$$Sim_{SUB}(S_i, S_j) = Sim_{\mu+\nu}(S_i, LC(S_j)) + Sim_{\mu+\nu}(RC(S_i), S_j) \quad \text{(Eq. 13)}$$

In this work, we simply take $LC(S_i)$ and $RC(S_i)$ to be the left adjacent sentence and right adjacent sentence of S_i in the source document. Note that $tc(S_i, S_j)$ does not apply here. In view of the chronological order widely accepted in MDS ordering, a time penalty, $tp(S_i, S_j)$, is used to discount the score by 0.8 if S_i's document date is later than S_j's document date. Finally, Eq. 14 summarizes our intra-cluster ordering method in a sentence cluster SC_k.

$$S_{i+1} = \text{argmax}_{S_j \in SC_k \setminus \{S_i\}} \left(\alpha_5 \times Sim_{ASS}(S_i, S_j) + (1 - \alpha_5) \times Sim_{SUB}(S_i, S_j)\right) \times tp(S_i, S_j) \quad \text{(Eq. 14)}$$

After all the sentences in the current sentence cluster are ordered, we move on by considering the similarity of sentence clusters. Given a processed sentence cluster SC_i, the next best sentence cluster SC_{i+1} is the one that maximizes the cluster similarity $Sim_{CLU}(SC_i, SC_j)$ among the set of all clusters U. Since clusters are collections of sentences, their similarity is the mean of cross-cluster pairwise sentence similarities, each calculated according to Eq. 14. Eq. 15 shows how SC_{i+1} is computed.

$$SC_{i+1} = \text{argmax}_{SC_j \in U \setminus \{SC_i\}} Sim_{CLU}(SC_i, SC_j) \quad \text{(Eq. 15)}$$

This is how we incorporate (block-style) global coherence into MDS sentence ordering. Starting from the second chosen sentence cluster, we choose the first sentence in the current cluster with reference to the last sentence in the previous processed cluster and apply Eq. 14. We continue the whole process until all the extract sentences are ordered.

6 Evaluation

In this section, we report the experimental result on the DUC 02 dataset.

6.1 Data

We use the dataset of the DUC 02 summarization track for MDS because it includes an extraction task for which model extracts are provided. For every document set, 2 model extracts are provided each for the 200w and 400w length categories. We use 1 randomly chosen model extract per document set per length category as the gold standard.

We intended to use all the 59 document sets on DUC 02 but found that for some categories, both model extracts contain material from sections such as the *title*, *lead*, or even *byline*. Those extracts are incompatible with our design tailored for news body extracts. Therefore we have to filter them and retain only those extracts with all units selected from the news body. As a result, we collect 42 200w extracts and 39 400w extracts as our experimental dataset.

6.2 Peer Orderings

We evaluate the role played by various key elements in our approach, including event, topic continuity, time penalty, and LSA-style dimensionality reduction. In addition, we produce a random ordering and a baseline ordering according to chronological and textual order only. Table 2 lists the 9 peer orderings to be evaluated, with their codes.

A	Random
B	Baseline (time order + textual order)
C	Entity only (no LSA)
D	Event only (no LSA)
E	Entity + Event – topic continuity (no LSA)
F	Entity + Event – time penalty (no LSA)
G	Entity + Event (no LSA)
H	Entity + Event (event clustering LSA)
I	Entity + Event (event + sentence clustering LSA)

Table 2. Peer Orderings

6.3 Metrics

A popular metric used in sequence evaluation is Kendall's τ (Lapata, 2006), which measures ordering differences in terms of the number of adjacent sentence inversions necessary to convert a test ordering to the reference ordering.

$$\tau = 4m/(n(n-1)) \qquad \text{(Eq. 16)}$$

where m is the number of inversions described above and n is the total number of sentences.

The second metric we use is the Average Continuity (AC) developed by Bollegala et al. (2006), which captures the intuition that the ordering quality can be estimated by the number of correctly arranged continuous sentences.

$$AC = \exp\left(\frac{1}{k-1}\sum_{n=2}^{k}\log(P_n + \varepsilon)\right) \qquad \text{(Eq. 17)}$$

where k is the maximum number of continuous sentences, ε is a small value in case $P_n = 1$. P_n, the proportion of continuous sentences of length n in an ordering, is defined as $m/(N - n + 1)$ where m is the number of continuous sentences of length n in both the test and reference orderings and N is the total number of sentences. We set $k = 4$ and $\varepsilon = 0.01$.

6.4 Result

We empirically determine all the parameters (α_i) and produce all the peer orderings. Table 3 lists the result, where we also show the statistical significance between the full model peer ordering "I" and all other versions, marked by * ($p < .05$) and ** ($p < .01$) on a two-tailed t-test.

Peer Code	200w		400w	
	Kendall's τ	*AC*	Kendall's τ	*AC*
A	0.014**	0.009**	-0.019**	0.004**
B	0.387	0.151*	0.259**	0.151*
C	0.369*	0.128*	0.264*	0.156*
D	0.380	0.163	0.270*	0.158*
E	0.375*	0.156*	0.267*	0.157*
F	0.388	0.159*	0.264*	0.157*
G	0.385	0.158*	0.269*	0.162
H	0.384	0.164	0.292*	0.170
I	**0.395**	**0.170**	**0.350**	**0.176**

Table 3. Evaluation Result

Almost all versions with entity and event information outperform the baseline. The LSA-style dimensionality reduction proves effective for our task, as the full model (Peer I) ranks first and significantly beats versions without event information, topic continuity, or LSA. Applying LSA to both event and sentence clustering is better than applying it only to event clustering (Peer H), which produces unstable results and is sometimes outperformed by no-LSA versions (Peer G).

Event (Peer D) proves to be more valuable than entity (Peer C) as the event-only versions outperform the entity-only version in all categories, which is predicable because events

1) Thursday's **acquittals** in the McMartin Pre-School **molestation** case outraged parents who said prosecutors botched it, while those on the defense side proclaimed a triumph of justice over hysteria and hype.
2) Originally, there were seven defendants, including Raymond Buckey's sister, Peggy Ann Buckey, and Virginia McMartin, the founder of the school, mother of Mrs. Buckey and grandmother of Raymond Buckey.
3) Seven jurors who spoke with reporters in a joint news conference after **acquitting** Raymond Buckey and his mother, Peggy McMartin Buckey, on 52 **molestation** charges Thursday said they felt some children who testified may have been **molested** _ but not at the family-run McMartin Pre-School.
4) ``The children were never allowed to say in their own words what happened to them,'' said juror John Breese.
5) Ray Buckey and his mother, Peggy McMartin Buckey, were found not guilty Thursday of **molesting** children at the family-run McMartin Pre-School in Manhattan Beach, a verdict which brought to a close the longest and costliest criminal trial in history .
6) As it becomes apparent that McMartin cases will stretch out for years to come, parents and the former criminal defendants alike are trying to **resign** themselves to the inevitability that the matter may be one they can never leave behind.

Figure 4. Extract sentences of d80ae, 200w

are high-level constructs that incorporate most of the document-level important entities.

When entity is used, extra bonus can be gained from topic continuity concerns from CT (Peer E vs. Peer G) because the centering transition effectively captures the coherence pattern between adjacent sentences. The effect of the chronological order seems less clear (Peer F vs. P) as removing it hurts longer extracts rather than short extracts. Therefore chronological clues are more valuable for arranging more sentences from the same source document.

Our ordering algorithm achieves even better result with long extracts because the importance of order and coherence grows with text length. Measured by Kendall's τ, the full model ordering in the 400w category is significantly better than all other orderings.

For a qualitative evaluation, we select the 200w extract d080ae and list all the sentences in Figure 4. The event terms are boldfaced and the event entities are underlined.

Limited by space, let's focus on the baseline (1 2 3 4 5 6), entity-only (3 5 2 4 6 1), and full-model versions (3 5 4 2 1 6). The news extract is about the acquitting of child molesters. Both the "acquitting" and "molesting" events are found in 1) and 3) but only the latter qualifies as the topic sentence because it contains important event entities. Choosing 3) instead of 1) as the leading sentence shows the advantage of our event-enriched model over the baseline. The same choice is made by the entity-only version because 3) happens to be also entity-intensive. In order to see the advantage of the full model over the entity-only model, let's consider 2) and 4). 2) is chosen by the entity-only model after 5) because of the heavy entity overlap between 5) and 2). But semantically, 2) is not as close to 5) as 4) because only 4) contains entities for both the "acquitting" ("juror") and "molesting" ("children") events and intuitively, 4) continues the main trial-acquittal event topic but 2) supplies only secondary information. We examined the sentence clusters before the ordering and found that 3), 5), and 4) are clustered together only by the full model, leading to better coherence, locally and globally.

7 Conclusion and Future Work

We set out by realizing the semantic deficiency of IR and propose a low-cost approach of building event semantics into sentence representation. Event extraction relies on shallow parsing and external knowledge sources. Then we propose a novel approach of two-layered clustering to use event information, coupled with LSA-style dimensionality reduction. MDS sentence ordering is guided by local and global coherence to simulate the block-style writing and is realized by a greedy algorithm. The evaluation shows clear advantage of our event-enriched model over baseline and event-agonistic models, quantitatively and qualitatively.

The extraction approach can be refined by deep parsing and rich verb (frame) semantics. In a follow-up project, we will expand our dataset and experiment with more data and incorporate human evaluation in comparative tasks.

Acknowledgment

The work described in this paper was partially supported by a grant from the HK RGC (Project Number: PolyU5217/07E).

References

Barzilay, R., Elhadad, N., and McKeown, K. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Barzilay, R., and Lapata, M. 2005. Modeling Local Coherence: An Entity-based Approach. In *Proceedings of the 43rd Annual Meeting of the ACL*, 141-148. Ann Arbor.

Barzilay, R., and Lapata, M. 2008. Modeling Local Coherence: An Entity-Based Approach. *Computational Linguistics*, 34:1–34.

Bollegala, D, Okazaki, N., and Ishizuka, M. 2006. A Bottom-up Approach to Sentence Ordering for Multi-document Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 385–392. Sydney, Australia.

Bromberg, I. 2006. Ordering Sentences According to Topicality. Presented at the Midwest Computational Linguistics Colloquium.

Chklovski, T., and Pantel, P. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*. 11–13. Barcelona, Spain.

Conroy, J. M., Schlesinger, J. D., and Goldstein, J. 2006. CLASSY Tasked Based Summarization: Back to Basics. In *proceedings of the Document Understanding Conference (DUC-06)*.

Filatova, E., and Hatzivassiloglou, V. 2003. Domain-independent detection, extraction, and labeling of atomic events. In *Proceedings of RANLP*, 145–152, Borovetz, Bulgaria.

Filatova, E., and Hatzivassiloglou, V. 2004. Event-Based Extractive Summarization. In *Proceedings of the ACL-04*, 104–111.

Grosz, B. J., Aravind K. J., and Scott W. 1995. Centering: A framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.

Jurafsky D., and Martin, J. H. 2009. *Speech and Language Processing, Second Edition*. Upper Saddle River, NJ: Pearson Education International.

Landauer, T., and Dumais, S. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104.

Lapata, M. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. In *Proceedings of the Annual Meeting of ACL*, 545-552. Sapporo, Japan.

Li, W., Wu, M., Lu, Q., Xu, W., and Yuan, C. 2006. Extractive Summarization Using Inter- and Intra-Event Relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 369–376. Sydney.

Liu, M., Li, W., Wu, M., and Lu, Q. 2007. Extractive Summarization Based on Event Term Clustering. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, 185–188. Prague.

Okazaki, N., Matsuo, Y., and Ishizuka, M. 2004. Improving Chronological Ordering by Precedence Relation. In *Proceedings of 20th International Conference on Computational Linguistics (COLING 04)*, 750–756.

Taboada, M., and Wiesemann, L., Subjects and topics in conversation. *Journal of Pragmatics* (2009), doi:10.1016/j.pragma.2009.04.009.

Tan, Y. F., Kan, M., and Cui, H. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context*, 49–56, Trento, Italy.

Yoshioka, M., and Haraguchi, M. 2004. Multiple News Articles Summarization Based on Event Reference Information. In *Working Notes of NTCIR-4*, Tokyo.