Multi Event Extraction Guided by Global Constraints

Roi Reichart **Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{roiri, regina}@csail.mit.edu

Abstract

This paper addresses the extraction of event records from documents that describe multiple events. Specifically, we aim to identify the fields of information contained in a document and aggregate together those fields that describe the same event. To exploit the inherent connections between field extraction and event identification, we propose to model them jointly. Our model is novel in that it integrates information from separate sequential models, using global potentials that encourage the extracted event records to have desired properties. While the model contains high-order potentials, efficient approximate inference can be performed with dual-decomposition. We experiment with two data sets that consist of newspaper articles describing multiple terrorism events, and show that our model substantially outperforms traditional pipeline models.

1 Introduction

Today, most efforts in information extraction have focused on the field extraction task, commonly formulated as a sequence tagging problem. When a document describes a single event, the list of extracted fields provides a useful abstraction of the input document. In practice, however, a typical newspaper document describes multiple events, and a flat list of field values may not contain the sufficient structure required for many NLP applications. Our goal is therefore to extract event templates which aggregate field values for individual events.

Consider, for instance, the New York Times article excerpt in Figure 1 that describes three related terrorist events. As this example illustrates, in order to populate the corresponding event templates, the model needs to identify segments that describe individual events. Such segmentation is challenging, as event boundaries are not explicitly demarcated in the text. Moreover, descriptions of different events are often intermingled, as in the above example, further complicating boundary recovery.

In this paper, we consider a model that jointly performs event segmentation and field extraction. This model capitalizes on the inherent connection between the two tasks in order to reduce the ambiguity of template-based extraction. For example, the distribution of field values in the text provides strong clues about event segmentation, such as the presence of multiple new fields strongly signaling a segment boundary. Likewise, knowledge of the boundaries enables the model to rule out mutually inconsistent predictions, such as extracting two distinct locations for the same event.

We formulate our approach as a joint model that marks each word with field and event labels simultaneously. At the sentence level, segmentation and field extraction taggers are implemented using separate sequence models operating over local features. At the document level, the model encourages global consistency via potentials that link the extracted event records and their fields. Some of these potentials are limited to fields of an individual event such as the "single city per event" constraint. Others encode discourse-level properties of the whole document and thus involve records of multiple events,

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 70–79,
Montréal, Canada, June 3-8, 2012.

A powerful **car bomb exploded** today in **Baghdad** inside the holiest **Shiite shrine** . As many as **95 people** were killed in the event, according to sources in Washington. The **blast** came only two days after another **car bomb exploded** in a crowded **street** in **Mosul** in the northern part of **Iraq**, killing **13 pedestrians**, in an attack carried out by **Al Qaeda**. Together with the previous attack by **Al Qaeda**, the **shooting** in **Najaf** three weeks ago that killed **15 American soldiers**, violence seemed to spike to its highest level. The **bombing** today, happened around 9am, when the roads are crowded with people. ...

	Organization	Tactic	Target	Weapon	Fatalities	City	Country
Event 1	—	bombing	Shiite shrine	car bomb	95 people	Baghdad	—
Event 2	Al Qaeda	bombing	—	car bomb	13 pedestrians	Mosul	Iraq
Event 3	Al Qaeda	shooting	—-	—	15 American Soldiers	Najaf	—

Figure 1: A New York times article describing three terrorist events and a table demonstrating the corresponding event records.

such as the tendency in newspaper reporting to feature the main event at the beginning and repeatedly throughout the document.

While these high-order potentials encode important linguistic properties of valid assignments, they greatly complicate learning and inference. Therefore, our method estimates the parameters of the local sequence models and the global potentials separately. Then, at inference time, it finds variable assignments that are most consistent with both the local models and the global potentials. Inference is implemented via dual-decomposition, an efficient algorithm shown to be effective for complex joint inference problems.

We evaluate our approach for event extraction on two data sets, one is a new collection of long newspaper articles and the other is a subset of the MUC-4 documents. Both data sets consist of articles that describe multiple terrorist events (40.3 and 12.4 sentences and 4.4 and 3.1 events per article for each data set on average). We demonstrate the benefits of the joint model for event extraction; it outperforms a traditional pipeline model by a significant margin. For instance, it yields an absolute gain of 8.5% for our new corpus when measured using document-level F-score. Our results show the effectiveness of global constraints in the context of template extraction and motivate their exploration in other IE tasks.

2 Previous Work

Event-Template Extraction Event template extraction has been previously explored in the MUC-4 scenario template task. Work on this task has focused on pipeline models which decouple the task into the sub-tasks of field extraction and event-based text segmentation. For example, rule-based methods (Rau et al., 1992; Chinchor et al., 1993) identify generalizations both for single field fillers and for relations between fields and use them to fill event templates. Likewise, classifier-based algorithms (Chieu et al., 2003; Xiao et al., 2004; Maslennikov and Chua, 2007; Patwardhan and Riloff, 2009) generally train individual classifiers for each type of field and aggregate candidate fillers based on a sentential event classifier. Finally, unsupervised techniques (Chambers and Jurafsky, 2011) have combined clustering, semantic roles, and syntactic relations in order to both construct and fill event templates.

In our work, we also address the sub-tasks of field extraction and event segmentation individually; however, we link them through soft global constraints and encourage consistency through joint inference. To facilitate the joint inference, we use a linear-chain CRF for each sub-task.

Global Constraints Previous work demonstrated the benefits of applying declarative constraints in information extraction (Finkel et al., 2005; Roth and tau Yih, 2004; Chang et al., 2007; Druck and McCallum, 2010). Constraints have been explored both at sentence and document level. For example, Finkel et al. (2005) employ document-level constraints to encourage global consistency of named entity assignments. Likewise, Chang et al. (2007) use constraints at multiple levels, such as sentence-level constraints to specify field boundaries and global constraints to ensure relation-level consistency. In our work we focus on document-level constraints. We utilize both discourse and record-coherence constraints to encourage consistency between local sequence models.

There has also been unsupervised work that demonstrates the benefit of domain-specific constraints (Chen et al., 2011). In our work we show that domain-specific constraints based on the common structure of newspaper articles are also useful to guide a supervised model.

3 Model

Problem Formulation Given a document, our goal is to extract field values and aggregate them into event records. The training data consists of event annotations where each word in the document is tagged with a field and with an event id. If a word is not a filler for a field, it is annotated with a default NULL field value. At test time, the number of events is not given and has to be inferred from the data.

Model Structure Our model is built around the connection between local extraction decisions and global constraints on event structure. Based on local cues, the model can identify candidate field fillers. However, connecting them to events requires a broader document context. To effectively capture this context, the model needs to group together portions of the document that describe the same event. Global constraints are instrumental in this process, as they drive the aggregation of contiguous segments computed by a local segmentation model. In addition, global constraints coordinate local decisions and thereby enable us to express important discourse dependencies between various assignments.

To implement these ideas in a computational framework, we define an undirected graphical model with a vertex set $V = X \cup Y \cup Z$. X is a set of observed nodes; x_i represents the $i$the word in a document. Y and Z are sets of unobserved nodes corresponding to the field and event assignments respectively of the ith word. The number of input words in a document is denoted by n.

We define three types of potentials:

- *Field-labeling Potentials* associate words in a document with field labels based on their local sentential context.
- *Event-labeling Potentials* associate words in a document with event boundaries based on the local surroundings of a candidate boundary.
- *Global Consistency Potentials* link the extracted event records and their fields to encourage global consistency. These potentials are defined over the entire set of variables related to a document.

The resulting maximum aposteriori problem is:

$$MAP(\theta) = \sum_{f \in F} \theta_f(r_f)$$

where θ_f are the potential functions and $\{r_f | f \subseteq \{1, \ldots, n\}, f \in F\}$ is the set of their variables.

3.1 Modeling Local Dependencies

Field Labeling The first step of the model is tagging the words in the input document with fields. Following traditional approaches, we employ a linear-chain CRF (Lafferty et al., 2001) that operates over standard lexical, POS-based and syntactic features (Finkel et al., 2005; Finkel and Manning, 2009; Bellare and McCallum, 2009; Yao et al., 2010).

Event Segmentation At the local level, event analysis involves identification of event boundaries which we model as linear segmentation. To this end, we employ a binary CRF that predicts whether a given word starts a description of a new event or continues the description of the current event, based on lexical and POS-based features. In addition, we add features obtained from the output of the field extraction CRF. These features capture the intuition that boundary sentences often contain multiple fields.

The potential functions of these components are given by the likelihoods of the corresponding CRFs.

3.2 Modeling Global Dependencies

The main function of the global constraints is to link extracted fields to the corresponding events. In addition, the model can use global constraints to resolve potentially inconsistent decisions of the local models by encouraging them to agree with global, document-level properties. We consider two types of global consistency potentials: *discourse potentials* that involve interactions between multiple records, and *record coherence potentials* that capture patterns at the level of individual records.

The general form of a global potential p is:

$$\theta_f(x_{f-p}, y_{f-p}, z_{f-p}) = \begin{cases} \alpha_p & \text{if potential-property holds} \\ 0 & \text{otherwise} \end{cases}$$

Where $f - p$ is the index set of variables over which the potential is defined. Table 1 gives a formal description of all the potentials. Below we describe the linguistic intuition behind these potentials.

Discourse Potentials To populate event records with extracted information, the model needs to

Discourse	
MAIN EVENT	*Two consecutive sentences without fields indicate a transition to the main event:* $(\exists S_i, S_{i+1}$ s.t. $(\forall k \in S_i, y_k = NULL) \wedge (\forall k \in S_{i+1}, y_k = NULL)) \rightarrow$ $(\forall l \geq i$ s.t. $(\forall u, u \geq i, u < l, 1_{f_{ME}(S_u)=1}), \forall p \in S_l, z_p = CENTRAL)$
SEGMENT BOUNDARY	*Event changes should take place in multi-field sentences:* $\forall i, j \in I, ((i = j + 1) \wedge (z_i! = z_j)) \rightarrow$ $(\exists i_1 \ldots i_t \in I$ s.t. $1_{[f_{s-SB}(i,i_1,\ldots i_t)=1]} \wedge 1_{[f_{f-SB}(i_1,\ldots i_t)=1]})$
EVENT REDUNDANCY	*Events should not significantly overlap:* $\forall i, j \in \{1, \ldots, \lvert Z\rvert\}, \exists k, l \in I$ s.t. $((y_k = y_l) \wedge (y_k! = NULL) \wedge (z_k = i) \wedge (z_l = j) \wedge (x_k! = x_l))$
Record Coherence	
FIELD SPARSITY	*Some fields take a single unique value per record:* $\forall K, L \subset I, C \in \xi, ((Y_K = C) \wedge (Y_L = C) \wedge (Z_K = Z_L)) \rightarrow (X_K = X_L)$
RECORD DENSITY	*Words associated with a field should fill the field if it is otherwise empty:* $\forall i \in \zeta, C \in \xi, (\exists k \in I$ s.t. $(1_{[C_{ind}(x_k)=1]}) \wedge (z_k = i)) \rightarrow (\exists l \in I$ s.t. $(y_l = C) \wedge (z_l = i))$

Table 1: Logical formulations of the properties encouraged by the global potentials. S_i is the set of indexes corresponding the the ith sentence. $f_{ME}(S_u) = 1$ *iff* there is no event change in sentence S_u. $f_{s-SB}(i_1, \ldots, i_t) = 1$ *iff* the corresponding words appear in the same sentence. $f_{f-SB}(i_1, \ldots, i_t) = 1$ *iff* the corresponding words have different, non-NULL, field values. $C_{ind}(x_k) = 1$ *iff* x_k is assigned to C in a training event record. $CENTRAL$ is the central event of the document, defined to be its first event. $I = \{1, \ldots, n\}, \xi = \{1, \ldots |Y|\}, \zeta = \{1, \ldots, |Z|\}$.

group together sentences that describe the same event. The local boundary model can only predict contiguous blocks of event descriptions, but it cannot link together blocks that appear in different parts of the document. Our approach towards this task is informed by regularity in the discourse organization of news articles. A typical news story is devoted to a single event, mixed with short descriptions of other events. Therefore, we prefer event assignments where long segments with no field values – e.g., background descriptions – are associated with the main event. This intuition is formalized in the *Main Event Potential* shown in Table 1.

The second discourse constraint concerns detection of event boundaries. We prefer assignments in which the boundary sentence contains a large number of fields. This preference is expressed in the *Segment Boundary Potential* shown in Table 1.

The final discourse constraint favors assignments that reduce redundancy in generated records. It is unlikely that a document describes several events with significant factual overlap. This constraint is implemented in the *Event Redundancy Potential* shown in Table 1.

Record Coherence Potentials These potentials capture properties of valid field assignments in the context of a given event record. The first potential in this group — *Field Sparsity Potential* — is applied to fields, such as City, that tend to take a single unique value per event record.[1] This potential discourages assignments that link this field with multiple values within the same event. Similar constraints have been effectively used in information extraction in the past (Finkel et al., 2005). In our work, we apply this constraint at the event level, rather than at the document level, thereby enabling multiple variable values for multi-event documents.

The second record coherence potential — *Record Density Potential* — aims to reduce empty fields in the event record. This potential turns on when a local extractor fails to identify a filler for a field when processing a given event segment. If this segment contains words that are labeled as potential fillers in the context of other events in the training data, we prefer assignments that associate them with the field that otherwise would have been empty. This potential is inspired by the *one sense per discourse* constraint (Gale et al., 1992) that associates all the occurrences of the word in a document with the same semantic meaning.

[1]The potential is defined for the following fields: Terrorist Organization, Weapon, City, and Country.

4 Inference

Dual Decomposition The global potentials encode important document level information that links together the extracted event records and their fields. Introducing these potentials, however, greatly complicates inference. Consider the MAP equation of Section 3. If the intersection between each pair of subsets, $f_i, f_j \in F$, had been empty, we could have found the MAP assignment by solving each potential separately. However, since many subset pairs do overlap, we must enforce agreement among the assignments which results in an NP-hard problem.

In order to avoid this computational bottleneck we turn to dual-decomposition (Rush et al., 2010; Koo et al., 2010), an inference technique that enables efficient computation of a tight upper bound on the MAP objective, while preserving the original dependencies of the model. Dual decomposition has been recently applied to a joint model for biomedical entity and event extraction by Riedel and McCallum (2011). In their work, however, events are defined in the sentence level. Here we show how this technique can be applied to a model which involves document-level potentials.

We first re-write the MAP equation, such that it contains a local potential for each of the unobserved variables, as required by the inference algorithm:

$$MAP(\theta) = \max_{y,z} \sum_{j \in J} \theta_j(r_j) + \sum_{f \in F} \theta_f(r_f)$$

where we denote the set of indexes of all unobserved variables with J and refer to each of them with r_j. We then define the dual problem:

$$\min_{\delta} L(\delta),\ L(\delta) = \sum_{j \in J} \max_{r_j} [\theta_j(r_j) + \sum_{f:j \in f} \delta_{fj}(r_j)] + \sum_{f \in F} \max_{r_f} [\theta_f(r_f) - \sum_{j \in f} \delta_{fj}(r_j)]$$

where for every $f \in F$ and $j \in f$, δ_{fj} is a vector of Lagrange multipliers with an entry for each possible assignment of r_j. We add the notation δ_f for the matrix of Lagrange multipliers for all the variables in f, and for an assignment M of the variables in f we define $\delta_f(M)$ to be the corresponding vector of Lagrange multipliers. The multipliers can be viewed as messages transferred between the potentials to encourage agreement between their assignments.

The dual objective, $L(\delta)$, forms an upper bound on the MAP objective. Our inference algorithm therefore searches for its minimum, i.e. the tightest upper bound of the original MAP objective. $L(\delta)$ is convex and non-differentiable and can therefore be minimized by the subgradient descent algorithm in Figure 2 (a).

Set $g^0_{fj} \leftarrow 0$ for all $j \in J, f \in F$
for $k = 1$ to K **do**
  for $j \in J$ **do**
    $rl^k_j = \arg\max_{r_j} [\theta_j(r_j) + \sum_{f:j \in f} \delta_{fj}(r_j)]$
  $end \leftarrow TRUE$
  for $f \in F$ **do**
    $rp^k_f = \arg\max_{r_f} [\theta_f(r_f) - \sum_{j \in f} \delta_{fj}(r_j)]$
    for $j \in f$ **do**
      if $rl^k_j \neq rp^k_{fj}$ **then**
        $g^k_{fj}(rl^k_j) \ += 1$
        $g^k_{fj}(rp^k_{fj}) \ -= 1$
        $end \leftarrow FALSE$
      $\delta^{k+1}_{fj} = \delta^k_{fj} - \beta_k \cdot g^k_{fj}$
  if end **then**
    return R^k
  $\beta_k \leftarrow 1/k$
return (R^K)

(a)

rl^k_j: Sort $[\theta_j(r_j) + \sum_{f:j \in f} \delta_{fj}(r_j)]$. Return the minimizing r_j.
rp^k_f:
$MMA^k_f \leftarrow$: Minimum-Message assignment
$PRA^k_f \leftarrow$: Property-Respecting assignment
if $(\alpha_p - sum(\delta_f(PRA)) > (-1) * sum(\delta_f(MMA))$ **then**
  $rp^k_f = PRA^k_f$
else
  $rp^k_f = MMA^k_f$

(b)

Figure 2: The inference algorithm. (a): The dual-decomposition algorithm. (b): Algorithms for the $\arg\max$ operations of the dual-decomposition algorithm.

Individual Potentials Maximization The inference algorithm requires efficient solvers for its $\arg\max$ problems. For the field labeling and event segmentation potentials, the messages are encoded into the feature space of the CRF, and exact maximization is achieved through standard CRF decoding. For the local potentials, (rl^k_j), the maximizing assignments are computed by sorting the messages for each unobserved variable (Figure 2 (b)).

The global potentials are more challenging. Ideally, we could find the optimal assignment, rp^*_f, that agrees with the assignments of the other potentials ($rp^*_f = argmin \sum_{j \in f} \delta_{fj}(rp_j)$) and at the same time respects the property encouraged by its own po-

tential ($\theta_p(rp_f^*) > 0$). In practice, however, there may be no such assignment, in which case the assignment conflict needs to be resolved.

We first compute the *minimum-message assignment (MMA)*, the assignment that minimizes the message sum. If this assignment respects the potential property then it is the optimal assignment. Otherwise, we compute the *property-respecting assignment (PRA)*, the assignment with the (approximate) lowest message sum under the condition that the potential property holds. From these two assignments we select the one with the higher score.

Finding the MMA is simple, as it is the minimum-message assignment of each unobserved variable separately. However, finding the global optimal PRA is computationally demanding, as it requires searching over a very large assignment space. We therefore trade accuracy for efficiency and restrict each potential to modify the MMA assignment for only one type of variables: Y (fields) or Z (events). The discourse potentials and the FIELD SPARSITY potential are restricted to changes of the event variables, while the RECORD DENSITY potential is restricted to changes of the field variables.

For the MAIN EVENT potential, consecutive sentences with no fields trigger a return to the main event. For the SEGMENT BOUNDARY potential, event changes that take place in sentences with a small number of fields are removed. For our work, this threshold is set to three. For the EVENT REDUNDANCY potential, redundant events are integrated with the largest event in which they are contained. For the RECORD DENSITY potential, words seen in both training records and event text are used to fill empty fields. For each empty field in each event, words labeled with event are scanned for candidate fillers, and those with the minimal impact on the message sum are assigned to that field.

Finally, for the FIELD SPARSITY potential, if a field contains more than one word or phrase per event, the event assignments of these words or phrases are recomputed. This computation is implemented as a minimum matching problem in a bipartite graph. One side of the graph consists of a vertex for every word or phrase assigned to the addressed field, and the other side consists of one vertex for each event in the document. If the number of phrases assigned to the field is larger than the number of events in the document, some of the event vertices will be assigned to new events. The edge weights are the sum of message changes corresponding to relabeling the word or phrase with the new event. We solve this problem efficiently ($O(n^3)$) using the Kuhn-Munkres algorithm (Kuhn, 1955).

5 Experiments

Data This work focuses on multi-event extraction. While some of the articles in the MUC test corpus do have multiple events, the majority contain only one (77.5%) or two (12%). We therefore created two corpora for our experiments. The first is a new corpus of 70 articles from New York Times (NYT) LDC corpus, each describing one or more terrorist events from various parts of the world. The second, also of 70 articles, consists of a subset of the MUC articles that describe more than one event. We stripped this corpus from the MUC annotation and annotated it according to our scheme.

Annotations were provided by two annotators with graduate school educations. Every word was tagged with a field and an event id. The 8 fields we use are: Terrorist Organization, Target, Tactic, Weapon, Fatalities, Injuries, Country and City.

We compared the agreement between annotators on 10 articles by computing the percentage of words for which the annotators gave the same labeling. The inter-annotator agreement was 90.9% (kappa = 0.9) when fields and events are evaluated together (i.e., the annotators are considered to agree only when they assign the same field and event id to the word), 97.8% (kappa = 0.97) for events only, and 92% (kappa = 0.91) for fields only.

The two corpora differ from each other with respect to several important properties. The New-York Times articles are longer (40.3 compared to 12.4 sentences per article) and describe a larger number of events (4.4 compared to 3.1 events per article on average). In addition, while our hypothesis about the predominance of the main (first) event coverage holds for both corpora, it better characterizes the New-York Times corpus, as is demonstrated by the following two statistics.

First, in the NYT corpus the average number of sentences containing field fillers for the main event is 14.7, while for any other event the average number

is 3.2. In the MUC corpus the corresponding numbers are 5.3 and 2.0. Second, in the NYT corpus the number of times an article goes back to a previously described event is 182 (average of 2.6 times per article), of which 154 (84.6%) are transitions to the main event. In the MUC corpus the number of times an article goes back to a previously described event is only 38 (average of 0.54 times per article), but, similarly to the NYT, in as much as 32 (84.2%) of these cases the transitions are to the main event.

Experimental Setup For both corpora, we used 30 articles for training (1218 sentences in NYT, 423 in MUC), 7 articles for development (358 sentences in NYT, 79 in MUC) and 33 articles for test (1244 sentences in NYT, 367 in MUC). The sentences were POS tagged with the MXPOST tagger (Ratnaparkhi, 1996) and parsed with the Charniak parser (Charniak and Johnson, 2005).

We trained our model with a two steps procedure. First, the local CRFs were separately trained on the training articles. Then, we trained the parameters of the global potentials using the structured perceptron algorithm (Collins, 2002) on the development data.

We perform joint inference over the local CRFs as well as the global potentials with dual decomposition. This algorithm is guaranteed to give the MAP assignment if it converges to a solution in which all the potentials agree on the label assignment for the variables in their scope. To deal with disagreements, we ran the algorithm for 200 iterations past the point of fluctuations around the dual minimum. The final label assignment is determined by a majority vote between the potentials in the 10 iterations with the highest total inter potential agreement (Sontag et al., 2010).

Baselines We compare our algorithm to two baseline models. The first baseline is related to previous techniques that decompose the task into field extraction and event segmentation sub-tasks (Jean-Louis et al., 2011; Patwardhan and Riloff, 2007; Patwardhan and Riloff, 2009). For this PIPELINE baseline, we run the CRF models described in Section 3.1, first the field CRF and then the event CRF. The field-based features of the event CRF are extracted from the output of the field CRF.

Our model incorporates global dependencies into a document level model. An alternative approach is to encode this information as local features that reflect global dependencies (Liang et al., 2008). We therefore constructed a second baseline, the bidirectional pipeline model (BI-PIPELINE), that considers global features which encode similar properties to those encouraged by our global potentials. We implement this by incorporating event-based features into the feature set of the field labeling CRF, while kipping the event segmentation CRF fixed. [2] As in the pipeline model, each CRF is trained separately on the training data. The BI-PIPELINE model, however, emulates our joint inference procedure by iteratively running a field labeling and an event segmentation CRFs. The number of iterations for this model was estimated on development data.

Evaluation Measures We follow the MUC-4 scoring guidelines (Chinchor, 1992). To compare between a learned and a gold standard event, we compute the word-level F-score between each of their fields and average the results. If a field is empty in both event records, it is not counted in the mutual event score, while if it is empty in only one of the event records, its F-score is 0.

Ideally, the measure should be able to capture paraphrases. For example, if the *T*actic field in a gold event record contains the words "bombing" and "blast", the measure is expected to give a perfect score to a learned record that contains one of these words. Therefore, as in the MUC-4 guidelines, we count pre-specified synonyms and morphological derivations of the same word only once.

For every document, we then map the learned events to the gold events in a greedy 1-1 manner using the Kuhn-Munkres algorithm (Kuhn, 1955). Once we have an event mapping, we can report an average recall, precision and F-score across the test set for all fields, events and documents (where the document F-score is the average F-score of its events). We use the sign test to measure the statistical significance for our results. Since the number of events described in a document is not given to the models as input, we also report the average ratio between the number of induced and gold events.

[2] Example additional features are: (1) whether a word with the same most frequent field (MFF) as the encoded word previously appeared in its event; (2) whether a new event is started in the sentence of the encoded word; and (3) whether the event of the encoded word contains at least one word annotated with the MFF of the encoded word.

NYT	Documents			Events			Fields			Event Number
	R	P	F	R	P	F	R	P	F	Ratio
Joint Model	**38.7**	**42.4**	**38.5**	**36.2**	**40.8**	**36.4**	**43.6**	**49.1**	**43.8**	0.95
Bi-pipeline Model	33.3	30.8	30.2	31.9	30.1	29.4	38.8	36.6	35.7	1.14
Pipeline Model	28.3	27.0	26.2	27.1	26.8	25.5	35.4	34.8	33.2	1.5

MUC	Documents			Events			Fields			Event Number
	R	P	F	R	P	F	R	P	F	Ratio
Joint Model	**49.8**	**43.2**	**43.5**	**48.7**	**43.0**	**42.7**	**53.6**	**45.9**	**46.2**	0.88
Bi-pipeline Model	38.1	38.6	36.3	34.3	33.9	32.2	41.5	40.5	38.6	0.92
Pipeline Model	30.8	32.8	29.7	29.9	32.0	28.9	37.9	40.1	36.6	0.89

Table 2: Performance of the joint model and the pipeline models on the event record extraction task. Top table is for the New-York Times data. Bottom table is for the MUC data. All results are statistically significant with $p < 0.05$.

NYT	TO	TAR	TAC	WEAP	INJ	FAT	CO	CITY
Joint Model	**21.9**	**23.4**	**49.0**	**39.6**	**40.8**	**49.1**	**43.1**	46.6
Bi-pipeline Model	8.4	19.7	47.5	20.9	25.9	18.3	38.8	38.1
Pipeline Model	7.1	18.1	41.9	36.9	19.1	16.5	38.0	**46.1**

MUC	TO	TAR	TAC	WEAP	INJ	FAT	CO	CITY
Joint Model	**49.0**	**25.2**	**63.6**	**62.0**	**43.3**	21.1	19.7	**38.3**
Bi-pipeline Model	28.0	24.7	38.2	55.8	42.7	**25.6**	**37.5**	37.2
Pipeline Model	34.9	23.4	50.3	56.5	10.4	12.4	30.0	32.0

Table 3: Comparison between the joint model and the pipeline models for the different fields. When the joint model is superior results are statistically significance with $p < 0.05$.

(a)

NYT	Fields			Events	
	R	P	F	GF	LF
Joint model	**47.3**	51.3	**49.2**	**54.8**	**61.3**
Bi-Pipeline	31.0	43.8	36.3	48.8	56.2
Pipeline Model	39.2	**55.4**	45.9	51.3	52.9

(b)

MUC	Fields			Events	
	R	P	F	GF	LF
Joint model	47.3	**51.3**	**49.2**	62.8	70.0
Bi-Pipeline	**49.5**	36.1	41.8	62.2	62.0
Pipeline Model	31.0	43.8	36.3	**65.5**	**70.3**

Table 4: Performance of the joint and the pipeline models on the labeling tasks of assigning words to fields (left) and to events (right). Field values are computed for words tagged with the non-NULL field. Events values are computed for words that are assigned to a non-NULL field by the gold standard (GF) or by the model (LF). When the joint model is superior, results for fields are statistically significant with $p < 0.01$ and for events with $p < 0.05$.

6 Results

Event-Records Results for event record extraction, the main task addressed in this paper, are presented in Table 2. For all measures, the model outperforms the pipeline baselines, with an F-score difference of up to 13.8%.

The rightmost column of the table demonstrates the tendency of our model to under-segment. For both corpora our model extracts a smaller number of events than the gold standard on average (5% for NYT, 12% for MUC). The pipeline baselines extract more events than our model on average. For NYT they over-segment (14% for bi-pipeline, 53% for the pipeline) while for MUC they under-segment (8% and 11% respectively). These differences are expected as the baselines cannot combine different text segments that describe the same event.

Table 3 presents per-field F-score performance. The joint model outperforms the pipeline baselines for 7 out of the 8 fields in the NYT experiments, and for 6 out of 8 fields in the MUC experiments.

Model Components Table 6 presents the performance of variants of the joint model created by excluding each potential type. The results demonstrate the significance of both discourse and record coherence potentials for the performance of the full model.

Sub-tasks Performance A model for our task

(a)

NYT	Gold Fields				Gold Events		
	Doc.	Events	Fields	Ratio	Doc.	Events	Fields
Joint Model	**69.1**	**62.5**	**64.4**	1.05	**45.7**	**46.5**	**50.0**
Bi-Pipeline	—	—	—	—	41.7	40.8	46.1
Pipeline	47.9	43.9	51.3	1.56	40.8	40.4	43.9

(b)

MUC	Gold Fields				Gold Events		
	Doc.	Events	Fields	Ratio	Doc.	Events	Fields
Joint model	**78.5**	**75.0**	**74.5**	0.76	**50.8**	**47.9**	**51.4**
Bi-Pipeline	—	—	—	—	37.0	34.3	39.9
Pipeline	76.1	71.1	72.0	0.78	32.6	31.2	36.0

Table 5: Performance of the joint model and the pipeline models when the gold standard for one of the labeling tasks is given at test time. Results are statistically significant with $p < 0.05$.

	NYT			
Excluded Component	Documents	Events	Fields	Event Rat.
Record Coherence	32.1	31.0	37.7	1.04
Discourse	26.7	26.3	34.3	1.5
	MUC			
Record Coherence	37.4	33.6	39.6	0.88
Discourse	37.7	36.6	42.7	0.89

Table 6: The effect of the record coherence potentials and of the discourse potentials on the performance of the joint model. Results are presented for F-scores, each line is for the full model when potentials of one type are excluded.

should determine both when a word is a good field filler and to which event the field belongs. Since our main evaluation collapses the effect of these decisions together, we performed two additional sets of experiments to analyze the model's accuracy on each sub-task separately.

Figure 4 presents the performance of the different models on the labeling tasks of assigning words to fields and to events. The number of words associated with a field differs between the gold standard and the models' output. For fields, we therefore report word level recall, precision and F-score between the set of words assigned a non-NULL field by a model and the corresponding gold standard set. For events, we compute the fraction of words assigned the correct event among the words assigned to a non-NULL field in either the gold standard or the output of the model.

Figure 5 presents the document F-score when the gold-standard fields (left) or events (right) of the test set are known at test time. Note that when the gold standard fields are known, the BI-PIPELINE model is not applicable anymore since it is designed to improve field assignment using event-informed features. The results demonstrate that encoding field information to the models is more valuable than encoding information about events. This provides us with an important direction for future improvement of our model.

Accuracy and Efficiency When we ran our algorithm on the joint task of the NYT data-set it converged after 89 iterations. For the MUC joint task and the ablation analysis experiments we ran the algorithm for 200 iterations past the point of fluctuations around the dual minimum.

On a 2GHz CPU, 2GB RAM machine, it took our dual-decomposition algorithm 15 minutes and 10 seconds to complete its run on the entire NYT test set. For the MUC joint task experiment, in the 10 iterations considered for the majority vote, there is full agreement between the potentials for 97.77% of the unobserved variables. That is, the voting scheme affects the assignment of only 2.23% of the unobserved variables.

7 Conclusions

In this paper we presented a joint model for identifying fields of information and aggregating them into event records. We experimented with two data sets of newspaper articles containing multiple event descriptions. Our results demonstrate the importance and effectiveness of global constraints for event record extraction.

Acknowledgements

The authors gratefully acknowledge the support of the DARPA Machine Reading Program under AFRL prime contract no. FA8750-09-C0172. Any opinions, findings and conclusions expressed in the material are those of the author(s) and do not necessarily reflect the views of DARPA, AFRL or the US government. Thanks also to the members of the MIT NLP group and to Amir Globerson for their suggestions and comments.

References

Kedar Bellare and Andrew McCallum. 2009. Generalized expectation criteria for bootstrapping extractors using record-text alignment. In *EMNLP*.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based information extraction without the templates. In *ACL*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint driven learning. In *ACL*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL*.

Harr Chen, Edward Benson, Tahira Naseem, and Regina Barzilay. 2011. In-domain relation discovery with meta-constraints via posterior regularization. In *ACL*.

Hai Leong Chieu, Hwee Tou Ng, and Yoong Keok Lee. 2003. Closing the gap: Learning-based information extraction rivaling knowledge-engineering methods. In *ACL*.

Nancy Chinchor, David Lewis, and Lynette Hirschman. 1993. Evaluating message understanding systems: an analysis of the third message understanding conference. *Computational Linguistics*, 19(3):409–449.

Nancy Chinchor. 1992. Muc-4 evaluation metrics. In *Fourth Message Understanding Conference (MUC-4)*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *EMNLP*.

Gregory Druck and Andrew McCallum. 2010. High-performance semi-supervised learning using discriminatively constrained generative models. In *ICML*.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *NAACL*.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*.

William Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*.

Ludovic Jean-Louis, Romaric Besancon, and Olivier Ferret. 2011. Text segmentation and graph-based methods for template filling in information extraction. In *IJCNLP*.

Terry Koo, Alexander M. Rush, Michael Collins, Tommi Jaakkola, and David Sontag. 2010. Dual decomposition for parsing with non-projective head automata. In *EMNLP*.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Percy Liang, Hal Daume, and Dan Klein. 2008. Structure compilation: trading structure for features. In *ICML*.

Mstislav Maslennikov and Tat-Seng Chua. 2007. A multi-resolution framework for information extraction from free text. In *ACL*.

Siddharth Patwardhan and Ellen Riloff. 2007. Effective ie with semantic affinity patterns and relevant regions. In *EMNLP*.

Siddharth Patwardhan and Ellen Riloff. 2009. A unified model of phrasal and sentential evidence for information extraction. In *EMNLP*.

Adwait Ratnaparkhi. 1996. A maximum entropy part-of-speech tagger. In *WVLC*.

Lisa Rau, George Krupka, Paul Jacobs, Ira Sider, and Lois Childs. 1992. Muc-4 test results and analysis. In *Fourth Message Understanding Conference (MUC-4)*.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *EMNLP*.

Dan Roth and Wen tau Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *CoNLL*.

Alexander M. Rush, David Sontag, Michael Collins, and Tommi Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *EMNLP*.

David Sontag, Amir Globerson, and Tommi Jaakkola. 2010. Introduction to dual decomposition for inference. In *Optimization for Machine Learning, editors S. Sra, S. Nowozin, and S. J. Wright: MIT Press*.

Jing Xiao, Tat-Seng Chua, and Hang Cui. 2004. Cascading use of soft and hard matching pattern rules for weakly supervised information extraction. In *COLING*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2010. Collective cross-document relation extraction-without labelled data. In *EMNLP*.