# Joint Inference for Cross-document Information Extraction

Qi Li, Sam Anzaroot, Wen-Pin Lin, Xiang Li, Heng Ji
Computer Science Department, Queens College and Graduate Center
City University of New York
New York City, NY, USA
{liqiearth, samanzaroot, danniellin, jackieiuu729, hengjicuny}@gmail.com

## ABSTRACT

Previous information extraction (IE) systems are typically organized as a pipeline architecture of separated stages which make independent local decisions. When the data grows beyond some certain size, the extracted facts become inter-dependent and thus we can take advantage of information redundancy to conduct reasoning across documents and improve the performance of IE. We describe a joint inference approach based on information network structure to conduct cross-fact reasoning with an integer linear programming framework. Without using any additional labeled data this new method obtained 13.7%-24.4% user browsing cost reduction over a state-of-the-art IE system which extracts various types of facts independently.

## Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis; H.3.4 [**Systems and Software**]: Information networks, Question-answering (fact retrieval) systems

## General Terms

Algorithms

## Keywords

Information extraction, global reasoning, Integer Linear Programming

## 1. INTRODUCTION

One of the initial goals for Information Extraction (IE) was to create a knowledge base from the entire input corpus, such as a profile or a series of activities about any entity, and allow further logical reasoning on the knowledge base. Unfortunately the knowledge base constructed from a typical IE pipeline often contains lots of erroneous and conflicting facts. Interestingly, when the data grows beyond some certain size, the extracted facts become inter-dependent and thus we can take advantage of information redundancy to conduct reasoning across documents and improve the performance of IE. Some recent work conducted inference to achieve extraction consistency across documents [4], entities [3], relations [11, 6] and events [7].

In this paper we propose to apply the new structure called "Information Networks" to conduct more complete and robust inference. An Information Network (InfoNet) is a heterogeneous network that includes a set of "information graphs" $G = G_i(V_i, E_i)$, where $V_i$ is the collection of entity nodes, and $E_i$ is the collection of edges linking one entity to the other, labeled by relation or event attributes, such as "Member_of" and "Family". This structured networked representation characterizes a dense graph of relations/events among all entities provides a strong graph theoretical framework to enable effective inference and propagation. From the InfoNet point of view, it's clear that most previous inference work addressed partial structures, such as the inferences between repeated nodes/links [4], a pair of nodes [3], or a pair of links [7]. The goal of this paper is to develop a uniformed global inference framework across all of the nodes and links in the entire InfoNet.

When a user distills knowledge from the IE output, coherent facts are preferred because they include fewer conceptual gaps and thus require fewer inferences and less prior knowledge to understand [9]. Based on this intuition, we propose the following hypothesis: if an extracted fact is more consistent with other facts to tell a *coherent* story, it's more likely to be correct. For example, Figure 1 depicts a partial InfoNet connecting inter-dependent relations. The solid lines present correct relations, while dash lines indicate incorrect relations. From this figure, we can observe that contradictions may happen due to the relational dependencies, for instance, *George W. Bush* is detected as the member of both *Republican Party* and *Hamas*, while these two organizations are located in different countries (*United States of America* vs. *Hamas*). Based on one possible global constraint that an organization and its members are unlikely to locate in different countries, we can determine that *George W. Bush* is unlikely to be a member of *Hamas*.

We propose a joint inference method based on cross-fact constraints in InfoNets to approach this goal. We gather together IE results from a large collection of documents, and then impose constraints in an integer linear programming framework to reach global optimization. The inference knowledge includes constraints among all kinds of IE stages. To demonstrate the power and generality of this approach, we evaluate it for two distinct relation types: membership and family relations. The effectiveness of this framework is demonstrated by substantially improving the performance of both relation types.

## 2. RELATED WORK

Some recent IE research has raised much interest in global inference [8, 4, 1, 2, 12, 7, 10, 3] based on integer linear programming,


*CIKM'11*, October 24–28, 2011, Glasgow, Scotland, UK.

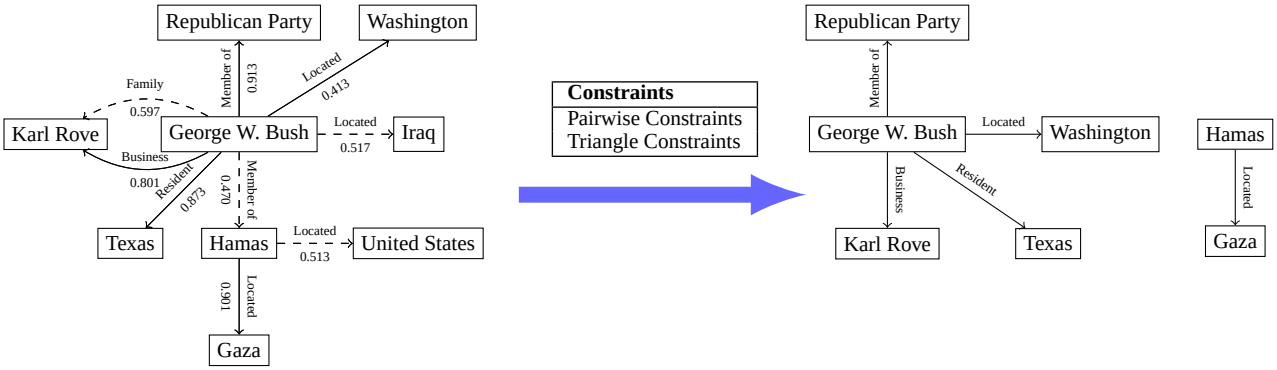


**Figure 1: Inference on a partial graph connecting inter-dependent relations**

heuristic rules or Markov Logic. In this paper, we extend the idea of global inference to the cross-document paradigm for enhancing IE performance. We also leverage relational constraints among all kinds of facts on diverse levels through a new information network representation. In addition, we have generalized the constraints to various categories, so that in the future one only needs to materialize these categories to new types of facts.

## 3. TASK AND BASELINE SYSTEM

We define a new cross-document IE task on top of the terminologies specified in the NIST ACE program. ACE defined 7 types of entities (persons, geo-political entities, locations, organizations, facilities, vehicles and weapons), 18 types of relations (e.g., "a town some 50 miles south of Salzburg" indicates a "Located" relation.); and 33 distinct types of relatively dynamic events (e.g., "Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment." indicates a "Personnel-start" event). We extend the ACE terminology from single document to cross-document setting as follows: Given a collection of source documents, our cross-document IE system should produce a networked knowledge base including unique (i.e. redundancy must be removed) facts.

We apply a state-of-the-art English ACE single-document IE system [4] as our baseline to extract facts from individual documents. This IE system includes Hidden Markov Model based name tagging, and Maximum Entropy based nominal mention tagging, entity coreference resolution, time expression extraction and normalization, relation extraction and event extraction , incorporating diverse lexical, syntactic, semantic and ontological knowledge. Finally a cross-document entity linking component based on graph clustering is applied to aggregate facts from different documents. Each component produces reliable confidence values.

## 4. GLOBAL INFERENCE

In this section, we will discuss the dependency and constraints of the IE outputs in detail, then present a novel cross-document global reasoning approach to enhance IE performance, based on Integer Linear Programming.

### 4.1 Dependency Constraints

We explore typical constraints to be defined across various types of facts. Usually there are heterogeneous relations and events existing among entities, as well as complex interaction patterns. We summarize the constraints on a more abstract level so that one can design domain-specific constraints with the map. Let $L_i$ denote a unique relation or event linking two entities $A$ and $B$. In this paper we consider pairwise and triangle dependency relations among various types of entities and link types, as depicted in Figure 2.

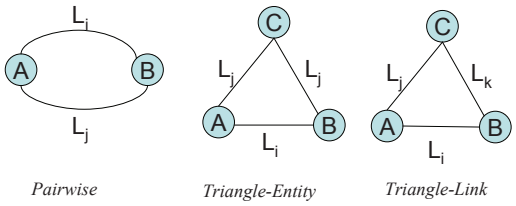


**Figure 2: Dependency Constraints among Entities, Relations and Events**

We compute pointwise mutual information (PMI) to automatically estimate the pairwise dependency between any two types of links from ACE2005 training data:

$$PMI(L_i, L_j) = log\frac{p(L_i, L_j)}{p(L_i)p(L_j)} \quad (1)$$

where $p(L_i)$ and $p(L_j)$ are the frequency of $L_i$ and $L_j$ respectively, and $p(L_i, L_j)$ is the co-occurence frequency of $L_i(A, B)$ and $L_j(A, B)$, for any two entities $A$ and $B$.

Similarly, we apply a multivariate generalization form of PMI [13], total correlation, to the triangle dependency:

$$PMI(L_i, L_j, L_k) = log\frac{p(L_i, L_j, L_k)}{p(L_i)p(L_j)p(L_k)} \quad (2)$$

for any two entities $A$ and $B$ or three entities $A$, $B$ and $C$ as shown in Figure 2.

If the PMI value is lower than a certain threshold (in our experiment we used -2.0 for pairwise and -3.0 for triangle), the links are considered as incompatible and used as a constraint for global inference.

In total we learned 34 pairwise constraints and 16 triangle constraints. Some constraint examples are listed in Table 1. For example, *Ariel Sharon* and *Mahmoud Abbas* are frequently involved in Contact.Meet events, so they are unlikely to be members of the same organization according to the pairwise constraint. If *Osama bin Laden* and *George W. Bush* are involved in a Conflict event with high confidence, then they are unlikely to be the members of the same organization according to the triangle-entity constraint. If *Washington* and *Iraq* are involved in a *Transport* event, then any member of *Washington* is unlikely to be located in *Iraq* according to the triangle-link constraint.

| | | | |
|---|---|---|---|
| Pairwise | $L_i$ | $L_j$ | |
| | Person A founded Organization B | Organization B hired Person A | |
| | Person A has a Business relation with Person B | Person A has a Person-Social relation (e.g. family) with Person B | |
| Triangle-Entity | $L_i$ | $L_j$ | |
| | Organization A is involved in a Justice/Conflict/Transaction event with Organization B | Person C is affiliated with/member of both Organization A and Organization B | |
| Triangle-Link | $L_i$ | $L_j$ | $L_k$ |
| | Entity A is involved in a Transport event originated from Location B | Person C is affiliated with/member of Entity A | Person C is located in Location B |

**Table 1: Examples for Incompatible Constraints**

## 4.2 Integer Linear Programming

Motivated by the intuition that facts are often dependent on each other, we consider the global inference as an optimization problem of all local predication confidence values, using the constraints automatically learned from the above PMI method as hard rules. These hard rules or constraints are designed to guarantee that the facts extracted from different documents are consistent with each other, hence weak predications which violate constraints can be filtered out.

First of all, we shall introduce the objective of our optimization problem. Assume we have a set of local outputs, $\mathcal{R} = \{r_i\}$, of inter-dependent relations and events, each output $r_i$ is associated with a number of entities $r_{i,j}$ with local confidence values $p_{i,j}$, where $p_{i,j} \in (0, 1]$.

From cross-document point of view, a reliable output should have high local confidence value as well as high global frequency. A correct fact may appear rarely within a single document but often appears frequently across documents. In contrast, an invalid fact often has low-frequency, simply because some entities accidentally co-occur in mis-leading contexts. We incorporate this hypothesis into a Integer Linear Programming framework by solving the following optimization problem:

$$
\begin{aligned}
& max \sum_{i=0}^{N} (x_i \cdot \sum_{j=0}^{M} p_{i,j}^{\theta}) \\
& \text{subject to :} \\
& x_i \in \{0, 1\} \quad \forall x_i \\
& \forall x_a \text{ and } x_b \text{ violate a pairwise constraint :} \\
& \qquad x_a + x_b \leq 1 \\
& \forall x_a, x_b, x_c \text{ violate a triangle constraint :} \\
& \qquad x_a + x_b + x_c \leq 2
\end{aligned}
\tag{3}
$$

Where $x_i$ is a binary value: 1 indicates $r_i$ is selected in the final output and 0 indicates $r_i$ is removed. $\theta$ determines to which extent we penalize low confidence values. If $\theta$ equals 0 then any confidence value should be considered equally as 1, since $p^0 = 1$ by definition. As $\theta$ grows, it gives more penalty to lower confidence values; a special case is $\theta$ equals 1, where $p^\theta$ equals $p$ itself.

# 5. EXPERIMENTS

In this section we present the results of applying this joint inference method to improve cross-document information extraction.

## 5.1 Data and Scoring Metric

We use the data set from the DARPA GALE distillation task, which contains 381,588 newswire documents, for our experiment. The baseline IE system extracted 18,386 person entities, 21,621 geo-political entities and 18,792 organization entities. Without loss of generality, we ask human annotators to evaluate the quality of all of the 1128 *Family* relations and 2854 *Member_of* relations (including ORG-Aff.Employment, ORG-Aff.Membership defined in ACE).

It's important to measure how well a system performs at extracting facts across documents accurately. However, it's time-consuming to manually extract all possible facts from a large collection of documents, and thus it's difficult to measure recall. To solve this problem, we apply the Browsing Cost metric defined in [5] to evaluate cross-document IE performance:

*Browsing Cost (i)* = the number of incorrect or redundant facts that a user must examine before finding *i* correct facts.

In our experiment, we consider a fact as a link $L^i(A, B)$, which is judged as correct if and only if both $A$ and $B$ have correct boundaries and entity types and $L^i$ has correct relation or event type.

## 5.2 Overall Performance

Figure 3 and 4 demonstrate the browsing costs. Compared to the baseline, on average our approach resulted in a 13.7% user browsing effort reduction for *Family* relations and a 24.4% user browsing effort reduction for *Member_of* relations.

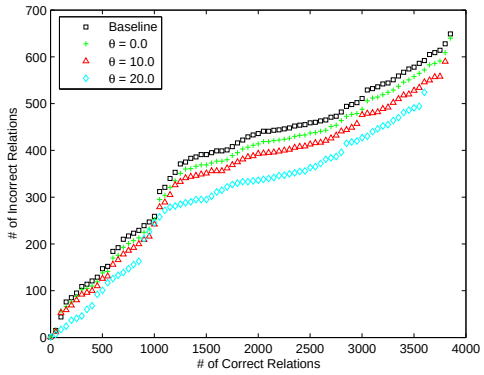


**Figure 3: Browsing cost comparison of *Member_of* relation.**

## 5.3 Discussion

Figure 5 depicts the correctness of the removal operations (i.e. the number of unique facts are removed corectly/incorrectly) for *Family* relation, varying the parameter $\theta$ of the objective function. Although our method mistakenly removed a few correct facts, it successfully removed many more incorrect instances using any parameter. For example, it removed the *Family* relation between "*Jack Straw*" and "*Tony Blair*" because they were involved in the *Business* relation ("*Jack Straw*" was in "*Tony Blair*"'s Cabinet). It occasion-

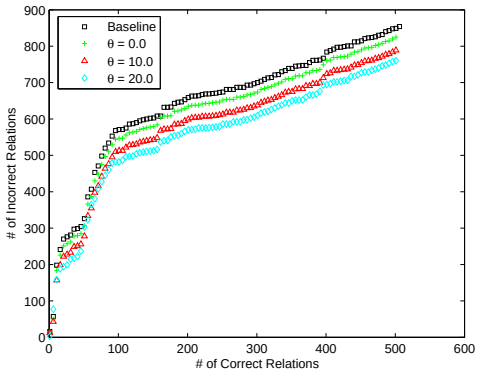


**Figure 4: Browsing cost comparison of *Family* relation.**

ally removed a few correct relations involving two person entities with multiple types of relations. For example, "*Mohammed Bakir Al-hakim*" and "*Abdul Aziz Al-hakim*" are family members as well as colleagues in the Iraqi government. Clearly overall the rewards of using joint inference significantly outweigh the risks.

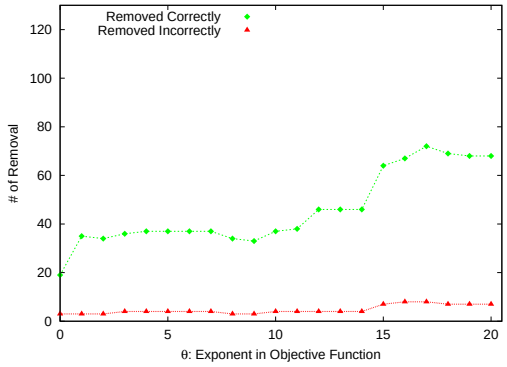


**Figure 5: Removal curve of *Family* relation w.r.t parameter $\theta$.**

In order to evaluate the impact of each constraint, we also conducted experiments using each constraint independently in the ILP model with $\theta = 20$ for *Family* relation. The results are presented in Table 2.

| Constraint type | # removed correctly | # removed incorrectly |
|---|---|---|
| Pairwise | 52 | 7 |
| Triangle | 64 | 5 |

**Table 2: Impact of different types of constraints on *Family* relation.**

## 6. CONCLUSION AND FUTURE WORK

Most previous IE research considered various types of relations and events independent of each other, and focused on individual predictions using local information. As a result, it's inevitable that contradictory facts can be extracted and the incorrect ones must then be removed in an effective way. We incorporated the interdependencies among various types of facts as latent constraints in an information network based inference framework using integer linear programming. Such joint inference analysis allowed us to incorporate information from a much wider context, going across documents and fact types, and utilize deeper semantic inferences to significantly enhance the extraction performance. In the future we are interested in applying statistical relational learning algorithms to capture more implicit constraints and soft inference rules. Finally in this paper we assumed that the constraints are relatively static, however, some facts may change over time (e.g. a person's employment or residence relation). Therefore we are also interested in extending our joint inference framework to capture temporal constraints.

## 7. ACKNOWLEDGMENTS

This work was supported in part by 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 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.

## 8. REFERENCES

[1] N. Chambers and D. Jurafsky. Jointly combining implicit constraints improves temporal ordering. In *EMNLP*, pages 698–706, 2008.

[2] P. Gupta and H. Ji. Predicting unknown time arguments based on cross-event propagation. In *Proc. ACL-IJCNLP2009*, 2009.

[3] Y. Hong, J. Zhang, B. Ma, J. Yao, G. Zhou, and Q. Zhu. Using cross-entity inference to improve event extraction. In *Proc. ACL2011*, 2011.

[4] H. Ji and R. Grishman. Refining event extraction through cross-document inference. In *Proc. ACL*, 2008.

[5] H. Ji, R. Grishman, Z. Chen, and P. Gupta. Cross-document event extraction, ranking and tracking. In *Proc. RANLP*, 2009.

[6] H. Ji, D. Westbrook, and R. Grishman. Using semantic relations to refine coreference decisions. In *Proc. HLT/EMNLP2005*, 2005.

[7] S. Liao and R. Grishman. Using document level cross-event inference to improve event extraction. In *Proc. ACL2010*, 2010.

[8] A. Mccallum. Information extraction, data mining and joint inference. In *Proc. KDD2006*, 2006.

[9] D. S. McNamara. Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. In *Canadian Journal of Experimental Psychology*, 2001.

[10] H. Poon and L. Vanderwende. Joint inference for knowledge extraction from biomedical literature. In *HLT-NAACL*, pages 813–821, 2010.

[11] D. Roth and W. tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proc. CoNLL*, pages 1–8, 2004.

[12] M. Tatu and M. Srikanth. Experiments with reasoning for temporal relations between events. In *Proc. COLING2008*, 2008.

[13] Van de Cruys Tim. Two Multivariate Generalizations of Pointwise Mutual Information. In *Proc. ACL2011*, 2011.