Open-domain Multi-Document Summarization via Information Extraction: Challenges and Prospects

Heng Ji[1], Benoit Favre[2], Wen-Pin Lin[1],
Dan Gillick[3], Dilek Hakkani-Tur[4], Ralph Grishman[5]

[1] Computer Science Department, Queens College and Graduate Center,
City University of New York, New York, NY, USA
hengji@cs.qc.cuny.edu

[2] LIF, Aix-Marseille Université, France
benoit.favre@lif.univ-mrs.fr

[3] Computer Science Department, University of California, Berkeley, Berkeley, CA, USA
dgillick@berkeley.edu

[4] Speech Labs, Microsoft, Mountain View, CA, USA
dilek@ieee.org

[5] Computer Science Department, New York University, New York, NY, USA
grishman@cs.nyu.edu

Abstract. Information Extraction (IE) and Summarization share the same goal of extracting and presenting the relevant information of a document. While IE was a primary element of early abstractive summarization systems, it's been left out in more recent extractive systems. However, extracting facts, recognizing entities and events should provide useful information to those systems and help resolve semantic ambiguities that they cannot tackle. This paper explores novel approaches to taking advantage of cross-document IE for multi-document summarization. We propose multiple approaches to IE-based summarization and analyze their strengths and weaknesses. One of them, re-ranking the output of a high performing summarization system with IE-informed metrics, leads to improvements in both manually-evaluated content quality and readability.

Keywords: Multi-document Summarization, Information Extraction

1 Introduction

Since about one decade ago Information Extraction (IE) and Automated Text Summarization have been recognized as two tasks sharing the same goal -- extract accurate information from unstructured texts according to a user's specific desire, and present the information to the user in a compact form [1]. Summarization aims to formulate this information in natural language sentences, whereas IE aims to convert the information into structured representations (e.g., databases). These two tasks have

been studied separately and quite intensively over the past decade. Various corpora have been annotated for each task, a wide range of models and machine learning methods have been applied, and separate official evaluations have been organized. There has clearly been a great deal of progress on the performance of both tasks.

Because a significant percentage of queries in the summarization task involve facts (entities, relations and events), it is beneficial to exploit facts extracted by IE techniques to improve automatic summarization. Some earlier work (e.g., [2], [3]) used IE as defined in the Message Understanding Conferences (MUC) [4] to generate or improve summaries. The IE task has progressed from MUC-style single template extraction to more comprehensive extraction tasks that target more fine-grained types of facts, such as the 18 types of relations and 33 types of events defined in NIST Automatic Content Extraction (ACE2005) [1] and the 42 types of slots defined in the Knowledge Base Population (KBP) track at the Text Analysis Conference (TAC2010) [29]. IE methods have also advanced from single-document IE to cross-document dynamic event chain extraction (e.g., [5]) and attribute extraction in KBP. In addition, recent progress on open-domain IE [7] and on-demand IE [8] can address the portability issue of IE systems and makes IE results more widely applicable. Furthermore, many current IE systems exploit supervised learning techniques, which enable them to produce reliable confidence values (e.g., [9]). Therefore allowing the summarization task to choose using IE results according to confidence values would improve the flexibility of this task. For these reasons we feel the time is now ripe to explore some novel methods to marry these two tasks again and improve the performance of the summarization task.

In this study, we test the following scenarios for combining these two tasks: IE-only based template filling and sentence compression for abstractive summary generation, IE for sentence re-ranking and redundancy removal, and IE-unit based coverage maximization. We start from a more ambitious paradigm which can generate abstractive summaries entirely based on IE results. Given a collection of documents for a specific query, we extract facts in both the queries and the documents. We implement two different approaches of utilizing these facts: template-filling and fact stitching based sentence compression. Both approaches obtain poor content and readability/fluency scores because IE still lacks coverage, accuracy and inference. Then we take a more conservative framework. We use a high-performing multi-document extractive summarizer as our baseline, and tightly integrate IE results into its sentence ranking and redundancy removal. Experiments on the NIST Text Analysis Conference (TAC) multi-document summarization task [11] show this integration method can achieve significant improvement on both standard summarization metrics and human judgment. In addition, we also provide extensive analysis on the strengths and weaknesses of these approaches.

1 http://www.nist.gov/speech/tests/ace/

2 Related Work

Our work re-visits the idea of exploiting IE results to improve multi-document summarization proposed by Radev et al. [2] and White et al. [3]. In [2], IE results such as entities and MUC events were combined with natural language generation techniques in summarization. White et al. [3] improved Radev et al.'s method by summarizing larger input documents based on relevant content selection and sentence extraction. They also formally evaluated the performance of this idea. More recently, Filatova and Hatzivassiloglou [21] considered the contexts involving any pair of names as general 'events' and used them to improve extractive summarization. Vanderwende et al. [22] explored an event-centric approach and generated summaries based on extracting and merging portions of logical forms. Biadsy et al. [23] exploited entity and time facts extracted from IE to improve sentence extraction for biographical summaries. Hachey [18] used generic relations to improve extractive summarization. Compared to these previous methods, we extend the usage of IE from a single template to wider types of relations and events. To the best of our knowledge our approach is the first work to apply KBP slot filling and event coreference resolution techniques to remove summary redundancy.

Recently there has been increasing interest in generating abstractive multi-document summaries based on template filling (e.g., [16]) and sentence compression (e.g., [24]; [12]). In this paper, we explore both of these methods entirely based on IE results. Rusu et al. [36] performed entity coreference resolution and generated a semantic graph with subject-verb-object triplets. Then they predicted which triplets should be included in the summary using Support Vector Machines based on diverse features including words, part-of-speech tags, sentence location, named entities, cosine similarity to centroid, pagerank scores and other graph-derived features. Our work is also related to the summarization research that incorporates semantic role labeling (SRL) results (e.g., [19], [25]). Semantic roles cover more event categories than IE, while IE can provide additional annotations such as entity resolution and event resolution which are beneficial to summarization. Furthermore, our approach of selecting informative concepts is similar to defining Summarization Content Units (SCUs) in the Pyramid Approach [26] because both methods aim to maximize the coverage of logical '*concepts*' in summaries.

3 Cross-document IE Annotation

We apply two English cross-document IE systems to extract facts from the query and source documents. These IE systems were developed for the NIST Automatic Content Extraction Program (ACE 2005) and the NIST TAC Knowledge Base Population (KBP 2010) Program [29]. ACE2005 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). The KBP Slot Filling task involves learning a pre-defined set of

attributes for person and organization entities. KBP 2010 defined 26 slot types for persons and 16 slot types for organizations. For example, "*Ruth D. Masters is the wife of Hyman G. Rickover*" indicates that the "*per:spouse*" of "*Hyman G. Rickover*" is "*Ruth D. Masters* "). Both systems produce reliable confidence values.

3.1 ACE IE System

The ACE IE pipeline ([5], [9], [10]) includes name tagging, nominal mention tagging, entity coreference resolution, time expression extraction and normalization, relation extraction and event extraction. Names are identified and classified using a Hidden Markov Model. Nominals are identified using a Maximum Entropy (MaxEnt)-based chunker and then semantically classified using statistics from the ACE training corpora. Entity coreference resolution, relation extraction and event extraction are also based on MaxEnt models, incorporating diverse lexical, syntactic, semantic and ontological knowledge. At the end an event coreference resolution component is applied to link coreferential events, based on a pair-wise MaxEnt model and a graph-cut clustering model. Then an event tracking component is applied to link relevant events on a time line.

3.2 KBP Slot Filling System

In addition, we apply a state-of-the-art slot filling system [13] to identify KBP slots for every person or organization entity which appears in the query and source documents. This system includes a bottom-up pattern matching pipeline and a top-down question answering (QA) pipeline. In pattern matching, we extract and rank patterns based on a distant supervision approach [37] using entity-attribute pairs from Wikipedia Infoboxes and Freebase [34]. Then we apply these patterns to extract attributes for unseen entities. We set a low threshold to include more candidate attribute answers, and then apply several filtering steps to remove wrong answers. The filtering steps include removing answers which have inappropriate entity types or involve inappropriate dependency paths to the entities. We also apply an open domain QA system, OpenEphyra [35] to retrieve more candidate answers. To estimate the relevance of a query and answer pair, we use the Corrected Conditional Probability (CCP) for answer validation. Finally we exploit an effective MaxEnt based supervised re-ranking method to combine the results from these two pipelines. The re-ranking features include confidence values, dependency parsing paths, majority voting values and slot types.

In the slot filling task, each slot is often dependent on other slots. For example, if the age of *X* is "*2 years old*", we can infer that there are unlikely any "employer" attributes for *X*. Similarly, we design propagation rules to enhance recall, for example, if both *X* and *Y* are children of *Z,* then we can infer *X* and *Y* are siblings. Therefore we develop a reasoning component to approach a real world acceptable answer in which all slot dependencies are satisfied. We use Markov Logic Networks (MLN) [28], a statistical relational learning language, to model these inference rules more

declaratively. Markov Logic extends first order logic in that it adds a weight to each first order logic formula, allowing for violation of those formulas with some penalty.

The general architecture of these two IE systems is depicted in Figure 1. Based on the assumption that the documents for a given query in a summarization task are topically related, we apply the extraction methods to each 'super-document' that includes the query and the source documents. As a result we can obtain a rich knowledge base including entities, relations, events, event chains and coreference links.

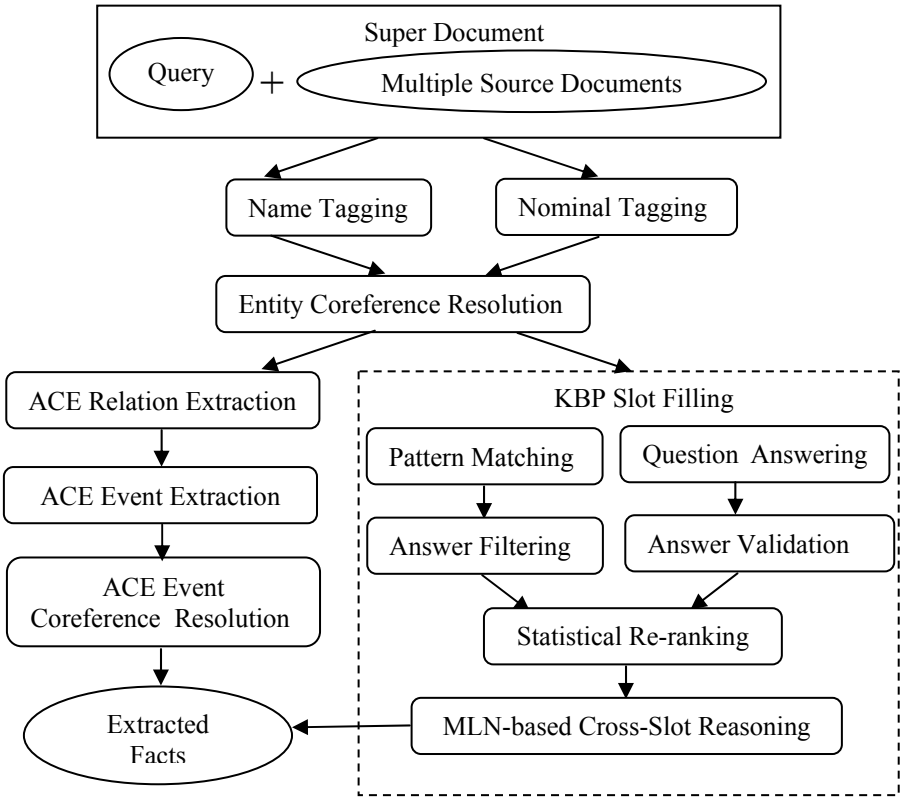

Fig. 1. Overview of IE Systems

4 Motivation for Using IE for Summarization

Using the combination of fact types in ACE and KBP, we can cover rich information in source documents. For example, among the 92 queries in the NIST TAC multi-document summarization task [11], 28 queries include explicit ACE events and their corresponding source documents include 2739 event instances. Some queries include specific events. For example, the query "*Provide details of the **attacks** on Egypt's*

Sinai Peninsula resorts targeting Israeli tourists." specifies "*attack*" events, one of the ACE event types. Some other queries inquire about general event series, such as "*Describe the views and **activities** of John C. Yoo.*" Previous research extensively focused on using entity extraction to improve summarization, so in this section we only present some concrete examples of using relations and events to improve summarization quality.

4.1 Relations/Events can Promote Relevant Sentences

Traditional sentence ranking methods in summarization used keyword matching, so the knowledge acquisition bottleneck [14] still remains due to sparse data. In order to learn a more robust sentence ranker, the method of matching query and sentences should go beyond the lexical and syntactic level in order to capture semantic structures. Several extractive summarizers (e.g., [30], [31], [32], [33]) used semantic relations in WordNet [15]. This approach has two main limitations: (1) It cannot address broader semantic relatedness; (2) It cannot address the semantic relations between two words with different part-of-speech tags. Semantic relation and event classification can provide a more flexible matching framework. Our basic intuition is that a sentence should receive a high rank if it involves many relations and events specified in the query, regardless of the different word forms indicating such relations and events. For example, for the following query sentences 1, 2 and 3 should receive high ranks according to the gold-standard summary:

> **[Query**]
> *Describe the July 7, 2005 **bombings** in **London, England** and the events, casualties and investigation resulting from the attack.*
>
> [***High-Rank Sentence 1***]
> *The **attacks**, the deadliest ever carried out on **London** in peacetime, coincided with a summit of the Group of Eight in Gleneagles, Scotland.*
>
> [***High-Rank Sentence 2***]
> *A group called Secret al-Qaida Jihad Organization in Europe claimed responsibility, saying the **attacks** were undertaken to avenge **British** involvement in the wars in Afghanistan and Iraq.*
>
> [***High-Rank Sentence 3***]
> *The **bomb exploded** in the lead car moments after the train pulled out of the **King's Cross station**, blowing apart the car and making it impossible to reach the dead and injured from the rear.*

In sentences 1 and 2, a summarizer without using IE may not able to detect "*attacks*" as the same event as "*bombings*" because they have different lexical forms. However, the IE system extracts "*conflict-attack*" events and labels "*London/British*" as "*place*" arguments in both sentences. This provides us much stronger confidence in increasing the ranks of sentences 1 and 2. Furthermore, even if the event triggers in sentence 3

"*bomb*" can be matched with "*bombings*" in the query, a summarizer may still assign a low weight to sentence 3 if it cannot detect the "*Located*" relation between "*King's Cross station*" and "*London*". But IE can successfully identify this "*Located*" relation from another sentence in the same document set: "*London - The subway tunnel between **King's Cross** and Russell Square is one of several "deep tubes" bored through **London**'s bedrock and clay more than a century ago*".

4.2 Relations/Events can Demote Irrelevant Sentences

Relations and events can also filter some irrelevant sentences by deep semantic structure analysis. For example,

> **[Query**]
> *Describe the **murders** of **Judge Joan Lefkow**'s husband and mother, and the subsequent investigation. Include details about any evidence, witnesses, suspects and motives.*
>
> [***Low-Rank Sentence 4***]
> *They remembered that he would sometimes show up at the federal courthouse to take his wife, U. S. District Judge **Joan Humphrey Lefkow**, to lunch and brought her flowers.*

A summarizer without using IE may mistakenly assign a high rank to sentence 4 because it involves a name "*Joan Humphrey Lefkow*". However, event extraction can be used to decrease the rank of this sentence because it does not include any "*Conflict-attack (murder)*" events as specified in the query.

4.3 Event Coreference can Remove Redundancy

What we have presented above is advancing summaries in terms of their *content* quality. Another central track of summarization research is the issue of *readability*, especially how to remove redundancy from multiple documents. In this paper we propose a novel approach based on event coreference resolution to reach this goal. Compared to similarity computation methods based on lexical features, our method can detect similar pairs of sentences even if they use completely different expressions. For example, we can fuse the following sentences because they include coreferential "*Conflict-attack*" events, with "*blasts/bombings*" as indicative words and "*London*" as their place arguments:

> [***Sentence 5***]
> *It was the deadliest of the four bomb **blasts** in **London** last week.*
>
> [***Sentence 6***]
> *The bus explosion was one of four co-ordinated **bombings**, the others on **London** Underground subway trains.*

It will be challenging for a summarizer without using IE to detect this redundancy because most words don't overlap in these two sentences.

4.4 Integrating IE and summarization

Methods for incorporating IE into summarization range from using IE alone to using IE to modify the behavior of existing summarization systems. Here we list five general approaches using facts extracted by IE, entities, relations and events, which we call IE units. These methods are schematized in Figure 2.

Template-based generation consists in detecting IE units, such as events and the entities involved in them, and feeding a generation module which uses templates for building summary sentences. Such templates could be: "*Attack: [Attacker] attacked [Place] on [Time-Within]*", which would result in, for instance, "*[Terrorists] attacked [the police station in South Bagdad] on [Friday]*". Such an approach is known for being susceptible to the coverage of the template rules.

IE-based compression is similar to template-based generation but it does not use pre-existing templates but rather takes advantage of the support of the IE units in the original documents. The process is less prone to the lack of good template coverage and generates sentences closer to the source but it requires very accurate detection of IE unit spans.

If an existing summarization system is available, one unobtrusive way to take advantage of IE is to pre-filter the input documents. For instance, if the summarization system is extractive, sentences that do not contain any IE elements can be dropped from its input. Various ways of performing such filtering can be devised depending on the type of summarization system.

The next step applies specifically to summarization systems that compute sentence-level similarities (like Maximal Marginal Relevance) by infusing these similarities with IE units. For instance, if two sentences involve coreferential events, then they may be marked as redundant and not used together in the summary. Various graph-based methods, unsupervised and supervised sentence relevance prediction methods can be used to reach this goal.

Finally, for coverage-based methods that account for "*concepts*" or "*information units*" present in the summary instead of sentence-level scores, IE can be used to infer relevant "*concepts*" and to deem redundant, for instance concepts that refer to the same entity ("*the president*" and "*Mr Obama*").

All these approaches can be used in conjunction, and we will, in the following sections, review a few possibilities.

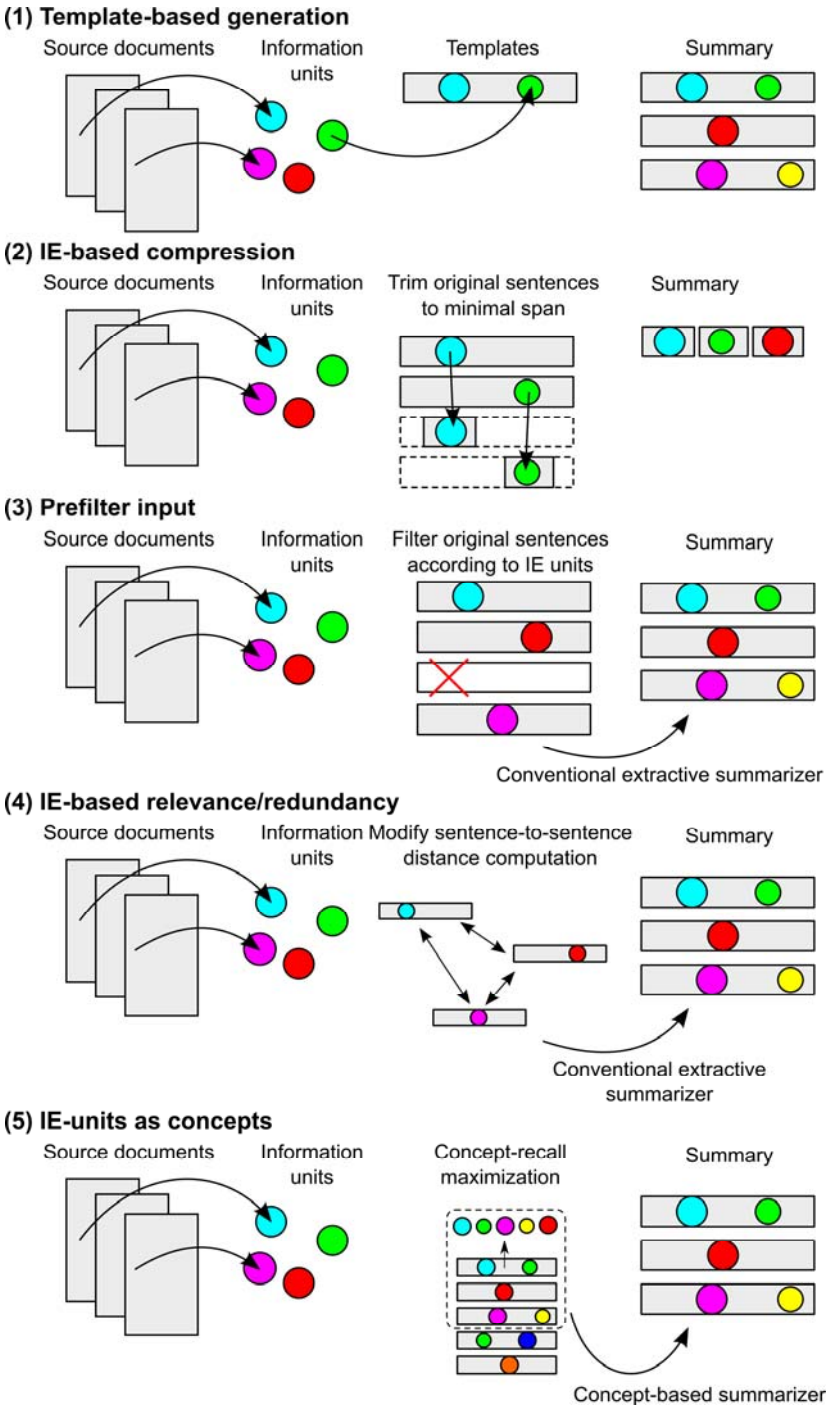

Fig. 2. Methods for Integrating IE with Summarization

5 Proposed Approaches

5.1 IE-based Template Filling

Traditional IE by definition is a task of identifying '*facts*' from unstructured documents, and filling these facts into structured representations (e.g., templates). Therefore, the most natural way of using IE results for summarization is to fill templates as described in some previous work ([2], [3], [16]).

Table 1. IE-based Template Examples

Event Type	Event SubType	Templates
Movement	Transport	[Agent] left [Origin].
Personnel	Elect	[Entity] elected [Person] as [Position] of [Place] on [Time-Within].
		[Person] was elected in [Place].
		[Person] was elected in [Place] on [Time-Within].
Personnel	Start-Position	[Person] was hired.
	End-Position	[Person] was fired
Life	Die	[Agent] killed [Victim].
		[Victim] died.
		[Victim] died on [Time-Within].
	Marry	The marriage took place in [Place].
		[Person] and [Person] married.
		[Person] and [Person] married in [Place] on [Time-Within].
Conflict	Demonstrate	There was a demonstration in [Place] on [Time-Within].
Conflict	Attack	[Attacker] attacked [Place] on [Time-Within].
		[Attacker] attacked [Place].
		[Attacker] attacked [Target].
		[Target] was attacked on [Time-Within].
		The attack occurred on [Time-Within].
		[Attacker] attacked.
Transaction	Transfer-Ownership	[Buyer] made the purchase on [Time-Within].
Justice	Arrest-Jail	[Person] was arrested.
	Trial-Hearing	[Defendant] was tried.

For ACE event types and some KBP slots, we construct specific templates. This approach can be summarized as follows: (1) apply IE to the query and source documents; (2) fill sentence templates with available pieces of information (event arguments, entities and attributes), and replace pronouns and nominal mentions with their coreferential proper names; (3) arrange sentences using temporal relations (if

there are no explicit temporal arguments, then using the text reporting order) up to the summary length constraint. Some examples of the templates generated based on ACE events are shown in Table 1.

For example, we can generate the following summary sentence using the "*Personnel/Elect*" event template "*[Person] was elected in [Place]*.":

[Original Sentence 7]
After a bitter and protracted recount fight in the ***Washington*** *governor's race,* ***elections*** *officials announced Wednesday that the Democratic candidate,* ***Christine O. Gregoire****, was leading her Republican opponent by 10 votes, a minuscule margin but a stunning reversal of the Nov. 2 election results.*

[Summary Sentence]
Christine O. Gregoire *was elected in* ***Washington*****.**

In addition, the summary sentences can be ordered based on their time arguments:

[Original Sentence 8]
Charles *announced he would* ***marry Camilla*** *in a civil ceremony at* ***Windsor Castle*** *on* ***April 8****.*
Charles *was previously* ***married*** *to Princess* ***Diana****, who* ***died*** *in a car crash in* ***Paris*** *in* ***1997****.*

[Summary Sentence]
Charles *and* ***Diana*** **married.** ***Diana*** *died in* ***Paris*** on ***1997. Charles*** *and* ***Camilla*** *married in* ***Windsor Castle*** *on* ***April 8.***

Ordering sentences based on event time arguments can produce summaries with better readability because the text order by itself is a poor predictor of chronological order (only 3% temporal correlation with the true order) [5].

5.2 IE-based Sentence Compression

IE-based template filling can only be fruitful if the template database has a large-enough coverage. In order to build a template-independent approach, we define IE-based sentence compression: instead of filling templates, use source words found in event and relation mentions to build summary sentences. For example, the following summary can be generated by compressing sentence 9:

[Original Sentence 9]
Four bombers *were among those* ***killed*** *in* ***Thursday****'s* ***attacks on a double-decker bus****, Sky News television reported* ***Tuesday****, quoting police sources.*

[Summary Sentence]
Four bombers *were among those* ***killed*** *in* ***Thursday****'s* ***attacks on a double-decker bus****, Sky News television reported* ***Tuesday.***

Using 'mentions' (the maximum span that covers the trigger and all arguments of a relation/event instance) to build summary sentences results in more faithful wording with respect to the source. However, the syntactic structure of source sentences is not always favorable to these kinds of extractions, generating verb-less sentences or decontextualized entity references. For example, the following summary was created from multiple sentences:

[Original Sentences 10]
Forty-four victims *of the London subway and bus* ***bombings*** *remained in hospitals Friday.*

[Summary Sentence]
Forty-four victims *of the London subway and bus* ***bombings.***

[Original Sentences 11]
British police said Friday they were "aware" of an ***arrest*** *in* ***Egypt*** *in connection with the investigation into* ***last week****'s* ***London bombings****.*

[Summary Sentences]
Arrest *in* ***Egypt*** *in connection with the investigation into* ***last week****'s* ***London bombings.***

5.3 IE-based Relevance Estimation

IE provides an effective way of modeling the central information described in the source documents. Even if the IE model describes such information perfectly, it does not tell us what subset of IE units should appear in a summary. As discussed earlier, IE can be integrated to existing summarization systems at the sentence similarity level to characterize relevance or redundancy. For the purpose of this work, we focus on a linear model to re-rank the relevance scores of a baseline summarizer with sentence-level IE scores.

For a given query Q and a collection of source documents D that includes N sentences (s_1, …, s_N), we generate a summary based on an integrated approach as follows.

Each IE component includes a statistical classifier which generates reliable confidence values. For example, for each event mention in D, the baseline Maximum Entropy based classifiers produce three types of confidence values:

- *Conf(trigger,etype)*: The probability of a string *trigger* indicating an event mention with type *etype*.
- *Conf(arg, etype)*: The probability that a mention *arg* is an argument of some particular event type *etype*.

- *Conf(arg, etype, role)*: If *arg* is an argument with event type *etype*, the probability of *arg* having some particular *role*.

For any sentence s_i in *D*, we extract the confidence values presented in Table 2.

Table 2. IE Confidence Values

Confidence	Description
$c_1(s_i, e_j)$	confidence of s_i including an entity e_j which is coreferential with an entity in *Q*
$c_2(s_i, r_k)$	confidence of s_i including a relation mention r_k which shares the same type and arguments with a relation mention in *Q*
$c_3(s_i, ev_l)$	confidence of s_i including an event mention ev_l which shares the same type and arguments with an event mention in *Q*
$c_4(s_i, kbp_m)$	confidence of s_i including a KBP relation kbp_m which shares the same slot type, entity and slot value with a KBP relation in *Q*

We then linearly combine them to form the final IE confidence for s_i as follows.

$$c_{ie}(s) = \alpha_1 \times \sum_j c_1(s_i, e_j) + \alpha_2 \times \sum_k c_2(s_i, r_k) + \alpha_3 \times \sum_l c_3(s_i, ev_l) + \alpha_4 \times \sum_m c_4(s_i, kbp_m)$$

The α parameters are optimized using a development set. Assuming the ranking confidence from the baseline summarizer for s_i is $c_{baseline}(s_i)$, we can get the combined confidence of using s_i as a summary sentence:

$$c_{summary}(s_i) = (1-\lambda) \times (c_{baseline}(s_i) / \sum_{i=1}^{N} c_{baseline}(s_i)) + \lambda \times (c_{ie}(s_i) / \sum_{p=1}^{N} c_{ie}(s_p))$$

We believe that incorporating these confidence values into a unified re-ranking model can provide a comprehensive representation of the information in the source collection of documents. Based on the combined confidence values, we select the top sentences to form a summary within some certain length constraint specified by the summarization task.

5.4 IE-based Redundancy Removal

While IE extracted entities, relation mentions and event mentions which might help introduce more relevant sentences, it does not prevent identical pieces of information from being represented multiple times in the summary under different wordings. We address redundancy removal by taking advantage of coreference links to drop sentences that do not bring new content.

This approach is implemented by filtering the ranked-list of sentences generated from the baseline summarizer. In particular, we conduct the following greedy search through any sentence pair of <s_i, s_j>:

- If all of the entity and event mentions in s_i are coreferential with a subset of the entity and event mentions in s_j, then remove s_i;
- If all of the entity and event mentions in s_i and s_j are coreferential, and s_i is shorter than s_j, then remove s_i.

For example, the following sentences include coreferential "*Personnel/End-Position*" events, so we remove the shorter sentence 13.

> [***Sentence 12***]
> ***Armstrong****, who* ***retired*** *after his seventh yellow jersey victory* ***last month****, has always denied ever taking banned substances, and has been on a major defensive since a report by French newspaper L'Equipe last week showed details of doping test results from the Tour de France in 1999.*
>
> [***Sentence 13***]
> ***Armstrong retired*** *from cycling after his record seventh straight Tour victory* ***last month****.*

5.5 IE-unit Coverage Maximization

Recent work in summarization has lead to the emergence of coverage-based models ([12] and references therein). Instead of modeling sentence-level relevance and redundancy, these models assess the value of information units, or "*concepts*", that appear in input sentences. A summary is created by concatenating sentences according to the concepts they contain, effectively tackling the problem of redundancy in sets of more than two sentences. Finding a selection of sentences in this model corresponds to solving a set-cover problem with a knapsack constraint (the length limit of the summary).

While concepts are mostly embodied by word n-grams, we suggest to cast the problem as finding the set of sentences that cover the most important IE units. We use frequency for characterizing the importance of IE units and perform inference with the following Integer Linear Program (ILP):

$$\text{Maximize} \quad \sum_i frequency(i) \times u_i$$

$$\text{Subject to} \quad \sum_j length(j) \times s_j \le length\_bound$$

$$s_j = 1 \Rightarrow u_i = 1 \quad \forall units \in sentence_j$$

$$u_i = 1 \Rightarrow \text{at least one } s_j = 1 \quad \forall \text{ sentences that contain } u_i$$

In this ILP, u_i is a binary variable indicating the presence of unit i in the summary; s_j is a binary variable indicating the presence of sentence j in the summary. Details on the formulation can be found in [12]. This model is particularly suited for incorporating IE results because IE extracted facts make particularly good concepts for the model. For instance, selecting the sentences that would cover all events central

to the topic would make a very relevant summary. In this approach, we perform cross-document IE to detect entities, relations and events, then associate each of the detected elements to corresponding units and find the set of sentences that maximizes weighted IE-unit coverage.

6 Experimental Results

In this section, we describe an experimental framework for evaluating the quality of the proposed approaches.

6.1 TAC Summarization Task

The summarization task we are addressing is that of the NIST Text Analysis Conference (TAC) multi-document summarization evaluation [11]. This task involves generating fixed-length summaries from 10 newswire documents, each related to a given query including a specific topic. While TAC also includes an update summarization task -- additional summaries assuming some prior knowledge -- we focus only on the standard task in this paper. For example, given a query "*Judge Joan Lefkow's Family Murdered/Describe the murders of Judge Joan Lefkow's husband and mother, and the subsequent investigation. Include details about any evidence, witnesses, suspects and motives.*" and 10 related documents, a summarization system is required to generate a summary about specific entities ("*Judge Joan Lefkow*"), relations ("*family*") and events ("*murder*" and "*investigation*").

In the TAC campaigns, the quality of system-generated summaries is evaluated through manual and automatic evaluations. In manual evaluation, judges give ratings on a Likert scale (1-5 or 1-10) on content responsiveness (informativeness given the input documents and the user query) and linguistic quality (grammaticality, non-redundancy, clarity of references, global organization...). Automatic evaluation compares each automatic summary to a set of expert-written summaries using distances such as word n-gram overlap (ROUGE) [20]. Multiple reference summaries are used to address the fact that there is no single good answer to the summarization problem.

6.2 Baseline Summarization System

As a baseline, we apply a top-performing TAC summarization system [12] using the principles of coverage-based summarization which was already described in section 5.5. In this model, a summary is the set of sentences that cover the most relevant concepts in the source document set, where concepts are simply word bigrams weighted by their document frequency. The concepts that include low-frequency words or stop-words are filtered. For sentence-level experiments, the value of a sentence is the sum of the concept values it contains. In addition, a sentence compression component is used to post-process the candidate sentences. The

compression step consists of dependency tree trimming using high-confidence semantic role labeling decisions. Non-mandatory temporal and manner arguments are removed and indirect discourse is reformulated in direct form.

6.3 Evaluation of IE-based Template Filling and Sentence Compression Approaches and Sentence-level Integration

We first evaluated the approaches that do not rely on an existing summarization system, IE-based Template Filling and IE-based Sentence Compression, and a system using sentence-level integration through IE-based Relevance Estimation followed by IE-based Redundancy Removal. In order to perform this evaluation, we randomly selected 31 topics from the TAC 2008 and TAC 2009 summarization tasks as our blind test set. The summaries are evaluated automatically with the ROUGE-2 and ROUGE-SU4 metrics [20]. In order to focus more on evaluating the ordering of sentences and coherence across sentences, we extend the length restriction in the TAC setting from 100 words to 20 sentences. Therefore the results are not directly comparable with those of the official evaluation.

We also asked 16 human subjects to manually evaluate summaries based on the TAC Responsiveness metrics [11] consisting of Content and Readability/Fluency measures. In order to compare different methods extensively, we asked the annotators to give a score in the [1, 5] range (1-Very Poor, 2-Poor, 3-Barely Acceptable, 4-Good, 5-Very Good).

In this evaluation, our baseline is the word-bigram based system described in section 6.2 in which sentences are first valued according to the concepts they contained, and then selected in order of decreasing value. The sentences output by this baseline are then rescored by IE-based Relevance Estimation and pruned using IE-based Redundancy Removal. Parameters of the relevance re-ranking module are estimated on a development set (the documents not used for scoring) in order to maximize ROUGE-2 recall. For processing the test set, we use α_1=1, α_2=2, α_3=3, α_4=1 (IE components) and λ=0.7 (IE weight respective to the baseline). If a query does not include any facts extracted by IE, we use the summaries generated from the baseline summarizer.

6.3.1 ROUGE Scores

The ROUGE-2 results of each system are summarized in Table 3. The IE-based Template Filling and Sentence Compression methods perform poorly in term of ROUGE score; the integrated approach using relevance estimation and redundancy removal yields improvement over the baseline, suggesting a benefit of taking advantage of IE output.

Table 3. TAC ROUGE-2 Scores

Method	Recall	Precision
Baseline with Dependency Tree Trimming-based Sentence Compression	0.1674	0.1058
IE-based Template Filling	0.1239	0.0825
IE-based Sentence Compression	0.1297	0.0901
IE-based Relevance Estimation and Redundancy Removal	0.1798	0.1084

Figure 3 presents the detailed ROUGE-2 and ROUGE-SU4 results of the integrated approach according to the λ parameter. It achieves significant improvement on Recall: when we use λ=0.7, which is also the best weight optimized from the development set, our methods achieve 7.38% relative ROUGE-2 gain. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on ROUGE scores for these 31 topics. The results show that we can reject the hypothesis that the improvements were random at a 95.7% confidence level. From these curves we can also conclude that using IE results only (λ=1) for sentence ranking produced worse ROUGE scores than the baselines.

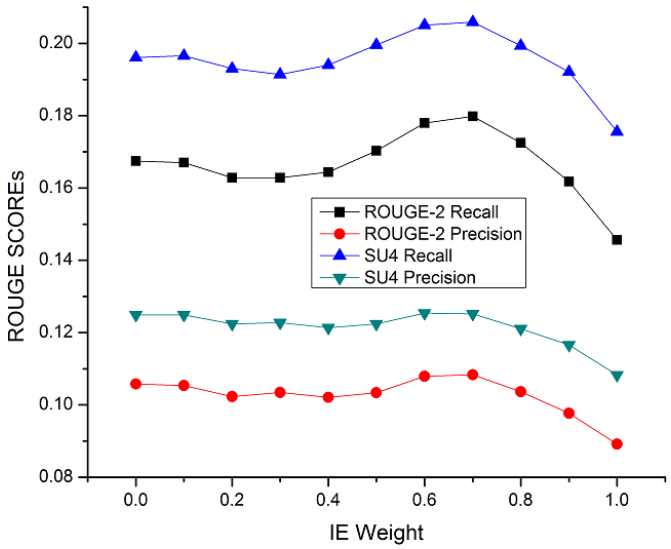

Fig. 3. IE-based Relevance Estimation and Redundancy Removal Results

6.3.2 TAC Responsiveness Scores

Table 4 presents the average scores across all topics based on manual evaluation using TAC Responsiveness metrics.

Table 4. TAC Responsiveness Comparison

Method	Content	Readability	Responsiveness
Baseline with Sentence Compression (λ=0)	3.11	3.56	3.39
IE-based Template Filling	2.24	3.08	2.64
IE-based Sentence Compression	2.73	2.85	2.76
IE-based Relevance Estimation (λ=0.7) and Redundancy Removal	3.89	3.67	3.61

The IE-only methods obtain lower responsiveness, content and readability scores, which is probably a combination of lack of coverage and bad linguistic quality. But Table 3 also shows that the IE-integrated method receives better content scores based on human assessment and even improves over the baseline. This is probably due to document sets involving facts that are ambiguous when using words only for modeling. For example, for the query "*Provide details of the kidnapping of journalist Jill Carroll in Baghdad and the efforts to secure her release*", the baseline summarizer received a score of "2" because of a mismatch between "*kidnapping*" in the query and the "*arrest*" events involving other person and place arguments in the source documents. In contrast, the IE-informed method received a score of "4", because of the effective integration of the "kidnap" event detection results when re-ranking sentences. Furthermore, according to the user feedback, our method produced fewer redundant sentences for most topics.

Error analysis shows that for 3 topics IE had negative impact because of incorrect event categorization for the queries, and missing/spurious extraction errors. For example, for the query "*BTK/Track the efforts to identify the serial killer BTK and bring him to justice.*", IE mistakenly recognized "*Justice*" as the main event type while it missed the more important event type "*Investigation*" which was not defined in the 33 event types. In these and other cases, we could apply salience detection to assign weights to different fact types in the query. Nevertheless, as the above results indicate, the rewards of using the IE information outweigh the risks.

6.4 Evaluating the IE-unit Coverage Maximization approach

For this set of experiments, we compared the coverage maximization system based on word bigrams as concepts with the same system with IE-units as concepts in addition to the word-based concepts. The behavior of this system can be tuned through a λ parameter which acts as a multiplicative factor of the frequency of IE units when computing their value. Therefore, λ=0 implies ignoring IE units and λ=1 gives an equal weight to word-based and IE-based concepts.

This time we processed 50 document sets from the non-update part of the TAC'09 evaluation in order to compare with state-of-the-art results. Table 5 shows the ROUGE-2 results for the IE-only system (where IE-units are used as concepts), the baseline system (Coverage maximization, words only) and the mixed system which takes advantage of both units.

Table 5. ROUGE-2 Scores for the Coverage-based Systems on TAC'09

System	ROUGE-2 Recall
Baseline (λ=0)	0.1237
IE only	0.0859
Baseline + IE (λ=1)	0.1199

Results show that neither IE-only nor the mixed system outperform the baseline in term of ROUGE. We tried with different values for the mixing parameter but none of them resulted in an improvement. However, we observed that sometimes the IE-infused system can outperform the baseline (for instance, in 10 out of 50 topics for the λ=1 system). A careful analysis of the results demonstrated that IE does not cover enough relevant events to gain the advantage over word n-grams, and that most events are only detected once which makes frequency a bad estimator of their relevance.

These results are very interesting because one of the most common criticisms of word-based content-coverage models is that they do not model meaningful real-world entities whereas IE would provide those natural representatives of the actual information. Clearly, more work has to be pursued in this direction in order to empower those already high-performing models.

7 Discussion

We have seen in the experiment section that IE is not a good contender for building summarization systems when used alone. However, when it is blended with existing summarization technology, IE can bring some interesting improvements. We will discuss in this section the limitations of the current approaches and devise new avenues for future work. These limitations are annotation coverage, assessment of importance, readability and inference.

7.1 Content Coverage

The first problem with targeting IE for open-domain summarization is that even though IE methods apply broadly to many kinds of entities, relations and events, actual systems are developed in the framework of evaluation campaigns and rely on the annotated data produced in these campaigns.

Unfortunately, none of the IE shared tasks (e.g., ACE, KBP) has a large-enough coverage of frequent relations and event types. To demonstrate that, we compare the event types represented in a large corpus with those of the ACE evaluation campaign. We cluster event verbs based on cross-lingual parallel corpora [17], creating classes from verbs that often align to the same words in foreign languages, to obtain 2504 English verb clusters supposed to represent important event types. Then we rank the clusters according to the frequency of their members appearing in the LDC English Gigaword corpus. In parallel, we look at the ranking position of each ACE event type among these clusters, which gives an idea of their coverage of the whole set.

Figure 4 presents the ranking results (each point indicates one event type). We can see that although most ACE event types rank very high, a lot of other important event types are missing in the ACE paradigm. The result of this observation is that IE systems trained on ACE-annotated data have a relatively poor coverage of the long tail of event verbs found in open-domain corpora like Gigaword.

It is also interesting to look at IE coverage on the TAC datasets. These datasets contain a manual annotation of the hand-written reference summaries with basic information units (called Summarization Content Units, SCUs). Whereas detecting the information represented by these SCUs is essential for generating relevant summaries, our IE system covers only 60.67% of their words, that is 25,983 out of 42,822 for the TAC'08 data. It seems clear that by being blind to a large part of the relevant information, IE cannot help summarization to its full extent.

According to our IE-unit Coverage Maximization system, the TAC'08 document-set with the fewest information units is D0827E-A. It has 241 information units, and only three of them are in its SCU annotation. The document-set that has the most information units is D0826E-A, with 3,229 units, but also only three of them overlap with the reference SCUs. This partially explains why IE does not seem to improve over the word-based coverage maximization system.

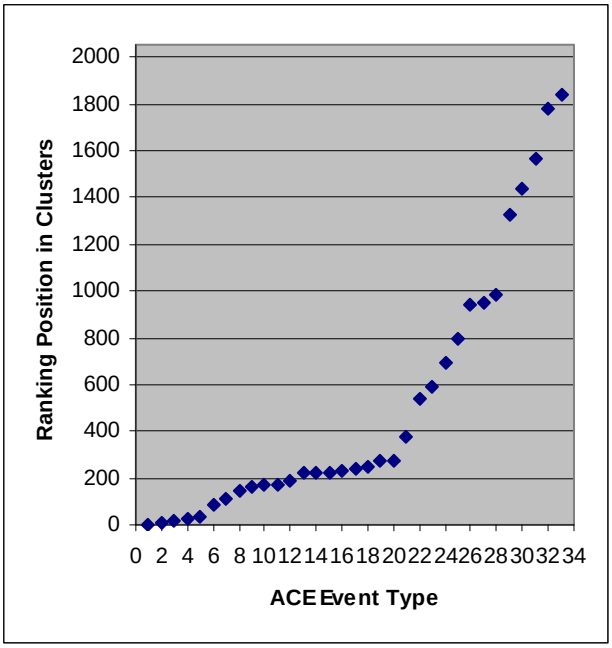

Fig. 4. Distribution of ACE Event Types among Verb Clusters

We can also make a comment on the template-based methods. They require very high accuracy in event argument role labeling, otherwise the generated sentences might contain wrong information. For example, in the sentence 8 in section 5.1, our template was not able to capture the tense information ("*would*") for the trigger word

"*marry*" and so produced the wrong tense in the summary "*Charles and Camilla married…*". For these methods, trading-off coverage for accuracy is not going to be an option.

In order to improve the coverage of current IE systems, we will have to devise new techniques for automatically expanding data from evaluation campaigns with new entity types and new events. Another approach would be to rely on semi-supervised learning to generalize knowledge learned from existing annotation to unseen types. Domain-independent techniques such as On-demand IE [8] can cover wider fact types, but have lower precision. An additional possible solution is to exploit more coarse-grained templates based on generic relation discovery [18] or semantic role labeling [19], but fine-grained annotations, such as event types and argument roles, are beneficial to select sentences that contain relevant facts specified in queries. Clearly, in order to get good-enough coverage for summarization applications, IE researchers will have to close the gap between fine-grained IE-elements for which only little training data is available and broader definitions of semantic roles and semantic frames.

7.2 Assessment of Importance

Summarization consists in gathering the most important pieces of information of a text in a limited space. This aspect is typically called "*relevance*" in the summarization literature, originating from the information retrieval literature. Relevance is measured as a combination of frequency and specificity. Frequent phenomena are good topical representatives (content words) unless they are part of the structure of the input (stop-words). The term frequency-inverse document frequency (TF-IDF) framework has been very successful for information retrieval and summarization, but does it directly apply to facts extracted by IE? In particular, in the word-based coverage maximization model, the document frequency of word n-grams is used to estimate their importance for the summary. We tried to apply the same framework for IE-units but it did not yield positive results. How are we supposed to deal with types of units and relational elements like the fact that a person is involved in a particular event? For instance, in the TAC'08 document sets, events occur at most once per document, making frequency a very crude proxy of importance. In addition it is difficult to infer the importance of an event or a relation according to the importance given to the involved participants. Devising good models of importance is critical in order to construct summaries focused on the most important facts if the length constraint does not allow for all facts to be presented. Using frequency from large corpora directly might not be a good solution since the documents being summarized are likely to deal with non-recurring events or events outside of the domain of available corpora. A generalization process could help for instance to infer the importance of an event from similar events, events from the same category or co-occurring events (e.g. "*what is the expected timeline following an earthquake event?*").

7.3 Readability

The IE-only summarization methods like Template Filling and Sentence Compression resulted in poor readability. Both methods suffer from misdetections that crudely insert spurious elements in the templates or extract wrong elements in the compression method. When coupled with extractive summarization, the effect is not as visible due to the fact that full original sentences are used in the final summary. In TAC-involved summarization systems, IE is also often used for replacing pronominal and nominal references with their antecedent. There again, IE errors lead to reduced readability. Improving IE accuracy seems the only remedy to those problems, but confidence scores could be a good source of information for relaxing IE-induced linguistic constraints when IE output is not estimated to be of high-enough quality. We showed an example of such processing in our IE-based relevance estimation method. Another way to improve readability is to take advantage of the progress in text generation, for instance by using language models for rescoring multiple summary hypotheses.

7.4 Inference and Generalization

Most of the current IE techniques do not tackle sophisticated inferences; that is, for instance, the fact that if somebody was somewhere on a given date, then this person cannot be at another place at that time. In fact, IE only detects stated information, and misses information that would be implied by world knowledge. None of the systems presented in this paper perform such processing whereas it would be particularly appropriate to assess the importance of IE elements. For instance, if it can be implied that a person participated in an important meeting then the importance of that person can be increased for the final summary. In addition, inferences could provide means of detecting inconsistencies in the input documents. Textual entailment sometimes uses IE as a tool, but both should really be considered as a joint problem.

8 Conclusions and Future Work

We investigated the once-popular IE-driven summarization approaches in a wider IE paradigm. We proposed multiple approaches to IE-based summarization and analyzed their strengths and weaknesses. We first concluded that simply relying upon IE for abstractive summarization is not in itself sufficient to produce informative and fluent summaries. Then we demonstrated that a simple re-ranking approach with IE-informed metrics can achieve improvement over a high-performing extractive summarizer. We expect that as IE is further developed to achieve higher performance in broader domains, the summarization task can benefit more from such extended semantic frames. We hope the experiments shown in this paper can draw some interest in both the IE and summarization communities. In the future we will attempt a hybrid approach to combine abstractive and extractive summarization techniques. In addition, we plan to incorporate high-confidence results from open-domain IE to increase the coverage of information units for summarization.

Acknowledgement

The first author was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, 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 Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

1. Grishman, R., Hobbs, J., Hovy, E., Sanfilippo, A., Wilks, Y: Cross-lingual Information Extraction and Automated Text Summarization. Linguistica Computazionale, Volume XIV-XV (1997)
2. Radev, D. R., McKeown, K. R.: Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3). pp. 469–500 (1998)
3. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., Wagstaff, K.: Multidocument Summarization via Information Extraction. Proc. Human Language Technologies (HLT 2001). pp. 263-269 (2001)
4. Grishman, R., Sundheim, B.: Message Understanding Conference - 6: A Brief History. Proc. the 16th International Conference on Computational Linguistics (COLING 1996). pp. 466-471 (1996)
5. Ji, H., Grishman, R., Chen, Z., Gupta, P.: Cross-document Event Extraction, Ranking and Tracking. Proc. Recent Advances in Natural Language Processing (RANLP 2009). pp. 166-172 (2009)
7. Banko, M., Cafarella, M., J., Soderland, S., Etzioni, O.: Open Information Extraction from the Web. Proc. International Joint Conferences on Artificial Intelligence (IJCAI 2007) (2007)
8. Sekine, S.: On-Demand Information Extraction. Proc. Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL 2006) (2006)
9. Ji, H., Grishman, R.: Refining Event Extraction Through Cross-document Inference. Proc. the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008) (2008)
10. Grishman, R., Westbrook, D., Meyers, A.: NYU's Chinese ACE 2005 EDR System Description. Proc. NIST Automatic Content Extraction Workshop (ACE2005) (2005)
11. Dang, H. T., Owczarzak, K.: Overview of the TAC 2009 Summarization Track. Proc. Text Analysis Conference (TAC 2009) (2009)
12. Gillick, D., Favre, B., Hakkani-Tur, D., Bohnet, B., Liu, Y., Xie, S.: The ICSI/UTD Summarization System at TAC 2009. Proc. Text Analysis Conference (TAC 2009) (2009)
13. Chen, Z., Tamang, S., Lee, A., Li, X., Lin, W., Artiles, J., Snover, M., Passantino, M.,Ji, H.: CUNY-BLENDER TAC-KBP2010 Entity Linking and Slot Filling System Description. Proc. Text Analysis Conference (TAC2010) (2010)

14. Yarowsky, D.: Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proc. the 14th International Conference on Computational Linguistics (COLING 1992) (1992)
15. Fellbaum, C. (Ed.). WordNet: An Electronic Lexical Database. Cambridge, MA: The MIT Press (1998)
16. Sauper, C., Barzilay, R.: Automatically Generating Wikipedia Articles: A Structure-Aware Approach. Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) (2009)
17. Callison-Burch, C.: Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP 2008) (2008)
18. Hachey, B.: Multi-Document Summarisation Using Generic Relation Extraction. Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP 2009). pp. 420-429 (2009)
19. Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A., Popowich, F.: Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task. Proc. Document Understanding Conference (DUC2005) (2005)
20. Lin, C., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. Proc. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003). pp. 150-156 (2003)
21. Filatova, E., Hatzivassiloglou, V.: A Formal Model for Information Selection in Multi-Sentence Text Extraction. Proc. the 20th International Conference on Computational Linguistics (COLING 2004) (2004)
22. Vanderwende, L., Banko, M., Menezes, A.: Event-Centric Summary Generation. Proc. Document Understanding Conference (DUC 2004) (2004)
23. Biadsy, F., Hirschberg, J., Filatova, E.: An Unsupervised Approach to Biography Production using Wikipedia. Proc. the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2008). pp. 807-815 (2008)
24. Liu, F., Liu, Y.: From Extractive to Abstractive Meeting Summaries: Can It Be Done by Sentence Compression? Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) (2009)
25. Melli, G., Shi, Z., Wang, Y., Liu, Y., Sarkar, A., Popowich, F.: Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2006 Summarization Task. Proc. Document Understanding Conference (DUC 2006) (2006)
26. Nenkova, A., Passonneau, R.: Evaluating Content Selection in Summarization: The Pyramid Method. Proc. Human Language Technology Conference-North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2004) (2004)
27. McKeown, K., Passonneau, R., Elson, D., Nenkova, A., Hirschberg, J.: Do summaries help? A task-based evaluation of multi-document summarization. Proc. the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005) (2005)
28. Richardson, M., Domingos, P.: Markov Logic Networks. Machine Learning. 62:107-136 (2006)

29. Ji, H., Grishman, R., Dang, H. T., Griffitt, K., Ellis, J.: An Overview of the TAC2010 Knowledge Base Population Track. Proc. Text Analysis Conference (TAC2010) (2010)

30. Dang, C., Luo, X., Zhang, H. : Wordnet-based Summarization of Unstructured Document. Journal of WSEAS Transactions on Computers. Volume 7 Issue 9, September 2008 (2008)

31. Chaves, R. P.: WordNet and Automated Text Summarization. Proc. the 6th Natural Language Processing Pacific Rim Symposium (2001)

32. Bellare, K., Sarma, A. D., Loiwal, N., Mehta, V., Ramakrishnan, G., Bhattacharyya, P.: Generic Text Summarization Using WordNet. Proc. the 4th International Conference on Language Resource and Evaluation (LREC2004) (2004)

33. Vikas, O., Meshram, A. K., Meena, G., Gupta, A.: Multiple Document Summarization Using Principal Component Analysis Incorporating Semantic Vector Space Model. Computational Linguistics and Chinese Language Processing, Volume 13, No. 2, June 2008, pp. 141-156 (2008)

34. Bollacker, K., Cook, R., Tufts, P.: Freebase: A Shared Database of Structured General Human Knowledge. Proc. National Conference on Artificial Intelligence (Volume 2) (2007)

35. Schlaefer, N., Ko, J., Betteridge, J., Sautter, G., Pathak, M., Nyberg, E.: Semantic Extensions of the Ephyra QA System for TREC2007. Proc. Text Retrieval Conference (TREC2007) (2007)

36. Rusu, D., Fortuna, B., Grobelink, M., Mladenic, D.: Semantic Graphs Derived from Triplets with Application in Document Summarization. Informatica, 33 (2009), pp 357-362 (2009)

37. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant Supervision for Relation Extraction without Labeled Data. Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009). (2009)