Using Topic Themes for Multi-Document Summarization

SANDA HARABAGIU and FINLEY LACATUSU
University of Texas at Dallas

The problem of using topic representations for multidocument summarization (MDS) has received considerable attention recently. Several topic representations have been employed for producing informative and coherent summaries. In this article, we describe five previously known topic representations and introduce two novel representations of topics based on topic themes. We present eight different methods of generating multidocument summaries and evaluate each of these methods on a large set of topics used in past DUC workshops. Our evaluation results show a significant improvement in the quality of summaries based on topic themes over MDS methods that use other alternative topic representations.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Algorithms, Performance, Experimentation

Additional Key Words and Phrases: Summarization, topic representations, topic themes

ACM Reference Format:

Harabagiu, S. and Lacatusu, F. 2010. Using topic themes for multi-document summarization. ACM Trans. Inform. Syst. 28, 3, Article 13 (June 2010), 47 pages.
DOI 10.1145/1777432.1777436. http://doi.acm.org/10.1145/1777432.1777436

1. THE PROBLEM

Today, information consumers are drowning in natural language text. While the Internet has increased access to text collections on a variety of topics, consumers now face a considerable amount of redundancy in the texts they encounter online. Now more than ever, consumers need access to robust multi-document summarization (MDS) systems, which can effectively condense information found in several documents into a short, readable synopsis, or summary.

Authors' address: Human Language Technology Research Institute and Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080; email: {sanda, finley}@hlt.utdallas.edu.

DOI 10.1145/1777432.1777436 http://doi.acm.org/10.1145/1777432.1777436

Recent work in multi-document summarization has leveraged information about the topics mentioned in a collection of documents in order to generate informative and coherent textual summaries. Traditionally, MDS systems have created informative summaries by selecting only the most relevant information for inclusion in a summary. In a similar fashion, coherent summaries have been created by ordering information extracted from texts in a manner that reflects the way it was originally expressed in a source document.

Our goal in this article is to provide a comprehensive explanation of how the choice of a topic representation impacts the quality—in terms of coherence, linguistic quality, and informativeness—of automatically generated summaries. While we assume that topic representations will continue to evolve as more and more sophisticated text understanding applications are developed we want to consider the properties of the ideal topic representation for MDS such that it best informs the components for information selection and information ordering simultaneously. Although a number of different topic representations [Lin and Hovy 2000; Harabagiu 2004; Barzilay and Lee 2004] have been proposed in the MDS literature, little work has specifically focused on how the choice of a particular topic representation (TR) impacts the quality of a generated summary.

In this article, we perform what we believe to be the first comprehensive evaluation of the impact that different topic representation techniques have on the performance of a multi-document summarization system. In order to identify the best possible topic representation for MDS, we consider a total of seven different topic representations of various levels of complexity, and use them in a total of forty different MDS methods. Five of these topic representations have been published before in the literature, while the sixth and seventh representations are novel approaches based on a new type of TR known as a *topic theme*. We define a *topic theme* as any type of knowledge representation that encodes some facet of the knowledge pertaining to a topic. In our work, we use a simple, easy to compute knowledge representation based on predicate-argument structures as our topic theme representation. We anticipate that other knowledge representations, if available, could be used as well. We expect that theme-based representations can be used to organize topic-relevant information from multiple sources, extracted from either, (1) a single sentence, (2) a cluster of sentences, (3) a discourse fragment, or even (4) a cluster of documents. In our work, we consider two topic structures based on themes: (a) a graph structure, in which themes are connected by binary relations; and (b) a linear structure, in which themes are ordered based on their position in texts, their density in the collection, and their connectivity.

The work presented in this article has two main goals. First, we introduce two novel topic representations that leverage sets of automatically-generated topic themes for MDS. We show how these new topic representations can be integrated into a state-of-the-art MDS system [Lacatusu et al. 2004] and provide quantitative results on the DUC[1] 2004 problem set.

[1] Document Understanding Conference; for more information, see the DUC Web site at http://duc.nist.gov.

Topic Representations

TR_1: Topic Signatures
TR_2: Enhanced Topic Signatures
TR_3: Topic Signatures Based on Document Structure
TR_4: Topics Represented as Content Models
TR_5: Topics Represented as Semantic Templates
TH_1: Topics Represented as Themes (organized in a Graph structure)
TH_2: Topics Represented as Themes (organized in a Linked List structure)

Sentence Extraction Methods

EM_1: Based on Topic Signature words (TR_1)
EM_2: Based on Enhanced Topic Signature relations (TR_2)
EM_3: Based on the Enhanced Topic Signature and Document Structure (TR_2 and TR_3)
EM_4: Based on the Semantic Templates (TR_5)
EM_5: Based on the Graph representation of Theme information (TH_1)
EM_6: Based on the Linked List representation Themes (TH_2)

Sentence Compression Methods

CM_0: No Compression
CM_1: Based on Basic Elements (BEs)
CM_2: Based on a word deletion function enhanced by Themes (TH)

Sentence Ordering Methods

OM_1: Based on grouping topically-relevant sentences
OM_2: Based on Content Models (TR_4)
OM_3: Based on the Linked List representation of Themes (TH_2)
OM_4: Based on the Graph Representation of Themes (T_1)

Fig. 1. Roadmap of algorithms.

Second, this article presents a comparison of 40 different configurations of an MDS system. We present experiments that investigate the use of six topic representation methods in six sentence extraction methods, two sentence compression methods, and four sentence ordering methods, over a total of 25 of the 50 DUC 2004 topics. (A roadmap of these algorithms is presented in Figure 1).

The remainder of the article is organized in the following way. Section 2 describes previous work on topic representations and their corresponding definition of topics. In Section 3, we motivate the need for topic themes and propose a theme-based representation. Section 4 describes how the six topic representations can be used to generate multiple-document summaries, whereas Section 5 presents and discusses the experimental results. Section 6 summarizes our conclusions.

2. TOPIC REPRESENTATION

Work in multi-document summarization (MDS) has long used automatically generated topic representations (TR) in order to approximate the information content of a collection of documents. While many of the top performing systems at the past DUC MDS evaluations have used TR to great effect, relatively little work has sought to identify the properties of an ideal TR for a generic MDS task. Since summaries need to be informative and coherent, we expect the ideal topic representation would need to capture both the most pertinent information

□ **Step 1:** Classify documents as relevant or nonrelevant according to the given topic. Since our experiments were performed on documents available from DUC, the relevant documents were given for each topic, and as non-relevant documents we selected the documents deemed relevant to the other 29 topics.
□ **Step 2:** Compute the value for $-2log\lambda$ as $-2log\frac{L(H_1)}{L(H_2)}$ for each term in the collection of relevant and non-relevant documents.
□ **Step 3:** Rank terms according to their $-2log\lambda$ value.
□ **Step 4:** Select first a confidence level from the χ^2 distribution table. Then determine the cutoff weight and number of terms to be included in the signature.

Fig. 2. The procedure of generating the topic signature TS_1.

about a topic, as well as the connections that are representative in a topic, such that an ideal order of presentation can be selected.

This section describes five of the topic representations that were used in MDS systems before. Each of them represents a different baseline. We describe them and show how they represent topics used in the Document Understanding Conference (DUC).

2.1 Topic Representation 1 (TR$_1$): Topic Signatures

In the earliest work on automatic topic representation, Lin and Hovy [2000] represent topics as weighted lists of topic-relevant terms, known as *topic signatures*, which could be considered to be generally relevant to any of the implicit topics mentioned in a collection of documents. In the notation of Lin and Hovy [2000], topic signatures were defined as $TS_1 = \{topic, < (t_1, w_1), \ldots, (t_n, w_n) >\}$, where the terms t_i were expected to be highly correlated to the topic, with an association weight w_i. Under this framework, topic signature terms are considered to be either stemmed content words, bigrams, or trigrams. Term selection and weight association are determined by the use of the *likelihood ratio* λ.

To find the candidate topic terms, a set of documents is preclassified into topic relevant texts, $\Re$, and topic nonrelevant texts, $\tilde{\Re}$. Based on this classification, the following two hypotheses can be made, as in Lin and Hovy [2000]:

- *Hypothesis 1* (H_1): $P(\Re|t_i) = p = P(\Re|\tilde{t_i})$ i.e. the relevance of a document is independent of t_i;
- *Hypothesis 2* (H_2): $P(\Re|t_i) = p_1 \neq p_2 = P(\Re|\tilde{t_i})$ i.e. the presence of t_i indicates strong relevance assuming $p_1 \gg p_2$;

Considering the following 2-by-2 contingency table:

	$\Re$	$\tilde{\Re}$
t_i	O_{11}	O_{12}
$\tilde{t_i}$	O_{21}	O_{22}

where O_{11} is the frequency of term t_i occurring in $\Re$, O_{12} is the frequency of term t_i occurring in $\tilde{\Re}$, O_{21} is the frequency of term $\tilde{t_i} \neq t_i$ occurring in $\Re$, O_{22} is the frequency of term $\tilde{t_i} \neq t_i$ occurring in $\tilde{\Re}$. The likelihood of both hypotheses is computed as:
$L(H_1) = b(O_{11}; O_{11} + O_{12}, p) \cdot b\big(O_{21}; O_{21} + O_{22}, p\big)$
$L(H_2) = b(O_{11}; O_{11} + O_{12}, p_1) \cdot b\big(O_{21}; O_{21} + O_{22}, p_2\big),$

where $b(k; n, x) = \left(\begin{array}{c} n \\ k \end{array}\right) x^k(1-x)^{n-k}$ represents the binomial distribution.

In our work, we generated topic signature terms (TR$_1$) using the four-step process outlined in Figure 2.

Figure 3 illustrates the topic signatures obtained for the DUC 2003 documents listed under the following four topics: T1 = Pinochet Trial, T2 = Leonid

TOPIC T1: PINOCHET TRIAL

term	Pinochet	Chile	Spanish	Chilean	immunity	arrest	diplomatic	British	Garzon	Argentina	human	dictator
weight	566.28	202.05	196.66	121.07	99.47	99.41	78.34	75.25	73.38	56.98	56.28	49.17

term	crime	London	Augusto Pinochet	Castellon	request	right	torture	authority	disappearance	magistrate	amnesty	Spain
weight	48.40	47.57	45.32	42.97	42.69	41.62	36.66	36.23	36.20	33.56	30.99	30.80

TOPIC T2: LEONID METEOR SHOWER

term	satellite	meteor	storm	Leonid	comet	shower	Earth	Tuttle	Sky	Kelly	meteorid	communication
weight	442.60	391.11	159.77	118.48	87.07	68.57	64.70	58.87	49.04	46.03	42.69	41.63

term	hour	Leo	constellation	Kirkhart	Temple	solar	fireball	peak	particle	printing	navigation	damage
weight	41.01	35.96	34.66	33.36	30.88	30.56	30.34	29.54	29.34	28.51	26.47	25.03

TOPIC T3: CAR BOMB IN JERUSALEM

term	Cabinet	attack	Palestinian	Israel	accord	assailant	Netanyahu	Jerusalem	explosion	peace	Israeli	shopper
weight	189.08	114.0	104.79	85.45	72.76	71.57	59.96	58.82	56.23	55.90	53.21	49.58

term	bomber	Arafat	Mahane Yehuda	explosive	suicide	agreement	militant	Hamas	market	Islamic	Jewish	bombing
weight	49.23	48.17	43.97	42.39	40.44	36.82	36.36	34.80	33.95	32.58	32.37	31.88

TOPIC T4: PAN–AM LOCKERBIE BOMBING TRIAL

term	Gadhafi	Libya	Libyan	suspect	Scottish	Lockerbie	Netherlands	Lockerbie, Scotland	Moammar Gadhafi	sanction	Cairo, Egypt
weight	246.50	195.9	184.03	96.05	91.71	77.81	56.94	54.27	50.49	47.30	44.96

term	U.N.	trial	Security Council	Lamen Khalifa Fhimah	bombing	Tripoli	ban	Abdel Basset Megrahi	accept	Egypt	Britain	PanAm
weight	37.6	32.2	32.19	30.03	29.94	27.66	27.4	26.43	26.10	25.7	24.00	22.36

Fig. 3. Examples of topic signature TS_1.

METEOR SHOWER, T3 = CAR BOMB IN JERUSALEM, and T4 = PANAM LOCKERBIE BOMBING TRIAL.[2]

2.2 Topic Representation 2 (TR$_2$): Enhanced Topic Signatures

In later work, Harabagiu [2004] introduced a new topic representation, known as *enhanced topic signatures* (TR$_2$), which used relations between topic-relevant terms (such as TR$_1$) to represent the topic of a collection of documents. This work featured two types of topic-relevant relations: (1) *syntax-based relations*, established between any two terminal nodes in a syntactic tree, which are immediately dominated by a nonterminal node (such as a head and its modifier or a preposition and its object), and (2) *context-based relations* (or *C-relations*), established between any two terminal nodes in a syntactic tree, which are not immediately dominated by the same nonterminal node.

The process of generating TR$_2$ begins with the selection of one or more *seed relations*. (Figure 4 details the procedure of generating a seed relation.) This process starts by using the CELEX database [Baayen et al. 1995] to perform morphological expansion of nouns and verbs (Step 1), followed by a semantic normalization (Step 2) using two sources of semantic information: (a) fifteen

[2] The DUC 2003 and 2004 MDS evaluations required sites to generate 100 word summaries from sets of newswire documents taken from popular news sources such as AP and NYT. Each set was associated with a topic label such as "Meteor Shower," which served to identify the topic.

□ **Step 1:** Morphological expansion of nouns and verbs from TS_1.
□ **Step 2:** Semantic normalization of nouns based on (a) name classes, and (b) ontological concepts.
□ **Step 3:** Count all syntactic relations between a verb/nominalization and nouns.
□ **Step 4:** Select the relation with the highest count.

Fig. 4. The procedure of selecting a seed relation for a topic.

Topic	*Seed Relation*
T1 = PINOCHET TRIAL	[question – PERSON]
T2 = LEONID METEOR SHOWER	[meteor – storm]
T3 = CAR BOMB IN JERUSALEM	[peace – accord]
T4 = PAN–AM LOCKERBIE BOMBING TRIAL	[NUMBER – suspect]

Fig. 5. Seed relations.

different name classes,[3] and (b) hypernymy relations derived from WordNet [Fellbaum 1998]. The selection of the seed relation is based on the discovery of the syntactic relation that most frequently connected concepts from the same class in the topic-relevant documents. Syntactic relations are automatically recognized with the publicly available chunk parser reported in Kudo and Matsumoto [2003].

Figure 5 presents examples of some of the seed relations discovered for some of the topics evaluated in DUC 2003.[4]

Step 3 generates candidate relations by considering both syntactic relations and C-relations. First defined in Harabagiu [2004], C-relations provide a mechanism for capturing potentially informative dependencies between pairs of terms from a sentence that may not be syntactically dependent on one another. In contrast, syntactic relations are deemed from any pair of terms that can be interrelated, given the structure of a syntactic tree, regardless of whether they belong to the same topic. Figure 6 provides examples of C-relations and syntactic relations extracted from the same passage. Syntactic relations are discovered again by using a chunk parser [Kudo and Matsumoto 2003]. C-relations are generated by considering a salience window of two sentences preceding and succeeding each topic-relevant verb from the document. (Topic-relevant verbs are available from TS_1.) Both syntactic relations and C-relations are normalized semantically, in the same way as in Step 2 of Figure 4.

Step 4 ranks relations based on their *Frequency* and *Relevance-Rate*, following a method introduced in Riloff [1996]. The *Frequency* of an extracted relation counts the number of times the relation is identified in the relevant documents. In a single document, one extracted relation may be identified multiple times. The *Relevance-Rate* = *Frequency*/*Count*, where *Count* measures the number of times an extracted relation is recognized in any document considered. Relations with *Relevance-Rate* $< \alpha$ are discarded as nonrelevant. Additionally, we only

[3] The name classes that are recognized are: People, Organizations, Locations, Dates, Times, Quantities, Money, Addresses, Phone Numbers, Passport Numbers, Airport Codes, Disease names, Medicine names, Chemical Components, Stock Exchange Names.

[4] All of these relations are instances of C-relations.

A car bomb blew up Friday in a Jerusalem market crowded with Israelis shopping for the Sabbath. The blast killed two assailants, wounded 21 Israelis and prompted Israel to suspend implementation of the peace accord with the Palestinians.

Syntactic Relations	*C-Relations*
killed – blast; killed – two assailants	kill – car bomb

Fig. 6. Examples of candidate relations.

```
□ Step 1: Select a topic relation seed; $Nr\_Iterations = 0$.
□ Step 2: Retrieve documents relevant to this topic relation seed.
□ Step 3: Generate candidate topic relations.
□ Step 4: Rank candidate topic relations.
□ Step 5: Select a new topic relation.
□ Step 6: If there are no new relations, or $Nr\_Iterations > 100$ then STOP.
□ Step 7: Else $Nr\_Iterations++$; Go To Step 2.
```

Fig. 7. Discovery of topic-relevant relations.

maintain relations with $\beta < Count/MaxCount < \gamma$, where *MaxCount* indicates the total number of instances observed for the most common relation.[5]

Step 5 selects a new topic relation by using the order produced in Step 4. Only the first relation is selected and added to the set of discovered relations. The rank of the relation represents its weight in the new topic signature TS_2. Steps 6 and 7 resume the discovery of the new topic relations if a new topic relation was selected at Step 5, performing the same sequence of operations as were performed on the seed topic relation.

The relations and their ranks constitute the new topic signature, TS_2, discovered in the procedure illustrated in Figure 7. Figure 8 illustrates the enhanced topic signatures (TS_2) obtained for the documents from DUC-2003 listed under the four topics, T1 = Pinochet Trial, T2 = Leonid Meteor Shower, T3 = Car Bomb in Jerusalem, and T4 = PanAm Lockerbie Bombing Trial. It can be noted that the TS_2 illustrated in Figure 8 uses some of the terms from TS_1, illustrated in Figure 3, but also some new terms, which were discovered by C-relations.

2.3 Topic Representation 3 (TR$_3$): Topic Signatures Based on Document Structure

The third TR we consider (TR$_3$) uses automatically recognized discourse segments (such as *TexTiles* [Hearst 1997]) as the unit of representation, as described in Harabagiu [2004]. Documents can be segmented into a set of units that can be assumed to correspond to the document structure. The TextTiling algorithm described in Hearst [1997] automatically determines the document structures as multiparagraph segments.[6]

When each segment is labeled by a topic-relevant concept, such structures give rise to a new topic representation in the form of signatures $TS_3 = \{topic, < (Tl_1, r_1), \ldots, (Tl_s, r_s) >\}$, where each label Tl_i is given a ranking weight r_i. To

[5] We used $\alpha = 0.7$, $\beta = 0.01$, and $\gamma = 0.4$. These values were set manually based on experiments conducted with a development set consisting of a small sample of DUC 2005 and 2006 topics.

[6] We used the TexTiling algorithm with the default parameters: $w = 20$ tokens per sequence, $k = 4$ sequences per block.

TOPIC T1: PINOCHET TRIAL

relation	[question – PERSON]	[request – question]	[genocide – terrorism]	[PERSON – regime]
weight	16.12	14.85	12.11	12.11

TOPIC T2: LEONID METEOR SHOWER

relation	[meteor – storm]	[Leonid – storm]	[satellite – company]	[constellation – Leo]	[present – target]	[NUMBER – meteor]
weight	34.80	24.23	21.61	16.39	16.39	13.80

relation	[NUMBER – satellite]	[printing – plant]	[meteor – trail]	[damage – satellite]	[Leonid – meteor]
weight	13.80	13.80	13.80	11.24	11.24

TOPIC T3: CAR BOMB IN JERUSALEM

relation	[peace –accord]	[Israeli – Cabinet]	[condemn – attack]	[ORG – millitant]	[NUMBER – Israeli]	[Mahane Yehuda – market]
weight	32.05	17.87	15.07	12.29	12.29	12.29

TOPIC T4: PAN–AM LOCKERBIE BOMBING TRIAL

relation	[NUMBER – suspect]	[ORGANIZATION – sanction]	[diplomat – say]	[leader – Moammar Gadhafi]	[Libyan – want]
weight	26.03	21.84	16.24	15.93	13.01

relation	[NUMBER – Libyan]	[PERSON – arrive]	[air – embargo]	[NUMBER – man]	[DATE – bombing]
weight	13.01	11.75	11.75	11.43	10.92

Fig. 8. Examples of topic signature TS_2.

assign labels to segments we used topic information provided by TS_1 and TS_2. In selecting the labels for text segments, we have encountered three cases:

Case 1. A single topic-relevant relation is identified in the segment. For example, for the topic $T_1 = \{$ARREST OF AUGUSTO PINOCHET$\}$ only the relation [*charge-murder*] is discovered. The segment is labeled by the nominalization of the verb, CHARGES. For verbs or nominalizations that express communication events, the label is a multiword expression, which also considers the theme of the communication (e.g. PEACE TALKS).

Case 2. Multiple topic-relevant relations are recognized in the segment. For example, if both the relation [*arrest-Pinochet*] and [*charge-murder*] are recognized, the nominalization/verb corresponding to the highest ranked relation becomes the label of the segment, e.g. ARREST.

Case 3. The segment contains topic-relevant terms derived from TS_1, but no topic relation. In this case the most relevant noun becomes the label of the segment, e.g. IMMUNITY.

To be able to produce the ranking of segment labels, we have used the following formula, as defined in Harabagiu [2004]:

$$R(Tl_i) = \begin{cases} w(r_{high}) & \text{when topic relations are recognized} \\ \sum_{t_k \in S_i} \frac{w(t_k)}{w(t_1)} + w(r^D_{low}) & \text{when no topic relations are recognized,} \end{cases}$$

where $w(r_{high})$ represents the weight in TS_2 of r_{high}, the relation with the highest weight in the segment; $w(r^D_{low})$ represents the lowest weight of a topic relation in the document, and $w(t_1)$ is the weight of the first term from TS_1, while $t_k \in S_i$ are terms from TS_1 recognized in segment S_i.

TOPIC T1: PINOCHET TRIAL

theme	arrest	immunity	investigation	genocide	amnesty	extradition
weight	99.41	99.47	13.84	26.30	30.99	23.33

TOPIC T2: LEONID METEOR SHOWER

theme	meteor shower	comet	damage	observation	communications
weight	68.54	87.08	25.03	18.44	41.63

TOPIC T3: CAR BOMB IN JERUSALEM

theme	explosion	peace talks	Hamas	Jerusalem	market
weight	56.23	55.90	34.80	58.82	33.95

TOPIC T4: PAN–AM LOCKERBIE BOMBING TRIAL

theme	suspect	victims	bombing	sanctions	trial
weight	96.05	22.36	29.94	47.30	32.20

Fig. 9. Topic signature TS_3 for topics T1, T2, T3, and T4.

Figure 9 illustrates the enhanced topic signatures, TS_3, for the documents corresponding to the following four topics: T1 = Pinochet Trial, T2 = Leonid Meteor Shower, T3 = Car Bomb in Jerusalem, and T4 = PanAm Lockerbie Bombing Trial.

2.4 Topic Representation 4 (TR$_4$): Topics Represented as Content Models

The fourth method of topic representation we consider follows Barzilay and Lee [2004] in using an iterative reestimation procedure in order to model the topic of a collection of documents.

As described in Barzilay and Lee [2004], this approach alternates between, (1) creating clusters of text spans with similar word distributions, and (2) computing models of word distributions and topic changes from the clusters obtained. The working assumption is that all texts describing a given topic are generated by a single *content model*. The content model is a Hidden Markov Model (HMM) wherein states correspond to subtopics and state transitions capture either orderings within that domain, or the probability of changing from one given subtopic to another. This model was trained for each topic on the set of documents to be summarized. More details about the the training and creation of these models can be found in Barzilay and Lee [2004].

The generation of TR$_4$ can be performed using the following three-step process:

□ **Step 1**. *Initial topic induction*, in which complete-link clustering is used to create sentence clusters by measuring sentence similarity with the cosine metric. All clusters smaller than a given threshold are merged into an "etcetera" cluster.
□ **Step 2**. *Model states and emission/transition probabilities*. Each cluster corresponds to a state. For each state s_i corresponding to cluster c_i the sentence emission probabilities are estimated using smoothed counts:

$$p_{s_i}(w'|w) = \frac{f_{c_i}(ww') + \delta_1}{f_{c_i}(w) + \delta_1|V|},$$

where $f_{c_i}(y)$ is the frequency with which word sequence y (e.g. $y = ww'$ or $y = w$) occurs within the sentences in cluster c_i, δ_1 is a smoothing parameter, and V is the vocabulary. When estimating the state-transition probabilities, if two clusters c_1 and c_2 are considered, $D(c_1, c_2)$ represents the number of documents in which a sentence from c_1 immediately precedes a sentence from c_2. $D(c_i)$ is the number of documents containing sentences from cluster c_i. For two states s_i and s_j, the probability of transitioning from s_i to s_j is estimated as:

$$p(s_j|s_i) = \frac{D(c_i, c_j) + \delta_2}{D(c_i) + \delta_2 m},$$

where m is the number of states and δ_2 is a smoothing constant.

☐ **Step 3**. *Viterbi re-estimation.* In this step, the *model parameters* are re-estimated using an EM-like Viterbi approach: the sentences are re-clustered by placing each sentence in a new cluster c_i that corresponds to a state s_i most likely to have generated it according to the Viterbi decoding of the training data. The new clustering is used as input to the procedure of estimating the HMM parameters. The cluster/estimate cycle is repeated until the clustering stabilizes or the iterations reach a predefined number of cycles.

Figure 10 illustrates samples from clusters corresponding to topic T_1. The clusters and the relations between them dictated by the HMM represent the topic representation TR_4. The unlabeled clusters from this representation indicate the different aspects of the topic. The weights associated with the relations between the clusters in Figure 10 represent transition probabilities in the HMM.

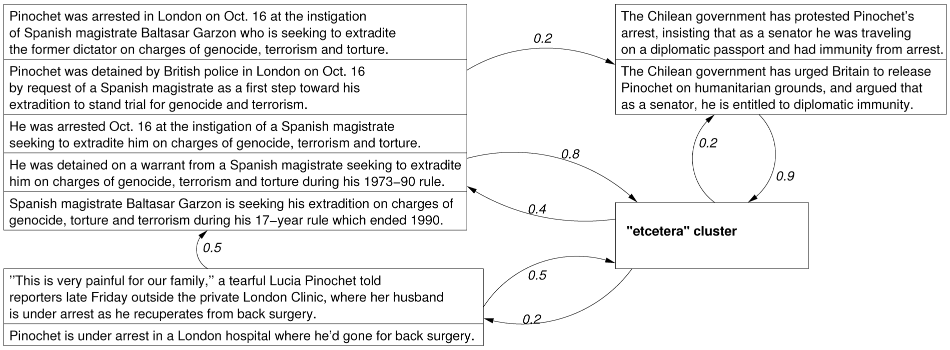
2.5 Topic Representation 5 (TR$_5$): Topics Represented as Semantic Templates

The idea of representing a topic as a frame-like object was first discussed in DeJong [1982], where underspecified (or “sketchy”) scripts were used to model a set of predefined particular situations, for example, DEMONSTRATIONS, EARTHQUAKES, or LABOR STRIKES.

Such frame-like objects were used for the past two decades in Information Extraction (IE), under the name of templates [Grishman and Sundheim 1996; Hirschman et al. 1999]. The idea is that if a frame-like template represents a topic, IE methods can identify all topic-relevant information from a set of documents. More recently, the ICSI Berkeley FrameNet project [Baker et al. 1998] has produced more than 825 semantic frames based on Charles Fillmore’s definition of semantic frames. The SemEval 2007 [SemEval 2007] and Senseval 2004 [SENSEVAL-3 2004] evaluations have shown that automatic labeling of semantic frames can be produced with good accuracy. Figure 11(a) shows the FrameNet definition of the TRIAL frame. Figure 11(a) also illustrates the Frame Elements (FEs) of this frame, which are recognized in a text relevant for the topic T_1 = PINOCHET TRIAL.

We have employed a two-tiered approach to recognizing frames (and frame elements) in texts. First, we have used a semantic parser developed for the SemEval-2007 evaluation, described in Bejan and Hathaway [2007]. More than one frame can be recognized in a text. Since FrameNet also encodes relations between frames, a set of interrelated frames may be used to represent a topic. Figure 11(b) illustrates several interrelated frames encoded in FrameNet. They may be used for representing information pertaining to topic T1 = PINOCHET TRIAL.

However, semantic frames have not yet been created for every topic. For example, the FrameNet frame SHOWER cannot be used for the topic LEONID METEOR SHOWER, as it captures only weather-related semantics. Moreover, no frame exists for METEORS or other extraterrestrial objects.


Pinochet was arrested in London on Oct. 16 at the instigation of Spanish magistrate Baltasar Garzon who is seeking to extradite the former dictator on charges of genocide, terrorism and torture.

Pinochet was detained by British police in London on Oct. 16 by request of a Spanish magistrate as a first step toward his extradition to stand trial for genocide and terrorism.

He was arrested Oct. 16 at the instigation of a Spanish magistrate seeking to extradite him on charges of genocide, terrorism and torture.

He was detained on a warrant from a Spanish magistrate seeking to extradite him on charges of genocide, terrorism and torture during his 1973–90 rule.

Spanish magistrate Baltasar Garzon is seeking his extradition on charges of genocide, torture and terrorism during his 17–year rule which ended 1990.

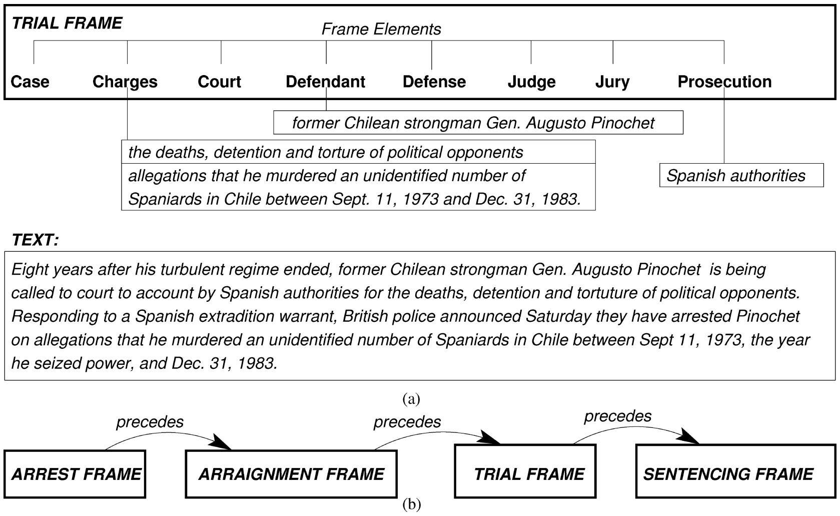
The Chilean government has protested Pinochet's arrest, insisting that as a senator he was traveling on a diplomatic passport and had immunity from arrest.

The Chilean government has urged Britain to release Pinochet on humanitarian grounds, and argued that as a senator, he is entitled to diplomatic immunity.

"This is very painful for our family," a tearful Lucia Pinochet told reporters late Friday outside the private London Clinic, where her husband is under arrest as he recuperates from back surgery.

Pinochet is under arrest in a London hospital where he'd gone for back surgery.

Fig. 10. Topic representation TR_4.


TRIAL FRAME
Frame Elements
Case
Charges
Court
Defendant
Defense
Judge
Jury
Prosecution
former Chilean strongman Gen. Augusto Pinochet
the deaths, detention and torture of political opponents
allegations that he murdered an unidentified number of Spaniards in Chile between Sept. 11, 1973 and Dec. 31, 1983.
Spanish authorities
TEXT:

Eight years after his turbulent regime ended, former Chilean strongman Gen. Augusto Pinochet is being called to court to account by Spanish authorities for the deaths, detention and tortuture of political opponents. Responding to a Spanish extradition warrant, British police announced Saturday they have arrested Pinochet on allegations that he murdered an unidentified number of Spaniards in Chile between Sept 11, 1973, the year he seized power, and Dec. 31, 1983.

(a)

(b)

Fig. 11. (a) Example of semantic frame recognized for topic T1 = Pinochet Trial; (b) Interrelated frames.

To be able to address this problem, we have used the method described in Harabagiu and Maiorano [2002], which automatically generates ad hoc frames. This method uses lexical relations encoded in the WordNet lexical database to mine relationships among topic-relevant semantic concepts.

In order to create topic representations based on frames, we first discover the topic-relevant information from WordNet and then use these concepts to generate ad hoc frames.

Discovery of topic-relevant information. The generation of ad hoc frames requires knowledge of topic-relevant concepts. This knowledge is mined automatically from WordNet. The building block of the WordNet lexical database is the *synset*, or the set of synonyms. WordNet encodes a vast majority of English nouns, verbs, adjectives, and adverbs and groups them in synsets. The WordNet 2.1 database additionally encoded a new relation, called Topic. Topic relations link words such as "trial," "prosecution," "defendant," to synsets representing

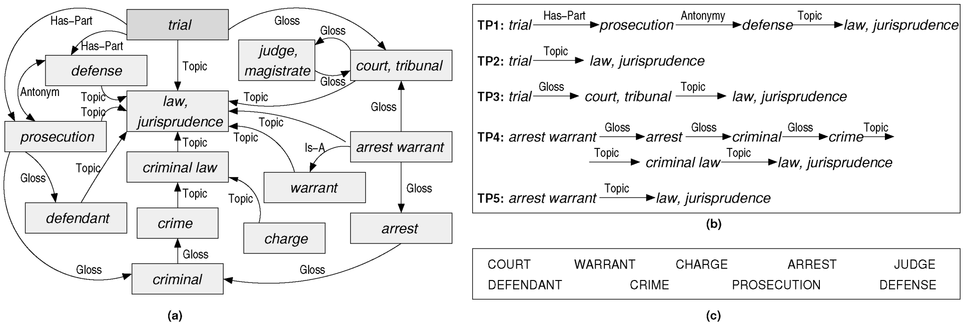

trial
Has–Part
Has–Part
defense
Antonym
prosecution
Topic
Topic
Topic
law, jurisprudence
Gloss
judge, magistrate
Gloss
Gloss
court, tribunal
Topic
Topic
Topic
Topic
Gloss
arrest warrant
Is–A
warrant
criminal law
Topic
Topic
Gloss
defendant
crime
charge
Topic
arrest
Gloss
Gloss
criminal
Gloss
Gloss
(a)
TP1: trial Has–Part prosecution Antonymy defense Topic law, jurisprudence
TP2: trial Topic law, jurisprudence
TP3: trial Gloss court, tribunal Topic law, jurisprudence
TP4: arrest warrant Gloss arrest Gloss criminal Gloss crime Topic
Topic criminal law Topic law, jurisprudence
TP5: arrest warrant Topic law, jurisprudence
(b)
COURT WARRANT CHARGE ARREST JUDGE
DEFENDANT CRIME PROSECUTION DEFENSE
(c)

Fig. 12. (a) Example of Semantic Graph containing topical paths; (b) Topical paths between a pair of concepts; (c) Example of slots of an ad hoc template. The topical paths and the ad hoc template correspond to the topic T_1 = PINOCHET TRIAL.

a certain domain, for example {*law, jurisprudence*}. In the current version of WordNet, there are 413 topic synsets, and a total of 6322 topic relations.

In order to discover topic-relevant concepts, we navigate the WordNet graph by uncovering *topical paths* created by at most four combinations of WordNet relations (e.g. IS-A, HAS-PART, ENTAIL, TOPIC, ANTONYM) and GLOSS[7] relations. The starting point is any noun concept, not a name, from the name of the topic. For the topic T_1, for example, the noun “trial” is the starting point. The graph upon which the navigation occurs for topic T_1 is represented in Figure 12(a). The graph is traversed by several topical paths. Two principles guide the uncovering of topical paths:

Principle 1. Redundant connections rule out a connection discovered by accident. Therefore, if at least two different paths of WordNet relations can be established between any two synsets, they are likely to be part of the representation of the same topic.

Principle 2. The shorter the paths, the stronger their validity. Consequently, we rule out paths of length larger than 5.[8] This entails the fact that each topic may be represented by at least five synsets.

The TOPIC relation in WordNet has the purpose of identifying the most relevant synsets to a given topic. By allowing the topical paths to be discovered, some new synsets that are relevant to the topic are uncovered. Figure 12(a) illustrates the topical paths determined when starting at the topic synset {*trial*} and ending at synset {*law, jurisprudence*}, assembled in a semantic graph. Figure 12(b) illustrates five topical paths. The first three paths connect synsets {*trial*} and {*law, jurisprudence*}. Concepts such as PROSECUTION, DEFENSE, or COURT are uncovered by these paths. The navigation is also permitted from other synsets that are related through TOPIC links. Topical paths $TP4$ and $TP5$ connect synsets {*arrest*} and {*warrant*} to synset {*law, jurisprudence*},

[7] We consider a GLOSS relation between each element of a synset and each content word from the gloss that defines the synset.

[8] In Harabagiu [1997] it has been shown the WordNet connectivity allows a path of at least 4 length between any pair of synsets.

□ **Step 1:** Identify all sentences in which words belonging to at least two synsets from a topical path are present. The semantic sense of the words is irrelevant.

□ **Step 2:** Identify all Subject-Verb-Object (SVO) triplets and prepositional attachments in which one of the topical concepts is used. For this purpose, we used the same chunk parser as those used for generating topic representation TR_2.

□ **Step 3:** Each noun from the structure identified at Steps 2 and 3 is classified in one of the following three classes: KNOWN, RELATED, or UNKNOWN. All nouns that have a semantic sense in the same hierarchy as any of the nouns from the topical paths are considered KNOWN. All nouns that have a semantic sense that can be related through at most four WordNet relations to any of the concepts from the topic paths are considered RELATED. All other nouns are considered UNKNOWN.

□ **Step 4:** We use the classification of nouns into the classes KNOWN, RELATED, and UNKNOWN for selecting the slots of the ad hoc template. For each concept from a topical path we compute three values: (1) *KFreq*: the number of nouns that are classified as KNOWN because they belong to the same WordNet hierarchy as the topical concept; (2) *RFreq*: the number of distinct nouns classified as RELATED because of the existence of a semantic chain to one of the concepts from the WordNet hierarchy of the topical concept; (3) *UFreq*: the number of distinct nouns classified as UNKNOWN because they were discovered due to the topical concept. A fourth value, *Afreq* represents the number of distinct nouns identified in the document collection. Similarly to Riloff and Schmelzenbach [1998], we define the probability that a topical concept defines a relevant slot in the ad hoc template as $PRel = \frac{KFreq}{KFreq+RFreq} > P = \frac{KFreq+RFreq+UFreq}{AFreq}$. A synset is selected as a slot of an ad hoc template if its $PRel > P$.

Fig. 13. The procedure of selecting the slots for TR_5.

uncovering additional topic relevant concepts, for example, ARREST, CRIME. In FrameNet, such concepts represent FEs of the ARREST frame, which is related to the TRIAL frame, as illustrated in Figure 11(b). All such concepts become candidate slots for the ad hoc frame. Figure 12(c) illustrates some of them.

Generation of ad hoc *frames*. The slots of an *ad hoc* frame are selected from all synsets that participate in some topical path. The selection procedure is represented in Figure 13.

The slots of the ad hoc frames are filled with the text spanned by subjects, objects, or prepositional attachments where the slot-concept is identified in Step 2 of the procedure illustrated in Figure 13.

3. A TOPIC REPRESENTATION BASED ON THEMES

The topic representations examined in Section 2 (e.g. TR_1–TR_5) differ both in terms of, (1) their *granularity*, and (2) their *connectivity*. In this section, we define granularity in terms of the manner and type of information unit (e.g. terms, relations, discourse fragments, themes) used in a TR. Connectivity is defined as the degree to which elements in a TR are linked to one another by one or more different types of relations.

In Section 2, we considered five different baseline topic representations that use terms (TR_1), relations (TR_2), document segments (TR_3), sentence clusters (TR_4), or semantic frames generated from multiple different documents (TR_5). Topics however, can rarely be considered in isolation. In order to fully characterize the set of topics derivable from a collection of documents, we expect that systems must also identify some subset of the relevant connections that exist between different topical elements. Many current TR methods capture few if any of the connections between topic elements. For example, TR_1 and TR_2 exhibit no interconnectivity, whereas TR_3 and TR_4 use a connectivity based on sentence cohesion and TR_5 is based on WordNet and FrameNet connectivity.

We believe the main drawback of the five baseline TRs that we consider in this article (TR_1–TR_5) stems largely from the lack of connectivity within the topic representation, although they use sophisticated language processing at syntactic and semantic level. To address this problem, we have considered a topic representation that:

(a) captures different aspects of a topic, that we call *topic themes*;
(b) uses semantic information that has more coverage than the current FrameNet database; and
(c) uses connections between frames motivated by discourse information.

In the rest of this section, we will revisit the notion of theme representation and propose a more linguistically-motivated definition and representation of themes.

3.1 Conceptual Representations of Themes

We believe that every aspect of a (set of) topic(s) can be represented by some semantic information that is conveyed by predicate-argument structures. This conceptual representation is similar to the dependency-based representation employed in Barzilay et al. [1999]. It consists of:

—the common predicate,
—the set of semantically consistent arguments, and
—the arguments that anchor the predicate in time and location, or describe the cause, manner or effect of the predicate.

Predicates and arguments are recognized by semantic parsers trained on PropBank[9] annotations [Palmer et al. 2005]. (While there are now several such parsers publicly available (e.g. Pradhan et al. [2005]; Surdeanu and Turmo [2005]), we used the parser reported in Moschitti and Bejan [2004] in all of our experiments.) Predicate-argument parsers recognize (a) verbal predicates, and (b) their arguments, which are labeled Arg0 to Arg5 or with some functional types. Typically, Arg0 denotes an argument that can be assigned an *agent* (or *experiencer*) thematic role, Arg1 for *patient* or *theme*, whereas Arg2 represents *recipient*, *benefactive*, or *instrument*. Additionally, the argument list may include functional tags from Treebank,[10] for example, ArgM-DIR indicates a directional, ArgM-LOC indicates a locative, and ArgM-TMP stands for a temporal, while ArgM-ADV is a subtype of general purpose used for encoding the adverbial modifier. Figure 15 illustrates the predicate-argument structures recognized in the sentences illustrated in Figure 14. It can be noted that in the same sentence, predicate-argument structures are connected. There are two types of connections, (1) nesting connections, identified when an entire predicate structure becomes the argument, or part of the argument, of another

[9] www.cis.upenn.edu/∼ace
[10] www.cis.upenn.edu/∼treebank

S_1: British police said Saturday they have **arrested** former Chilean dictator Gen. Augusto Pinochet on allegations of murdering Spanish citizens during his years in power.
S_2: Responding to a Spanish extradition warrant, British police announced Saturday they have **arrested** Pinochet on allegations that he murdered an unidentified number of Spaniards in Chile between Sept. 11, 1973, the year he seized power, and Dec. 31, 1983.
S_3: Pinochet, 82, was **placed under arrest** in London Friday by British police acting on a warrant issued by a Spanish judge.

Fig. 14. A collection of similar sentences.

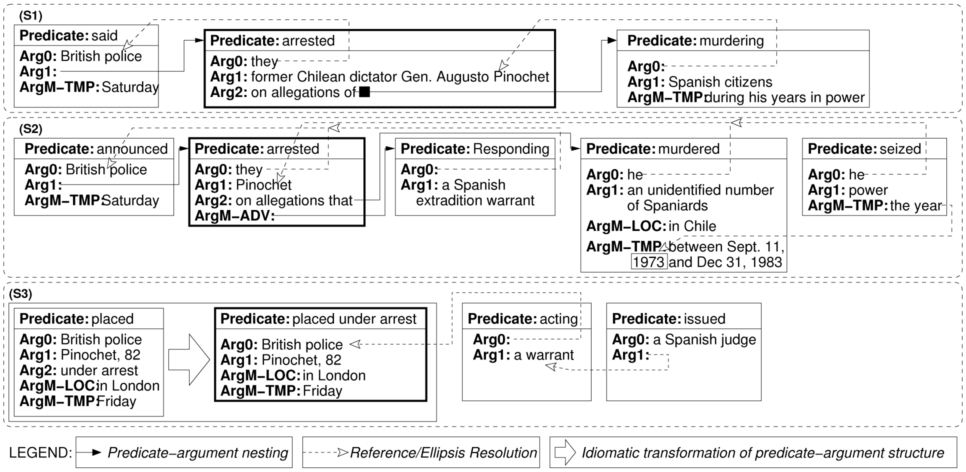


Fig. 15. The predicate-argument structures of the theme illustrated in Figure 14.

predicate; and (2) connections between arguments recognized by reference or ellipsis recognition. Additionally, some predicates are recognized as idiomatic constructions, and they need to be transformed to capture the correct idiom. The recognition of idiomatic predicates is processed as a paraphrase recognition problem.

Several predicate-argument structures can be used to create a conceptual representation of a theme only if, (a) they have a common predicate (or paraphrased predicates); and (b) their arguments are semantically consistent. We granted semantic consistency of arguments only when they refer to the same objects. To identify the common reference, we had to rely on coreference connections identified by coreference resolution software. We have used the coreference resolution method reported in Nicolae and Nicolae [2006], but any other coreference resolution system could have been used as well, for example, Ng [2004].

The conceptual representation of a theme indicates, (1) the common predicate, its frequency in the sentence cluster, as well as the cluster size; (2) the consistent arguments, their textual span and their position in the cluster, marked by, (a) the sentence where the argument was discovered, (b) the word count within the sentence where the argument starts, and (c) the word number where the argument ends; (3) grounding arguments, for example, ArgM-LOC, ArgM-TMP, and ArgM-ADV.

Predicate: Arrested (Placed Under Arrest) cluster size:38 frequency:15)
Arg0: (S_3,12,13) British police (S_1,4,4) they (S_2,11,11) they
Arg1: (S_1,7,12) former Chilean dictator Gen. Augusto Pinochet
(S_3,0,2) Pinochet, 82 (S_2,14,14) Pinochet
Arg2: (S_2,15,43) on allegations that he murdered an unidentified number of Spaniards in Chile between Sept. 11, 1973, the year he seized power, and Dec. 31, 1983
(S_1,13,23) on allegations of murdering Spanish citizens during his years in power
ArgM-LOC: (S_3,8,9) in London
ArgM-TMP: (S_3,10,10) Friday
ArgM-ADV: (S_2,0,5) Responding to a Spanish extradition warrant

Fig. 16. Conceptual representation of a theme.

Phase I: Discovering Topic Themes

Step 1: *Identify the predicate–argument structures of each sentence in the document collection*
Step 2: *Cluster all sentences that, (a) have at least one common predicate; and*
(b) the arguments of the common predicate are semantically consistent
Step 3: *Generate conceptual representations for each theme*
Step 4: *Select topic themes based on multiple evidence sources*

Phase II: Discovering the Topic Structure based on Themes

Step 5: *Identify relations between themes based on cohesion and coherence information*
Step 6: *Generate the topic structure based on the relations between themes*
Step 7: *Evaluate the relevance of the topic themes based on the topic structure*

Fig. 17. Procedure for generating topic representations based on themes.

Figure 16 illustrates the conceptual representation of a theme derived from the sentence cluster illustrated in Figure 14.[11] The conceptual representations of themes may be considered as building blocks of a new topic representation. To be able to represent topics, we also need to establish which themes should be selected, and what connections exist between them.

3.2 Using Themes for Representing Topics

A topic representation based on themes can be created in two phases of processing, as illustrated in Figure 17. Our method for discovering topic themes is based on the first four steps of the procedure illustrated in Figure 17.

Step 1. identifies the predicate-argument structures. We use a semantic parser (trained on PropBank) to identify all of the predicates and their corresponding arguments.

Step 2. concerns the clustering of sentences for topic themes. In order to capture semantic similarity between pairs of sentences that may convey similar meanings (but include different lexical items), we have followed Hickl et al. [2006] in using a classifier-based paraphrase recognition method in order to identify pairs of predicate-argument structures that are high-quality paraphrases of one another.[12]

The recognition of predicate paraphrases improves the identification of sentences that have a common predicate, and semantically consistent arguments.

[11] The numbers in parentheses refer to the start and end offsets for the tokens corresponding to each argument in each sentence.

[12] The classifier features we use are the same as those reported by Hickl et al. [2006].

Table I. The Number of Predicates and Themes Generated for T_1, T_2, T_3, and T_4

Topic	Predicates	Themes
T_1 = PINOCHET TRIAL	973	853
T_2 = LEONID METEOR SHOWER	1270	1063
T_3 = CAR BOMB IN JERUSALEM	963	780
T_4 = PANAM LOCKERBIE BOMBING TRIAL	891	767

Semantic consistency of arguments can be checked by cross-document coreference because the sentences from the cluster originate in different documents. We have approximated cross-document coreference in the following way: (1) we have considered only within-document coreference; and (2) we checked semantic consistency only between the semantic classes of the arguments, using the WordNet information. For example, in Figure 15, the argument Arg1 of the predicate "murdering" in S_1 is consistent with Arg1 of the predicate "murdered" in S_2 because both "Spanish citizens" and "Spaniards" are mapped in the same WordNet synset, defined as native or inhabitant of Spain.

Step 3. generates the conceptual representation of themes by considering all consistent arguments of the predicates as well as the arguments that are recognized in only one of the sentences.

Step 4. involves the selection of the most representative themes from the potentially large number of theme representations generated in Steps 1, 2, and 3.

The first three steps produce a large number of conceptual representations. Table I lists the number of predicates and the number of themes produced for the representations of the four topics, T_1 = PINOCHET TRIAL, T_2 = LEONID METEOR SHOWER, T_3 = CAR BOMB IN JERUSALEM, and T_4 = PANAM LOCKERBIE BOMBING TRIAL.

In our work, we have cast the selection of candidate themes as a binary classification problem that can be solved through inductive methods. We have implemented the classification with decision trees, considering eight features for each candidate theme, as illustrated in Table II. Training data for this selection was provided by three human annotators, who judged sentences associated with a set of 25 DUC 2003 topics as either relevant or irrelevant to the topic.

The eight features used for training the classifiers use (a) frequency and relevance information employed in Information Retrieval (IR) (F_1, F_8); (b) statistical relevance information for the topic (F_2, F_3, F_4); (c) position information (F_5); and (d) information obtained directly from other representations of themes (F_6 and F_7).

Table III illustrates the values of these features for the theme represented in Figure 16. When this classifier was trained, it selected, on average, 33% of the candidate themes produced at Step 3.

Step 5. identifies relations between the selected themes.

We anticipate that themes can be linked by two kinds of relations, including (1) cohesion relations, and (2) coherence relations. As with natural language discourse, we assume that the principles of cohesion and coherence that hold between discourse segments can also be applied to topic themes derived from the content of a collection of documents.

Table II. Features Used in the Selection of Candidate Themes

F_1	*Theme frequency*. This feature is computed as the frequency of the theme in the topic normalized by the sum of the frequencies of all the themes in the topic.
F_2	*Theme coverage*. F_2 is inspired from the TF-IDF measure from information retrieval and is computed as: $Theme\_frequency \times \log \frac{\lvert T \rvert}{Inverse\_theme\_frequency}$, where *Theme_frequency* is the unnormalized theme frequency, $\lvert T \rvert$ is the number of themes in the topic, and *Inverse_theme_frequency* is the number of themes containing the same predicate as the current theme.
F_3	*Predicate signature*. Its value is the weight of the theme's predicate in the topic signature TS_1.
F_4	*Argument signature*. It is a vector having the length equal to the number of arguments of the theme. The value of the index i of this vector corresponds to the weight in the TR_1 of the head of Arg_i.
F_5	*Theme relation signature*. It is a vector having the length equal to to the number of arguments of the theme. The value of the index i of this vector corresponds to the weight of the relation between the theme predicate and the argument arg_i of the theme in TR_2. Topic signature TS_2 has an important role in the topic representation, and we use it for feature F_5. It is similar to F_4, but for each argument we store in the vector the TS_2 weight of the relation between the predicate and the argument. Since the values for this set of features can be sparse, we chose to add to the vector another value representing the sum of the weights of all the predicate-argument relations that appear in the theme.
F_6	*Position*. Inspired by the results from single-document summarization, where the position of the sentence within the document plays an important role in the selection of the sentence to appear in the summary, we added a new feature F_6 that represents the average position of the theme in the set of documents, normalized by the average length of the documents.
F_7	*Content signature*. This feature models the TF-IDF measure and is computed as $\frac{\max_i \lvert theme\_sentences \cap C_i \rvert}{\lvert C_{\arg\max_i \lvert theme\_sentences \cap C_i \rvert} \rvert} \times \log \frac{N}{n}$, where N is the number of content states, n is the number of content states produced in TR_4 that contain at least one sentence from the theme, $\max_i \lvert theme\_sentences \cap C_i$ is the largest number of sentences from the theme, present in a content state, and $\lvert C_{\arg\max_i \lvert theme\_sentences \cap C_i \rvert} \rvert$ is the number of sentences in that content state.
F_8	*Frame mapping*. From topic representation TR_5 we are counting the number of frame slots that are semantically consistent with the theme arguments. This value becomes F_8.

We assume that *cohesion relations* connect topic themes that co-occur in the same document segment and represent an account for common information. In contrast, we assume that coherence relations can be used to account for the logical connections between discourse units. Following Marcu [1998], we propose that access to this kind of discourse-level information can improve the quality of MDS in any possible domain. We expect that by organizing topic themes using either (or both) cohesion and coherence relations, we can generate more relevant MDS, which better reflects the actual organization of topics in texts.

In the rest of this section, we describe how we recognized both types of relations for use in a summarization system.

We implemented procedures for recognizing both forms of relations.

Cohesion relations. Often themes co-occur, (1) in the same sentence, (2) in successive sentences, or (3) in the same text segment. Themes may co-occur

Table III. Values of the Features Presented in Table II for the Theme Illustrated in Figure 16

F_1	*Theme frequency* = 15 / 973 = 0.0154
F_2	*Theme coverage* = $15 \times \log \frac{853}{2} = 39.45$
F_3	*Predicate signature* = 99.41
F_4	*Argument signature* Arg0: 20.34; Arg1: 566.28; Arg2: 21.82; ArgM-ADV: 23.58; ArgM-LOC: 47.57
F_5	*Theme relation signature* For this theme there are no relations from TS_2.
F_6	*Position* = 3.3 / 22.3 = 0.15
F_7	*Content signature* = $\frac{5}{6} \times \log \frac{23}{7} = 0.43$. There are 23 clusters in total, this theme is present in 7 of them with the following frequencies: 5, 2, 2, 2, 1, 1, and in the "etcetera cluster" there are 2 sentences that are part of this theme. There are 6 sentences in the cluster and the theme appears 5 times.
F_8	*Template mapping* = 3 (Defendant, Crime, Prosecution)

in the same sentence because of nestings of predicate-argument structures. Additionally, coreference between arguments of two different themes marks another form of cohesion. To establish cohesion relations between themes, we create for each theme three different lists as follows.

—$L^{S}_{Coh}(Th_i)$ is an inverted list of a theme Th_i containing information about (a) the themes Th_j that co-occur with Th_i in the same sentence, and (b) the frequency of co-occurrence;

—$L^{Succ}_{Coh}(Th_i)$ is an inverted list of a theme Th_i containing information about (a) the themes Th_j that appear in sentences succeeding or preceding the sentences in which Th_i appears, and (b) the frequency of co-occurrence;

—$L^{Ref}_{Coh}(Th_i)$ is an inverted list of a theme Th_i containing information about (a) the themes Th_j that have arguments co-referring with arguments from Th_i, and (b) the number of coreference relations between arguments of Th_i and Th_j. To establish coreference, we have used the coreference resolution system described in Nicolae and Nicolae [2006].

Given any pair of themes Th_i and Th_j, the strength of the cohesion relation between themes is given by:

$$\begin{aligned} W^{Coh}(Th_i, Th_j) \;=\; & \alpha_1 \mathit{Inf}\big(L^{S}_{Coh}(Th_i), Th_j\big) + \alpha_2 \mathit{Inf}\big(L^{Succ}_{Coh}(Th_i), Th_j\big) \\ & + \alpha_3 \mathit{Inf}\big(L^{Ref}_{Coh}(Th_i), Th_j\big), \end{aligned}$$

where $\mathit{Inf}(L^x(Th_i), Th_j)$ represents the contribution of the information provided by the list $L^x(Th_i)$ about the cohesion between Th_i and Th_j. To quantify this contribution, we have used the formula $\mathit{Inf}(L^x(Th_i), Th_j) = \frac{f(Th_j, L^x(Th_i))}{\max_k f(Th_k, L^x(Th_i))}$, where $f(Th_j, L^x(Th_i))$ represents the frequency value of theme Th_j in list $L^x(Th_i)$. To determine the values α_1, α_2, and α_3 we have the choice of using some empirical values or to induce them. In our representation, we selected $\alpha_1 = 0.5$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.2$.[13] The strengths of the cohesion relations obtained for some of the themes of topic T_1 when these values were used, are illustrated in Figure 18.

[13] The values assigned to α_1, α_2, and α_3 reflect the descending order of strength of cohesion based on (a) same-sentence co-occurrence, (b) same-paragraph co-occurrence, and (c) anaphora.

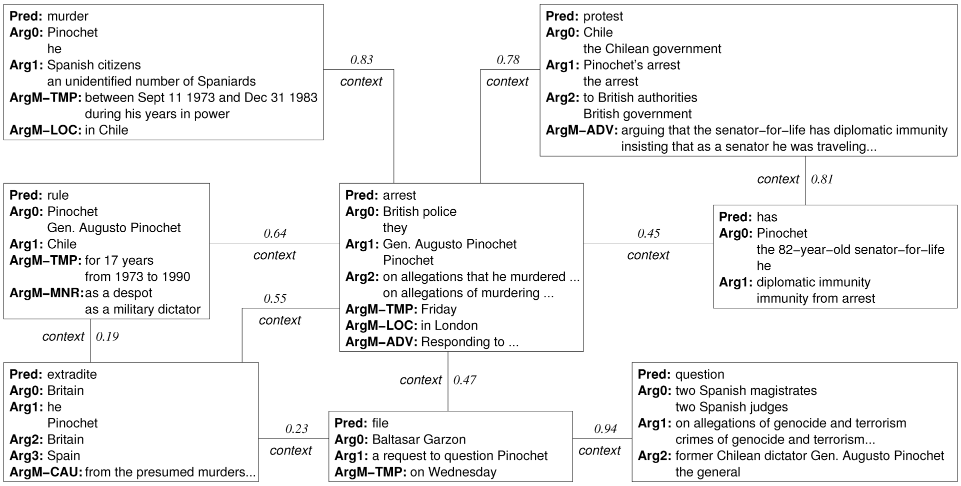


Fig. 18. Cohesion relations between themes from topic T_1.

The relations are labeled as CONTEXT and the weights of each cohesion relation are also illustrated.

Coherence relations. We assume that themes can often be interrelated using the same coherence relations that link the discourse segments from which they are derived. Following Marcu and Echihabi [2002] (who explored using only a small set of discourse relations), we considered only four types of coherence relations, namely (1) CONTRAST relations, (2) CAUSE-EXPLANATION-EVIDENCE relations, (3) CONDITION relations, and (4) ELABORATION relations. Each type of discourse relation was recognized by a separate classifier.

As described in Kehler [2002], most coherence relations can be recognized by the presence of specific cue-phrases, for example, "*but*" or "*although*" for CONTRAST discourse relations; "*because*" or "*therefore*" for CAUSE-EXPLANATION-EVIDENCE relations; "*therefore*" or "*and*" for CONDITION relations; and "that is" or "*in general*" for ELABORATION relations. As noted by Knott and Sanders [1998], the same cue phrases may signal different coherence relations. Moreover, discourse relations often exist between textual units without the presence of any cue phrase. There are two possible solutions to this problem. The first solution, reported in Marcu and Echihabi [2002], relies on very large training corpora bootstrapped from a corpus of human-annotated discourse relations, which allows the implementation of a Bayesian classifier that selects between pairs of possible coherence relations. In the work reported in this article, we have adopted an alternative approach, which constitutes a second possible solution. First reported in Harabagiu et al. [2006], for the recognition of CONTRAST relations, this approach argues that discourse relations can be discovered without considering only the distribution of cue phrases. Harabagiu et al. [2006] demonstrated that different forms of inferential information—such as *opposition* or *causation*—identified from text, can be used as an invaluable source of features when recognizing discourse relations such as CONTRAST or CAUSE-EXPLANATION-EVIDENCE. For example, this approach argues that CONDITION relations can be

- *Step 1*: Discover discourse units where information relevant to the themes is expressed.
- *Step 2*: Identify discourse relations between discourse units.

Fig. 19. Recognizing discourse-relevant information.

identified by *enablement* information, while ELABORATION relations are discovered by *event structure* information.

In order to identify each form of information that pertains to corresponding discourse relations, we used a two step procedure, as illustrated in Figure 19.

Discourse units are identified by the SPADE discourse parser described in Soricut and Marcu [2003]. SPADE performs discourse segmentation by using a statistical model in order to detect the boundaries of each discourse unit in a text. In our work, we only consider those discourse segments that contain information relevant to a set of topical themes.

The second step of the procedure presented in Figure 19 uses relation-specific information. For example, in the case of CONTRAST relations, *antonym* relations are used.[14] In addition, we have automatically generated *antonymy chains*, with a technique reported in Harabagiu et al. [2006].

The recognition of discourse relations between text units can be performed by using a combination of features that represent (a) alignment between entities and predicates from each text that encode similar types of lexico-semantic information; (b) alignment based on functional similarity or inclusion; (c) alignment based on semantic information specific to each discourse relation (e.g. antonymy for CONTRAST); (d) cue phrases (e.g. “but” or “although” for CONTRAST, or “because” or “thus” for CAUSE-EXPLANATION-EVIDENCE). These features are used to train a maximum entropy-based classifier, which determines whether or not the pair of text units being considered represents a case for the discourse relation being tested. The maximum entropy-based classifier also determined the strength (weight) of the discourse relations.

Table IV illustrates the coherence relations established between the themes of topic T_1.

Step 6. organizes themes and their relations into *topic structures*. In our work, we generated topic structures in two different ways. First, we explored how topic themes could be organized into a graph-based topic structure using the cohesion and coherence relations recognized in Step 5. Second, we also explored how a linked list-based structural representation could be used, which ordered themes based on a combination of their, (1) position in a text, (2) density in a collection, and (3) observed connectivity to other themes generated for the same topic.

We used both cohesion and coherence relations between themes in order to generate graph-based topic structures.[15] Each link between themes was

[14] In WordNet there are 1713 antonym relations between nouns, 1025 antonym relations between verbs, 3758 antonyms between adjectives, and 704 antonyms between adverbs.

[15] All cohesion relations were assigned the same label (CONTEXT). We used four different coherence relations in our work, including (1) CONTRAST, (2) CAUSE-EXPLANATION-EVIDENCE, (3) CONDITION, and (4) ELABORATION.

Table IV. Coherence Relations Between Themes from Topic T_1

CONTRAST: grant (amnesty) – arrest; *Strength:* 0.73 Pinochet, who ruled Chile as a military dictator from 1973 to 1990, was granted amnesty in his homeland but was arrested on Friday in London at the request of Spanish authorities, who want him extradited to Spain.
CONDITION: arrest – travel; *Strength:* 0.81 That meant countries that are part of the Interpol agreement would have to arrest the officers if they traveled outside Argentina.
ELABORATION: arrest – murder; *Strength:* 0.88 British police said Saturday they have arrested former Chilean dictator Gen. Augusto Pinochet on allegations of murdering Spanish citizens during his years in power.
CAUSE-EXPLANATION-EVIDENCE: arrest – send (request); *Strength:* 0.69 Pinochet was placed under arrest after a senior Spanish judge, Baltasar Garzon, sent a request to detain him in connection with possible extradition hearings.

assigned both a label and a weight. Themes with no links are then removed from the representation and are not considered further. This approach captures our observation that links between themes are a powerful indicator of potential salience for a summary.

In contrast, we generated a linked list-based representation of topic themes by performing a radix sort of themes based on the following four scores.

◇ Score 1: *The average position of the theme in the documents*. For each instantiation of a predicate-argument structure that is mapped in the conceptual representation of a theme, we register the position as the sentence number. The position is normalized by the document length.

◇ Score 2: *The density of the theme in the collection* is provided by the number of themes with which it has cohesion relations, normalized by the maximum number of cohesive relations in which any theme participates.

◇ Score 3: *The cohesive connectivity of a theme to other themes* is measured by the number of cohesive relations that are established between the current theme and all other themes from the same topic.

◇ Score 4: *The discourse connectivity of the theme to other themes* is measured by $\sum_{k \in \{Discourse\_relations\}} \frac{n_k}{c_k}$, where n_k represents the number of relations of type k between the theme and other themes, whereas c_k represents the total number of relations of type k discovered between any pair of themes.

4. USING TOPIC REPRESENTATIONS FOR MDS

We cast multi-document summarization as a three-stage process. First, documents are sent to a sentence extraction module, that identifies the set of sentences that contain information, which corresponds best to the topic of the document collection. Second, extracted sentences are sent to a sentence compression module, which processes sentences in order to retain only the most important pieces of information found in each sentence. Finally, compressed sentences are then sent to an information ordering module, which constructs a coherent, fixed-length summary of the document collection.

Sentence extraction techniques are used to identify all of the important information that should be included a multi-document summary. Two techniques have traditionally dominated the task of sentence extraction.

First, work in MDS [McKeown et al. 1999; Radev et al. 2000; Marcu and Gerber 2001] used clusters of similar sentences in order to select content for a summary. Once a set of clusters was generated, sentences were then selected by clustering and ranking information for inclusion in a summary. This approach was first formalized using the theme-based representations introduced in McKeown et al. [1999] or the elementary discourse unit-based representations featured in Marcu and Gerber [2001].

A second method [Carbonell et al. 1997; Carbonell and Goldstein 1998] began the process of sentence extraction by ranking document passages using an informativeness measure (such as Maximal Marginal Relevance (MMR)). Sentences that were deemed to be most dissimilar to the sentences already included in a summary were then selected for inclusion in the summary.

More recently, a third method of sentence extraction [Lin and Hovy 2000; Biryukov et al. 2005] has emerged, which leverages automatically-generated topic representations (such as the TRs described in Section 2). This method ranks sentences for inclusion in an MDS based on their topic scores computed from the sum of all of the topical terms included in each candidate sentence. In this article, we describe how this extraction method can be applied for each of the six topic representation methods we presented in Section 2.

The second phase of MDS, sentence compression, is either based on linguistically motivated heuristics, like those reported in Zajic et al. [2004] and Euler [2002], or using the noisy-channel model introduced in Knight and Marcu [2000]. The noisy-channel model finds the most probable short string that generates an observed full sentence. This model for compression was further improved in Turner and Charniak [2005]. In this article, we take advantage of the conceptual representation of themes for deciding on how to compress the extracted sentences.

The final step of MDS, information ordering, organizes sentences originating from different documents into a coherent multi-document summary. The strategy employed for ordering sentences combines constraints from the chronological order of events with measures of topical relativeness, performed on the topic representation TR_3.

The rest of this section provides an overview of a number of different methods for performing, (1) sentence extraction, (2) sentence compression, and (3) information ordering for MDS.

We first discuss a total of six different sentence extraction methods (EM_1–EM_6) which correspond to the four baseline topic representation techniques (TR_1–TR_4), introduced in Section 2, plus the two topic theme-based representations based on TR_5 introduced in Section 3.[16] For ease of exposition, we will refer to the graph-based theme representation as TH_1 and the linked list-based theme representation as TH_2.

In addition, we describe a total of three different sentence compression methods (referred to as CM_0–CM_2) and a set of four different information ordering methods (referred to as OM_1–OM_4).

[16] TR_4 is used only to order sentences (with OM_2). The other topic representations (TR_1, TR_2, TR_3) are used explicitly to retrieve and rank sentences for an extractive summary.

4.1 Extraction Methods

We have considered six sentence extraction methods in our work. The first four methods are based on topic representations as described in Section 2. The fifth method leverages the graph-based theme representation described in Section 3, while the sixth method leverages the linked-list theme representation described in Section 3.[17]

EM_1; Extraction Method 1. Similar to Lin and Hovy [2000], sentence extraction is based on topic identification and interpretation, provided by TR_1 in the form of topic signatures TS_1. Each sentence from the collection receives a topic signature score equal to the total of the signature word scores it contains, normalized by the highest sentence score. Only the sentences with the highest topic scores are extracted. Since topic signatures TS_1, can be represented as ordered linked lists, we consider that the MDS methods that use EM_1 will employ a linked-list representation of topics.

EM_2; Extraction Method 2. The same procedure as in EM_1 can be used when the enhanced topic signature TS_2, pertaining to TR_2, is used to score sentences according to the weights of the topic-relevant relations. Each sentence from the collection receives a topic signature score equal to the weight of the relations from the topic signature TS_2, which are recognized in the sentence. Only the sentences with the highest topic scores are extracted. Since the enhanced topic signatures TS_2, can be represented as ordered linked lists, we consider that the MDS methods that use EM_2 will employ a linked-list representation of topics.

EM_3; Extraction Method 3. EM_3 uses output from TR_2 and TR_3 to select segments for an MDS. Sentences are first assigned a topic score based on TS_2; the sentence with the highest score for each TR_3 theme is then extracted. Since topic signatures TR_3 can be represented as ordered linked lists, we consider that the MDS methods that use EM_3 will employ a linked-list representation of topics.

EM_4; Extraction Method 4. With EM_4, we use a graph-based topic representation based on TR_5 [Harabagiu and Maiorano 2002] in order to generate multi-document summaries. As with Harabagiu and Maiorano [2002], sentences are extracted based on the importance of the template that they partially match. For each slot S_j, of a template T_i, we count the frequency with which a text snippet filled that slot and any other slot of another template. The importance of T_i equals the sum of all frequencies for each slot S_j.

EM_5; Extraction Method 5. The first of two topic theme-based extraction methods, EM_5 uses the graph-based representation of theme information (first described in Section 3) in order to extract sentences for a multi-document summary.

Under this method, each sentence receives a reputation score which takes into account, (1) the reputation of the themes it contains, and (2) its coverage of the theme arguments it contains. Within each theme, certain arguments are recognized in multiple sentences, thus they represent repetitive information. The argument weight in this case is 1/*number of appearances in cluster*. Other

[17] Each EM employs the corresponding baseline TR with the same index; for example, EM_1 uses TR_1, EM_2 leverages TR_2. EM_5 and EM_6 correspond to TH_1 and TH_2, respectively.

arguments are recognized in only one sentence of the theme. The argument weight in this case is 1. The sentence relevance is given by $R^S = \sum_{Th\in T} R_{Th} * \sum_{A\in Arg(Th)} W_A$, where T is the set of themes recognized in sentence S, R_{Th} is the relevance rank of theme Th, *Arg(Th)* represents the set of arguments from its conceptual representation, and W_A is the weight of argument A.

EM$_6$*; Extraction Method 6.* Like EM$_5$, EM$_6$ uses a topic theme-based representation for sentence extraction; unlike EM$_5$, however, EM$_6$ leverages the linked list representation of theme structure generated in Step 6 of the algorithm described in Section 3. Under this method, each sentence receives a score based on, (1) the position in the linked list of the themes it covers, and (2) its coverage of the theme arguments it contains. Since the themes are ordered in the linked list based on their relevance to the topic, themes that appear at the beginning of the list are more important, and the sentences that cover these themes have a higher probability of appearing in the summary. The sentence relevance is computed in EM$_6$ as $R^S = \sum_{Th\in T} \frac{1}{Pos_{Th}} * \sum_{A\in Arg(Th)} W_A$, where T is the set of themes recognized in sentence S, Pos_{Th} is the position of theme Th in the linked-list representation, *Arg(Th)* represents the set of arguments from its conceptual representation, and $W_A = 1/$ nr. *of appearances in the theme* is the weight of argument A.

4.2 Compression Methods

We experimented with three different compression methods in order to reduce the size of a multi-document summary. While the first method (CM$_0$) uses no compression, the second and third methods are based on, (1) an empirical method based on linguistic heuristics (CM$_1$), and (2) a method based on the theme representation (CM$_2$).

CM$_0$*; Compression Method 0.* In order to provide a control for our experiments with summary compression, we created multi-document summaries that were not compressed. We will refer to summaries that use this technique as employing compression method CM$_0$.

CM$_1$*; Compression Method 1.* CM$_1$ starts with a candidate summary that is larger than the intended length and then performs compression by removing content that is redundant, or otherwise unnecessary, given the overall content of the summary.

Following Hovy et al. [2005], we represented the content of a MDS using a syntactic representation, known as a *basic element* (or BE). Like Hovy et al. [2005], we assume that a BE can be defined as a triplet of syntactic constituents consisting of, (1) the head of a syntactic phrase, (b) the modifier of the head, and (3) the syntactic relation between the head and the modifier.

Summaries are compressed by removing redundant information. Redundancy is checked by maintaining a "have-seen" table, which lists all top-ranked BEs of sentences already processed. The order of processing sentences is given by their extraction scores. When a BE from a candidate sentence is found in the "have-seen" table, it is marked for removal. To determine the syntactic constituent in which the tree reduction should occur, the parse tree is traversed from the lowest tree level that covers the "remove" BE up until a decision to backtrack is made on one of the upper-level ancestors. Backtracking occurs when the children of the current node cover unseen top-ranked BEs. For each BE marked as "remove," a tree reduction operation enables the compression of the sentence that is added to the summary.

CM_2; Compression Method 2. Inspired by the word-based compression method of Hori and Furui [Hori and Furui 2004] and the superior results it obtained in the comparisons conducted by Clarke and Lapata [2006], CM_2 uses a word deletion function that maximizes a scoring function in order to compress an MDS. This score, known as a *compression score* is given by:

$$S(V) = \sum_{i=1}^{M} (\lambda_I I(v_i) + \lambda_{SOV} SOV(v_i) + \lambda_L L(v_i|v_{i-1}, v_{i-2}) + \lambda_{Th} Th(v_i)).$$

We assume the sentence $V = v_1, v_2, \ldots, v_m$ (of M words) that maximizes the score $S(V)$ represents the best compression of an original sentence consisting of N words ($M < N$). The first three terms of the sum defining $S(V)$ were reported in Clarke and Lapata [2006]; we added the fourth term to model the role of the topic themes. The lambdas ($\lambda_I, \lambda_{SOV}, \lambda_L, \lambda_{Th}$) are parameters used to balance the weight of each term in the scoring function.[18]

—$I(v_i)$ measures the significance of each word with a formula similar to *tf-idf*: $I(v_i) = f_i \log \frac{F_A}{F_i}$, where f_i is the frequency of v_i in the document, F_i is the corpus frequency of v_i, and $F_A = \sum_i F_i$.

—$SOV(v_i)$ is based on the intuition that subjects, objects and verbs should not be dropped:

$$SOV(v_i) = \begin{cases} f_i & \text{if } w_i \text{ in subject, object or verb role,} \\ \lambda_{default} & \text{otherwise} \end{cases}$$

where $\lambda_{default}$ is a constant weight assigned to all other words, and f_i is the document frequency of a verb, or a word bearing the subject/object role.

—$L(v_i|v_{i-1}, v_{i-2})$ is the linguistic score, which helps to select function words while ensuring that the compressions remain grammatical. It measures the n-gram probability of the compressed sentence.

—$Th(v_i)$ is the theme score, which prefers words that belong to a theme representation:

$$Th(v_i) = \begin{cases} n_i & \text{if } v_i \text{ belongs to the argument of a theme,} \\ \lambda_{default} & \text{otherwise} \end{cases}$$

where n_i is the document frequency of a word relevant to any of the themes.

4.3 Ordering Methods

In this section, we describe the four ordering methods we considered when constructing multi-document summaries. These methods are based on, (1) the ordering introduced in Barzilay et al. [2002], (2) an ordering based on the

[18] The values for the λ parameters differ from one topic to another. While λ_I, λ_{SOV}, λ_L are chosen to normalize the values of their respective terms in the same range, λ_{Th} is chosen such that $\lambda_{Th} Th(v_i)$ has a range twice as large. This is based on our observation that, in general, words belonging to a theme representation are relevant content words and should not be dropped from the summary.

Viterbi algorithm, reported in Barzilay and Lee [2004], and two orderings based on the theme representation using, (3) a linked list-based representation, and (4) a graph representation.

OM_1; Ordering Method 1. This ordering method, introduced in Barzilay et al. [2002], aims to remove disfluencies from the summary by grouping together topically-relevant sentences.

We applied this algorithm in the following way. Sentences extracted by any of the five extraction methods were first clustered using a complete-link clustering algorithm based on the distribution of non-stop words in each sentence.[19] We then computed three types of pairwise relations for each of the possible pair clusters generated. We assumed a relation *Rdoc* held between a pair of clusters (C_1, C_2) (when C_1 was comprised of a set of sentences $(s_1, s_2, \ldots, s_n)$ and C_2 was comprised of the set of sentences $(t_1, t_2, \ldots, t_m)$) if there exists at least one pair of sentences (s_i, t_j) that come from the same document. Likewise, we assume that a second relation, *Rpara*, holds between (C_1, C_2) if there exists at least one pair of sentences (s_i, t_j) that occur in the same paragraph within the same document. Finally, we consider a relation, *Rnull*, to hold if there are no pairs of sentences drawn from (C_1, C_2) that co-occur in the same paragraph or document.

We compute a relation weight for each $Rdoc(C_1, C_2)$ and $Rpara(C_1, C_2)$ based on the number of sentences drawn from each cluster that instantiate that relation. We then compute an overall cluster relatedness score for (C_1, C_2) as $(Rpara/Rdoc)$; if this score is greater than 0.6, we consider the two clusters to be related.

We then use transitive closure to select all of the sentence clusters that are related to one another. We then sort all selected sentences according to, (1) the publication date of that document, and (2) their position in a document; duplicate sentences are removed.

OM_2; Ordering Method 2. OM_2 uses a method first introduced in Barzilay and Lee [2004] to order sentences based on their correspondence to a content model. Under this approach, topic representation TR_4 was used to model the content structure of texts belonging to a domain. Sentences are assumed to follow the order induced by the content structure of the full texts that they were extracted from. Once a content model has been computed, sentences are then added to a summary in descending order of the probability that they correspond to elements of the model.

OM_3; Ordering Method 3. OM_3 uses the inherent order of themes in a linked list-based theme representation (TH_2) to order extracted and compressed sentences.

OM_4; Ordering Method 4. OM_4 uses the order of themes in the graph-based theme representation (TH_1) to order extracted and compressed sentences.

4.4 Generating Multi-Document Summaries

Our experiments evaluate the performance of a total of 40 multi-document summarization systems. (See Figure 20 for the combination of extraction methods (EM), compression methods (CM), and ordering methods (OM) considered in

[19] We used a standard lexicon of stopwords to filter non-content words from consideration.

		EM1	EM2	EM3	EM4	EM5	EM6
OM1	CM0	MDS1	MDS3	MDS5	MDS7	MDS17	MDS21
	CM1	MDS9	MDS11	MDS13	MDS15	MDS25	MDS29
	CM2	CM2 depends on themes				MDS33	MDS37
OM2	CM0	MDS2	MDS4	MDS6	MDS8	MDS18	MDS22
	CM1	MDS10	MDS12	MDS14	MDS16	MDS26	MDS30
	CM2	CM2 depends on themes				MDS34	MDS38
OM3	CM0	OM3 depends on themes				MDS19	MDS23
	CM1					MDS27	MDS31
	CM2					MDS35	MDS39
OM4	CM0	OM4 depends on themes				MDS20	MDS24
	CM1					MDS28	MDS32
	CM2					MDS36	MDS40

Fig. 20. 40 Multi-document summarization methods.

each system.) Since we introduced 6 EMs, 3 CMs, and 4 OMs, a careful reader might expect that we would consider a total of 72 ($6 \times 3 \times 4 = 72$) systems in our work. However, since one compression method (CM$_2$) and two ordering methods (OM$_3$ and OM$_4$) depend on topic theme-based representations, we could not consider 32 systems ($(4 \times 1 \times 2) + (4 \times 2 \times 3) = 32$) for which themes were not computed during sentence extraction or content ordering. The 40 systems we consider represent all of the possible combinations afforded by the algorithms we have defined.[20]

5. EVALUATION RESULTS

Our experimental evaluations had two main goals. First, we sought to quantify the impact that topic information has on each phase of a multi-document summarization (MDS) system. To this end, we conducted a set of experiments that assessed the contribution that the availability of topic information had on three components of a traditional MDS system: (1) sentence extraction, (2) sentence compression, and (3) information ordering. Second, we we wanted to assess the quality of topic representations that produced the largest improvements in quality for automatically-generated multi-document summaries.

In our work, we have considered both, (1) component-based evaluations, which evaluated each phase in the creation of a multi-document summary separately, and (2) intrinsic evaluations, which evaluate the quality of each individual multi-document summary generated by a summarization system.

We performed the following three component-based evaluations.

—*Component-Based Evaluation 1; sentence extraction*. We evaluated the quality of sentence extraction based on different representations of topics.

—*Component-Based Evaluation 2; summary compression*. We separately evaluated the quality of compression when different topic representations were available to the summarization system.

[20] It is true, however, that CM$_2$ could have been computed for EM$_1$–EM$_4$ simply by excluding the theme-based term from Hori and Furui [2004] the scoring function of. Since themes were an important and heavily-weighted component in this computation, we felt that there was little to be learned from performing compression in this way without the theme-based term.

—*Component-Based Evaluation 3; sentence ordering*. We evaluated the quality of ordering with different topic representations.

We performed intrinsic evaluations of summaries using the following three methods.

—*Intrinsic Evaluation 1; recall-oriented understudy for gisting evaluation (ROUGE)*. Following Lin and Hovy [2003], we evaluated summaries using five variants of the ROUGE automatic summary evaluation algorithm: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU.

—*Intrinsic Evaluation 2; pyramid evaluation*. As in the DUC 2005 and DUC 2006 summary evaluations [Dang 2005], we performed a manual evaluation of the quality of summaries using the Pyramid Method outlined in Nenkova and Passonneau [2004].

5.1 Evaluating the Quality of Summaries

Multi-document summaries have traditionally been evaluated both in terms of their information content and their linguistic quality. Content evaluation methods, including ROUGE [Lin and Hovy 2003], Basic Elements (BE) [Hovy et al. 2006], and the Pyramid Method [Nenkova and Passonneau 2004], assume that multi-document summaries can be evaluated based on their correspondence to a summary content model that represents the ideal content that an automatically- or manually-generated summary should contain. Summaries that contain more key information elements (or nuggets) from a content model are presumed to be better summaries than those containing fewer elements.

Summaries have also been evaluated in terms of “linguistic quality” [Dang 2005], as well. In these evaluations, summaries have been judged by human annotators based on their correspondence to other gold-standard, human-authored summaries. Machine-generated summaries have traditionally been evaluated in terms of grammaticality; non-redundancy; referential clarity; focus; and structure and coherence.

5.1.1 *Automatic Evaluation.*

Automatic text summarization has drawn a lot of interest in the natural language processing and information retrieval communities in recent years. Sponsored by the U.S. National Institute for Standards and Technology (NIST), the Document Understanding Conferences (DUC) have organized yearly evaluations of automatically-produced summaries by comparing the summaries created by systems against those created by humans. In 2004, multi-document summaries were produced for 50 different topics.

Following the recent adoption of automatic evaluation techniques (such as BLEU/NIST) by the machine translation community, a similar set of evaluation metrics—known as ROUGE[21] and BE[22]–were introduced for both single and multi-document summarization [Lin and Hovy 2003; Hovy et al. 2006]. ROUGE includes five automatic evaluation methods that measure the

[21] ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. For more information, see http://www.berouge.com.

[22] BE stands for Basic Elements. For more information, visit http://haydn.isi.edu/BE/.

Table V. Description of ROUGE Scores

ROUGE-N measures the *n*-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows: $ROUGE-N = \frac{\sum_{S\in\{Ref.Summ.\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Ref.Summ.\}}\sum_{gram_n\in S} Count(gram_n)}$
ROUGE-L uses the longest common subsequence (LCS) metric in order to evaluate summaries. Each sentence is viewed as a sequence of words, and the LCS between the automatic summary and the reference summary is identified. ROUGE-L is computed as the ratio between the length of the LCS and the length of the reference summary.
ROUGE-W (Weighted Longest Common Subsequence) is an improvement of the basic LCS method. It favors LCS with consecutive matches.
ROUGE-S (Skip-Bigram Co-Occurrence Statistics) measures the overlap ratio of skip-bigrams between a candidate summary and a set of reference summaries. A skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. ROUGE-S with maximum skip distance is called ROUGE-S*N*, where *N* is the distance.
ROUGE-SU is an extension of ROUGE-S that solves the problem of ROUGE-S of not giving credit to sentences that do not have any word pair co-occurring with their references. ROUGE-SU adds the unigram as counting unit.

similarity between summaries: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU (illustrated in Table V).

In a recent DUC evaluation (DUC 2006), three scores were used as part of the official automatic evaluation of multi-document summaries: ROUGE-2, ROUGE-SU4, and BE-HM. As noted in Dang [2005], these three measures have traditionally shown the best correlation with human evaluations of summary quality, with ROUGE-SU4 and BE-HM demonstrating slightly more robustness than ROUGE-2 across summaries and topics examined in previous evaluations.

In the rest of this section, we will use output from ROUGE-SU4 to present the results of the automatic evaluation of the 40 different summarization methods our extraction, compression, and ordering methods enable us to generate for each of the 50 topics considered as part of the DUC 2004 evaluations.[23]

In order to evaluate the quality of summaries with ROUGE-SU4, we take into account that this scoring method is insensitive to the quality of ordering methods used in MDS. Consequently, we select from the 40 MDS methods listed in Figure 20, only the methods that use OM_1. These methods are MDS_1, MDS_3, MDS_5, MDS_7, MDS_9, MDS_{11}, MDS_{13}, MDS_{15}, MDS_{17}, MDS_{21}, MDS_{25}, MDS_{33}, and MDS_{37}. Only MDS_{17}, MDS_{21}, MDS_{25}, MDS_{33}, and MDS_{37} use the novel theme-base topic representations (TH_1, or TH_2).

Figure 21 presents average ROUGE-SU4 results from the 15 MDS methods that use OM_1.[24] Summaries generated using baseline TR methods (TR_1–TR_4) are presented with solid, shaded bars, while summaries generated using theme-based representations (TH_1, TH_2) are presented with solid, white bars.

[23] Although we believe that BE-HM holds much promise for the automatic evaluation of multi-document summaries, we were unable to use it to evaluate the summaries we generated, as no official set of basic elements was generated for the set of topics considered as part of the DUC 2004 evaluations. ROUGE-SU4 produces better correlation with human assessors than ROUGE-2; thus we selected this measure for our automatic evaluation.

[24] HiDUC04 signifies the top-scoring system (in terms of ROUGE-SU4 score) reported in the DUC 2004 official results.

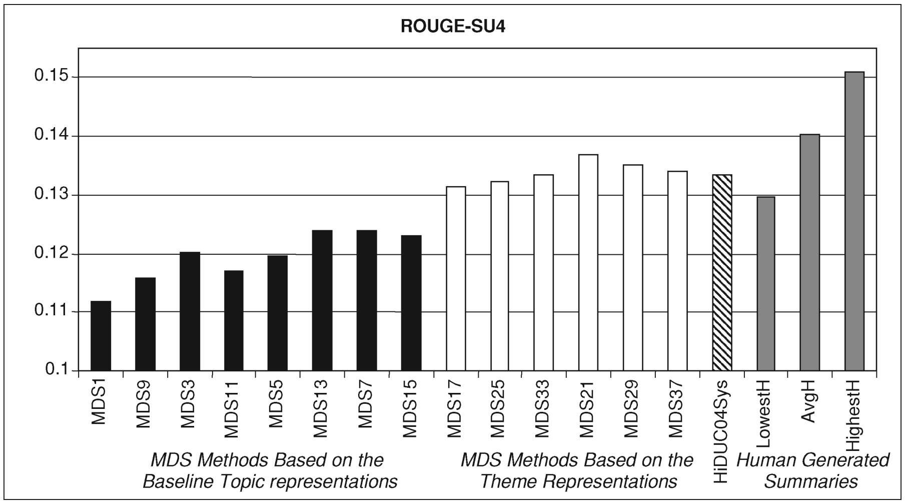


Fig. 21. Summary of the ROUGE-SU4 scores.

Human-generated summaries are presented with gray bars. In Figure 21, LowestH indicates the lowest ROUGE-SU4 scores obtained by human summaries, HighestH represents the highest ROUGE-SU4 scores obtained by human summaries, whereas AvgH represents the average ROUGE-SU4 scores for human summaries.

We can make three preliminary observations based on these results. First, MDS methods using novel topic representations appear to consistently outperform ($p < 0.05$) MDS methods using any baseline topic representations. Second, MDS methods using topic themes generate ROUGE-SU4 scores that are competitive with the scores assigned to human summaries. We found that theme-based summaries score higher than the lowest-scoring human summaries, but not as high as the best-scoring human summaries. Third, we have found that the MDS methods using topic themes also outperform the best system that was evaluated in DUC 2004, which in turn produced better results than those obtained by MDS methods using any of the baseline topic representations.

A closer inspection of the results obtained by MDS methods using baseline topic representations suggests that more complex topic representations produce higher ROUGE scores. If we assign scores to the level of complexity for the extraction methods ($EM_1 = 1$, $EM_2 = 2$, $EM_3 = 3$, $EM_4 = 4$, $EM_5 = 5$, $EM_6 = 5$) and for the compression methods ($CM_0 = 0$, $cM_1 = 0.5$, $CM_1 = 1$), the complexity score for any given MDS method can be computed as the sum of the scores of that method's components. For example, MDS_{29} uses EM_6 and CM_1 and receives a complexity score of $5 + 0.5 = 5.5$. Using this approach, the correlation between the representation complexity and the ROUGE score is very high ($r = 0.94$).

In our results, MDS_{21} (which used a linked-list topic theme representation with CM_0) received the highest overall ROUGE-SU4 score of 0.1368, just slightly ahead of MDS_{33} (graph-based topic theme with CM_2), which scored 0.1335. On average, the three topic theme-based representations that used linked-list-based structures (MDS_{21}, MDS_{29}, MDS_{37}) scored ahead (0.1353) of the three topic theme-based representations (MDS_{17}, MDS_{25}, MDS_{33}) that were implemented using graph-based structures (0.1324). While differences due to extraction method and theme structuring proved to be significant, we found that the choice of compression methods did not significantly affect ROUGE-SU4 scores for any of the 14 methods.[25]

5.1.2 *Manual Evaluation using Pyramids.* The Pyramid summarization evaluation methodology [Nenkova and Passonneau 2004] depends on the assumption that no single reference summary can effectively model an ideal (model) summary for any particular collection of documents. For example, while a single model summary can be used to provide an inventory of some of the information that should be in a machine-generated summary, it provides no mechanism that can be used to identify the most vital pieces of information that necessarily should be included in a summary. For example, we expect that information that is common among a larger number of human summaries should be weighted more in the evaluation process than information particular to fewer reference summaries. The Pyramid method manually decomposes the set of reference summaries into summarization content units (SCUs). A content unit consists of a set of contributions from different summaries that express the same semantic content. The more reference summaries contribute to an SCU, the higher its weight.

In our case, where there are four human summaries for each cluster of documents, the maximum weight for a single SCU is 4 (when its semantic content is present in all four summaries). Given this model, we define a *pyramid* as a set of SCUs that have been arranged into tiers corresponding to their weight. Once a set of ideal SCUs has been identified for a collection of model summaries, each peer summary (a summary generated by an MDS method) is manually annotated with all the matching SCUs in the pyramid. All information that appears in a peer summary that does not appear in the pyramid is labeled as non-matching and does not contribute to the overall score computed for that peer summary. An *OBServed* score is computed for the peer summary as $OBS = \sum_{i=1}^{n} i * O_i$ where n is the number of tiers in the pyramid, and O_i is the number of SCUs of weight i that appeared in the summary.

We consider two maximum scores for each summary: (1) MAX_O, which represents the maximum score an ideal summary can obtain, given the set of SCUs found in the peer summary, and (2) MAX_M, computed as the maximum score possible for a summary, given the average number of SCUs in the set of model summaries. These scores are used to compute two overall pyramid scores, including

- the original pyramid score (Pyr_{orig}), which models precision, and is computed as $\frac{OBS}{MAX_O}$;
- the modified pyramid score (Pyr_{mod}), which models recall, and is computed as $\frac{OBS}{MAX_M}$.

[25] A Student's paired, two-tailed t-test was used to test significance.

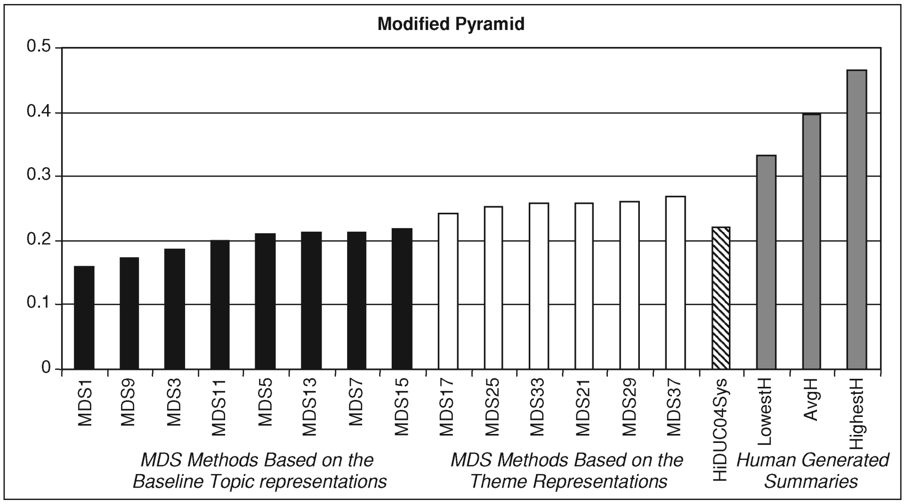


Fig. 22. Summary of the modified pyramid scores.

Since computing the modified pyramid score [Passonneau et al. 2005] does not require the labeling of non-matching SCUs (thereby significantly simplifying the annotation process), we only consider Pyr_{mod} in this section.

We followed the guidelines for the manual creation and scoring of pyramids first described in Passonneau et al. [2005] in order to create pyramids for all of the topics considered in the DUC 2004 evaluations. Model summaries were obtained from the five human-generated "gold-standard summaries" assembled by the DUC 2004 organizers. A team of three human annotators was used to create pyramids from each set of five model summaries. Following the creation of each individual pyramid, annotators met in conference to resolve any discrepancies in their annotation and to create "gold standard" pyramids that could be used for the scoring of the 700 peer summaries we consider in this article.[26] Each peer summary was scored by a total of two annotators to ensure consistency of annotation. Pyramid scores were then averaged; any pair of pyramid scores that differed by more than 25% was discussed in conference by the pair of annotators until a consensus score could be achieved.

Figure 22 illustrates the average modified pyramid scores for the 14 MDS methods considered in the results from the ROUGE-SU4 evaluation presented in Section 5.1.1. As was observed in Figure 21, multi-document summaries based on topic themes received higher modified pyramid scores than summaries based on any other topic representation, regardless of the structural model used to rank themes. Again, as with the ROUGE-SU4 results, theme-based summaries based on a linked-list structure (TH2) outperformed the theme-based summaries generated from graph-based structures of themes (TH1). The highest-scoring list-based summary method, MDS_{37}, received a modified

[26] A "gold" pyramid was created for each topic evaluated in DUC 2004 (50 in total). Each of the 14 summarization methods considered in the content evaluation produced one summary per topic. Thus, the total number of summaries considered for the Pyramid evaluation was 700.

Table VI. Evaluation of the Summary Content for the 14 Content Selection Methods

EM	CM	OM	MDS Method	ROUGE-SU4	Modified Pyramid
1	0	1	MDS_1	0.1117	0.1585
1	1	1	MDS_9	0.1159	0.1741
2	0	1	MDS_3	0.1201	0.1865
2	1	1	MDS_{11}	0.1171	0.1983
3	0	1	MDS_5	0.1197	0.2104
3	1	1	MDS_{13}	0.1238	0.2122
4	0	1	MDS_7	0.1239	0.2137
4	1	1	MDS_{15}	0.1231	0.2189
5	0	1	MDS_{17}	0.1313	0.2226
5	1	1	MDS_{25}	0.1323	0.2514
5	2	1	MDS_{33}	0.1335	0.2571
6	0	1	MDS_{21}	0.1368	0.2569
6	1	1	MDS_{29}	0.1351	0.2601
6	2	1	MDS_{37}	0.1339	0.2689

Table VII. Evaluation of the Ordering Methods

Ordering Method	EM_6		
	Poor	Fair	Good
OM_1	3	7	15
OM_2	4	7	14
$\mathbf{OM_3}$	**2**	**6**	**17**
OM_4	6	5	14

pyramid score of 0.2689, which topped the best graph-based method, MDS_{33}, which received a score of 0.2571. On average, differences between extraction and theme structuring methods proved to be significantly different ($p < 0.05$); as with the automatic evaluation, differences between compression methods, however, did not prove to be significant.

Full results from these two summary evaluation methods are presented in Table VI.

5.1.3 *Manual Evaluation of the Ordering Methods.*

We used the evaluation method proposed in Barzilay et al. [2002] in order to evaluate the performance of each of the sentence ordering methods we employed (OM_1–OM_4). Output from each OM was given to a set of three human assessors who had to assign a score of POOR, FAIR, or GOOD to the summary. We found this score takes into account only the readability of the summary (in terms of sentence ordering), and not how good the summary is for the given set of documents.

We investigated each of the four different OMs using the output of EM_6, the only one of the EMs that does not utilize any form of sentence compression. This evaluation was performed on 25 of the 50 document clusters from the DUC 2004 data. (25 topics were selected at random. Not all of the 50 topics could be completed due to resource and time constraints.) The results of the evaluation of the sentence ordering methods are presented in Table VII.

Ordering method OM_3 had the best results, producing the most summaries with a GOOD ordering (17), and also the fewest POOR ordered summaries (2). If

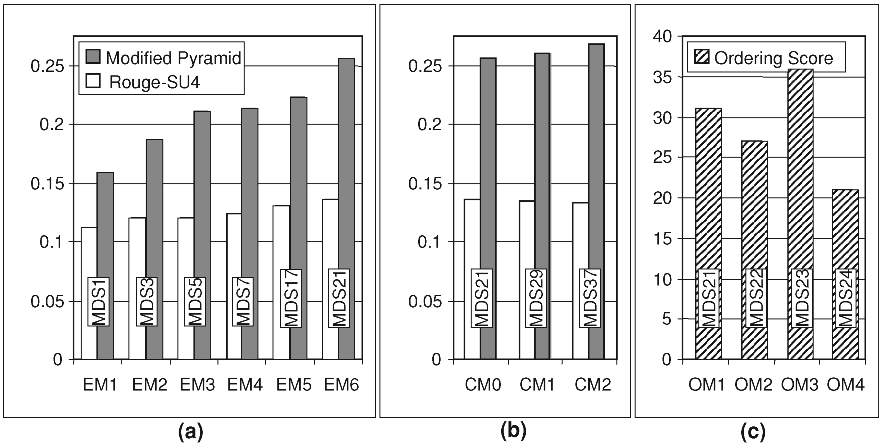


Fig. 23. Evaluation of (a) extraction methods, (b) compression methods, and (c) ordering methods.

we take into consideration the fact that OM_3 is the sentence order induced by the linked-list representation of themes, the ordering evaluation results also prove the positive influence of using the theme representation.

5.2 Evaluating the Impact of Topic Information on Summaries

Topic representations can be used in any of the three stages of multi-document summarization. We also separately evaluated the performance of different extraction methods, compression methods, and ordering methods. Figure 23 illustrates the evaluation results obtained using both the Pyramid scores and ROUGE-SU4.

To evaluate extraction methods, we considered all MDS methods that used CM_0 and OM_1. The results illustrated in Figure 23(a) show that extraction methods using more complex topic representations perform better. To evaluate compression methods, we selected the best extraction method (EM6) and OM_1. The selection of the best extraction method in evaluating compression methods is motivated by the increased likelihood of replacing compressed information with good content. The results of the evaluation of compression methods, illustrated in Figure 23(b), do not show any significant impact of topic information on the quality of compression, because after compression, new sentences that are selected and added to the summary do not bring significant information. The ordering methods, as described in Section 5.1.3, were evaluated by using EM_6 and CM_0. The best results were obtained by OM2, which is based on the topic representation based on themes. In order to be able to represent the results of the ordering methods we chose the following formula: $-2 * \mathit{Poor} + \mathit{Fair} + 2 * \mathit{Good}$ for the chart illustrated in Figure 23(c).

5.3 Evaluation of the Topic Themes

In this section, we present results from experiments designed to evaluate the quality of automatically-generated topic themes in an MDS context.[27]

[27] The topic themes we consider are generated using the methods described in Section 3.2.

Table VIII. Data Sets Used in the Semantic Parsing Evaluation

Data Set 1. Hand-corrected syntactic parses available from the Treebank component of PropBank

Data Set 2. Automatic parses derived from the output of the syntactic parser [Collins 1999]

Data Set 3. Syntactic parses from Data Set 1 augmented with hand-identified arguments.

Data Set 4. Hand-corrected syntactic parses on the test data.

Data Set 5. Automatic parses from the Collins syntactic parser on the test data.

We assessed the performance of our algorithm for discovering topic themes in four separate evaluations:

(1) *Semantic Parsing*. We evaluated the predicate-argument structures generated by our system's semantic parser using the gold standard annotations available in PropBank [Palmer et al. 2005]. Traditionally, the performance of semantic parsers has been evaluated on at least two tasks.[28]

The first subtask, *argument identification*, involves the discovery of the boundaries of each argument from a putative predicate argument-structure. The second subtask, *argument classification*, involves the assignment of the appropriate argument role labels to the known (previously discovered) arguments of a predicate.

The semantic parsing model introduced in Gildea and Jurafsky [2002] and [Gildea and Palmer 2002] operates on the output of a probabilistic syntactic parser, described in Collins [1999]. Because semantic parsing utilizes features derived from valid syntactic parse, it is assumed that the performance of a semantic parser (both at the level of argument identification and argument classification) can be adversely impacted by errors introduced by the syntactic parser. In order to evaluate the impact of syntactic parse errors on the quality of semantic parsing, we have considered three different data sets as inputs—data sets 1, 2, and 3 from Table VIII.

Each of the two subtasks were evaluated in terms of precision (P), recall (R), F1-Measure (F1),[29] and classification accuracy (A). Table IX presents the evaluation of our parser for the first three data sets.

Table IX shows that results from our semantic parser are competitive with results from the state-of-the-art semantic parser described in Pradhan et al. [2005]. In order to evaluate the performance of our semantic parser on the kinds of data featured in the DUC evaluations, we manually annotated a set of 30 documents from the DUC 2004 corpus with predicate-argument information. We then used this annotated corpus to create two additional data sets to evaluate the performance of our semantic parser—data sets 4 and 5 from Table VIII. Results for these two data sets are presented in Table X.

[28] Here, we follow the evaluation methods proposed in the semantic parsing literature, including work by [Pradhan et al. 2005] and [Gildea and Palmer 2002]. The training set consisted on Wall Street Journal sections 02-21 from PropBank, while the testing was done on section 23.

[29] $F1 = \frac{2PR}{P+R}$.

Table IX. Evaluation Results for the Detection of Predicate-Argument Structures with Shallow Semantic Parsing

<table>
<tr><th colspan="5">TASK 1: <i>Argument Identification</i></th></tr>
<tr><td>Syntactic Parse</td><td>P</td><td>R</td><td>Fβ1</td><td>A</td></tr>
<tr><td>Data Set 1</td><td>95.9%</td><td>91.2%</td><td>93.5%</td><td>93.9%</td></tr>
<tr><td>Data Set 2</td><td>97.8%</td><td>94.3%</td><td>96.0%</td><td>93.0%</td></tr>
<tr><td colspan="5">TASK 2: <i>Argument Classification</i></td></tr>
<tr><td>Data Set 1</td><td>93.7%</td><td>87.9%</td><td>90.7%</td><td>93.8%</td></tr>
<tr><td>Data Set 2</td><td>95.1%</td><td>90.0%</td><td>92.5%</td><td>95.1%</td></tr>
<tr><td>Data Set 3</td><td>96.9%</td><td>93.2%</td><td>95.0%</td><td>97.1%</td></tr>
</table>

Table X. Performance of Semantic Parser on DUC Data

<table>
<tr><th colspan="5">TASK 1: <i>Argument Identification</i></th></tr>
<tr><td>Syntactic Parse</td><td>P</td><td>R</td><td>Fβ1</td><td>A</td></tr>
<tr><td>Data Set 4</td><td>95.2%</td><td>90.5%</td><td>92.8%</td><td>93.4%</td></tr>
<tr><td>Data Set 5</td><td>96.9%</td><td>94.1%</td><td>95.5%</td><td>92.8%</td></tr>
<tr><td colspan="5">TASK 2: <i>Argument Classification</i></td></tr>
<tr><td>Syntactic Parse</td><td>P</td><td>R</td><td>Fβ1</td><td>A</td></tr>
<tr><td>Data Set 4</td><td>93.1%</td><td>87.8%</td><td>90.4%</td><td>93.3%</td></tr>
<tr><td>Data Set 5</td><td>95.1%</td><td>89.7%</td><td>92.3%</td><td>95.2%</td></tr>
</table>

We believe the comparable results our parser achieves on the PropBank and DUC data sets can be explained from the similarities in style and linguistic structure between these two corpora.

(2) *Paraphrase Recognition*. We evaluated the accuracy of our paraphrase recognition module using hand-annotated examples that we extracted from the DUC 2003 data. We considered 10 clusters from the DUC 2003 data, and for each of them we marked all the predicate-argument paraphrases, creating the gold-standard set of paraphrases.[30] We ran the paraphrase identification procedure over the 10 clusters, and for each cluster, each potential paraphrase was compared with the gold set of paraphrases. For each topic we computed the recall and precision of our procedure. The recall (R) is the number of correctly identified paraphrases divided by the number of gold paraphrases. The precision (P) is equal to the total number of correctly identified paraphrases divided by the number of pairs marked as paraphrases by the system. We computed an average recall of 74.2% and an average precision of 69.8%, which yields an F_1 score of 71.9%.

(3) *Sentence Clustering*. We measured the quality of our sentence clustering algorithms using the normalized mutual information measure introduced in Ji et al. [2006]. Given two sets of clusters, C and C', the normalized mutual information is defined as:

$$NMI(C, C') = \frac{\textsc{mi}(C, C')}{\textsc{max}(H(C), H(C'))},$$

[30] Annotators were instructed to consider a pair of predicates (p_i, p_j) as paraphrases iff p_j could be substituted for p_i in any sentence s_i containing p_i without changing the meaning of s_i significantly.

where $MI(C, C')$ is the mutual information between clusters C and C', and $H(C)$ and $H(C')$ are the entropies of C and C', respectively. The mutual information is defined as:

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

where $p(c_i, c'_j)$ denotes the joint probability that a randomly selected sentence belongs to both c_i and c'_j (it is computed as the ratio between the number of sentences the two clusters have in common and the total number of sentences); and $p(c_i)$ and $p(c'_j)$ are the probabilities that a randomly selected sentence belongs to clusters c_i and c'_j, respectively (these probabilities are computed as the ratio between the number of sentences in the cluster and the total number of sentences).

The entropy for a set of clusters, C, is computed as:

$$H(C) = -\sum_{i=1}^{|C|} \frac{|c_i|}{n} \cdot \log_2 \frac{|c_i|}{n},$$

where n is the total number of sentences.

To evaluate the performance of clustering the theme sentences, we used the same 10 topics as in the evaluation of the paraphrase discovery. For each topic we computed $NMI(C, C')$ using the formula described previously, where C is the clustering produced by the system and C' is the gold clustering for the given topic. The average value for the *NMI* measure was 0.8839, with the highest value of 0.9031 achieved for topic T1 = Pinochet Trial, and the lowest value of 0.8572 for topic T2 = Leonid Meteor Shower.

We believe that the discrepancy between these values can be explained by at least two factors. First, topic T1 is focused on one person (*Pinochet*) and one event (*the arrest*), while although topic T2 is focused on only one event (*the Leonid meteor shower*), it deals with the impact the event has around the globe. Second, while the documents for topic T1 were published in a temporal window of three days, the documents for topic T2 were published in a window of one month.

(4) *Theme Selection*. We evaluated the accuracy of our theme selection techniques using clusters annotated with theme information. We used human annotators in order to evaluate the precision of our theme selection and ranking modules.

In order to evaluate the precision of our theme generation algorithm, annotators were asked to classify each theme generated for a topic as either, (1) *relevant* to the overall topic (REL), (2) *irrelevant* (IRR) to the topic, or (3) *unclassifiable* (UNC) (as either relevant or irrelevant), given the information available. Themes were deemed unclassifiable when there was insufficient evidence available to the annotators to make a relevance judgement.

Annotators were presented with a total of 125 different automatically-generated themes, ordered randomly from the top 25 themes generated for

Table XI. Theme Relevance: Intra- and Inter-Annotator Agreement

	Intra-Annotator				Inter-Annotator
	A_1	A_2	A_3	Average	Average
T_1	0.94	0.78	0.81	0.843	0.725
T_2	0.91	0.93	0.85	0.897	0.705
T_3	0.83	0.92	0.84	0.863	0.602
T_4	0.85	0.90	0.92	0.890	0.755
T_5	0.88	0.79	0.83	0.833	0.698
Average	0.882	0.864	0.850	0.865	0.697

Table XII. Theme Relevance: Annotators' Judgments

	Annotator Judgments								
	A_1			A_2			A_3		
	Rel	Irr	Unc	Rel	Irr	Unc	Rel	Irr	Unc
T_1	61	10	4	60	15	0	58	16	1
T_2	41	34	0	39	36	0	35	38	2
T_3	66	9	0	62	10	3	51	24	0
T_4	57	15	3	44	31	0	55	19	1
T_5	49	26	0	48	26	1	56	15	4
Total	274	94	7	253	118	4	255	112	8
Precision	73.1%	25.1%	1.8%	67.5%	31.5%	1.0%	68.0%	29.8&	2.1%

each of five different DUC 2004 summarization topics.[31] In addition to a theme, annotators were also presented with a short prose description of the summarization topic. A total of three annotators were used in this experiment and each experimental session was repeated three times over a course of three weeks with each annotator. While the same themes were used in each repetition of the experiment, themes were presented in a randomized order for each experimental session, and no two sessions featured the same order of themes. Kappa statistic measurements of the level of intra-annotator and inter-annotator agreement are presented in Table XI.

In our experiments, we found that individual annotators' judgments of relevance remained stable across each of three experimental sessions high Kappa scores were observed for intra-annotator agreement, ranging from 0.78 to 0.94. Inter-annotator agreement was measured by comparing the most frequently returned judgment for each theme returned by each annotator.[32] Although slightly lower than the agreement observed with the intra-annotator comparisons, Kappa scores for inter-annotator agreement remained relatively high, ranging from 0.698 to 0.755 for individual topics.

We believe that these high levels of intra- and inter-annotator agreement suggest that human annotations can be used to reliably evaluate the quality of theme generation and selection algorithms. Table XII presents results from annotators' classification of the 125 themes evaluated in this experiment.

When results from all three annotators were aggregated by topic, we found that annotators judged 69.5% (782/1125) of themes evaluated in the three

[31] T1 = Pinochet Trial, T2 = Cambodian Government Coalition, T3 = Hurricane Mitch, T4 = Car Bomb in Jerusalem, and T5 = NBA Labor Disputes.

[32] In our experiments, no annotator returned all three classifications—*relevant*, *irrelevant*, and *unclassifiable*, for a particular theme, although there was nothing a priori preventing this outcome.

Table XIII. Theme Relevance: Annotators' Judgments

	Topic		
	Rel	Irr	Unc
T_1	179	41	5
T_2	115	108	2
T_3	179	43	3
T_4	156	65	4
T_5	153	67	5
Total	782	324	19
Precision	69.5%	28.8%	1.7%

Table XIV. Theme Ranking: Wilcoxon Ranked-Sums Test

	Differences			Total	Error
	A_1	A_2	A_3	Errors	Rate
T_1	0	0	0	0	0.0%
T_2	3	2	0	5	55.5%
T_3	0	1	0	1	11.1%
T_4	0	3	1	4	44.4%
T_5	3	2	3	8	88.8%
Total	6	8	4	18	40.0%

experiments to be *relevant*, 28.8% (324/1125) of themes were judged to be *irrelevant*, and a total of 1.7% of the themes (19/1125) were judged *unclassifiable*. (Table XIII presents details from these experiments.)

While these results are encouraging, they only reflect each individual annotator's attitudes towards the relevance of an individual theme. When judgments from all three annotators were combined, a total of 81.6% (101/125) of themes were judged to be relevant by at least one annotator in at least two of the three experimental trials.

In order to evaluate the precision of our theme ranking modules, we had annotators rank the top 25 themes selected for a topic in order of their perceived relevance to the topic itself. (The same 5 topics and 125 themes used in the previous experiment were used in this experiment as well.) Annotators ranked the themes associated with each of the five summarization topics a total of three times over a three week period; rankings were performed immediately following the classification task. (Annotators were used here who had not previously seen the theme ranking produced by the system.)

We used the Wilcoxon ranked-sums test to compare each human ranking for each of the 5 topics against the theme ranking output by our algorithm. A significant difference ($p < 0.05$) between ranked lists of themes was assumed to signal that the algorithm's ranking was suboptimal, given the set of themes and the individual annotator's interpretation of the summarization topic. If no significant difference between the ranked lists of themes was found, we assumed that the algorithm's ranking was a sufficiently accurate approximation of the annotator's ranking. Table XIV provides a comparison of rankings for the 15 rankings generated by our human annotators.

In our experiments, we found that our theme ranking algorithm produced rankings that were statistically indistinguishable from annotators' ranking in 60% (27/45) of cases. Although our system differed in terms of its theme ranking from annotators' ranking 40% of the time, we believe that these results demonstrate the effectiveness of our theme-based approach in identifying the content that should be incorporated in a multi-document summary, especially given the potentially large number of possible rankings that could exist for a set of themes.

In addition, we measured the performance of our approach to identifying topic structures in three additional evaluations:

(5) *Recognition of Cohesion Relations*. We evaluated the precision of our approach to recognizing cohesion relations between themes using a new hand-annotated corpus developed from the DUC 2004 data.

Cohesive relations between themes are inspired by the notion of lexical chains [Morris and Hirst 1991], which show that sequences of words $w_1, w_2, \ldots, w_i$ in a text can be considered to represent the same topic when the words are connected by any of a set of lexico-semantic relations. Since topic themes contain information that originates from different documents, we argue that the notion of *lexico-semantic cohesion* can be replaced by the notion of *contextual cohesion*. Since cohesion relations consider contextual information, whenever a theme is incorrectly selected to represent a topic, it may also lead to the identification of incorrect cohesion relations. Furthermore, whenever a representative theme is not recognized, cohesion relations to other themes representing the same topic are missed. We described the evaluation of the theme selection in item 4, and thus we consider that cohesion relations should be evaluated only on correctly selected themes. In order to evaluate the precision (P) and recall (R) of the cohesion relations, we used the same data as in the evaluation of theme selections. The results indicate P = 83%, R = 75%, and F1 = 79%. The only error source is the quality of reference resolution, which performs with P = 92.7% and R = 92.4% (F1 = 92.55%) on benchmarked data, as reported in Nicolae and Nicolae [2006].

(6) *Recognition of Coherence Relations*. We measured the accuracy of the coherence relations between themes using another hand-annotated corpus derived from the DUC 2004 data.

Coherence relations exist between discourse units that belong to the same text. When we identify discourse relations between themes, we argue that they correspond to the relations between discourse units that contain information relevant to those themes themselves. The procedure of recognizing discourse-relevant information between themes consists of two different steps, illustrated in Figure 19.

The identification of discourse units where information relevant to a theme is performed by using a discourse segmenter available from SPADE [Soricut and Marcu 2003]. When evaluated on the RST Discourse Treebank,[33] the discourse segmenter achieves a precision of 84.1%, a recall of 85.4%, and an Fβ1 score of 84.7%. We did not evaluate the discourse segmenter separately on our test set of DUC data.

[33] The data, but not the annotations are available from the Linguistic Data Consortium http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07.

Table XV. Performance of Discourse Relation Classifier

Relation	P	R	F1
CAUSE	73%	69%	71%
CONTRAST	64%	70%	67%
CONDITION	71%	69%	70%
ELABORATION	77%	81%	79%

Table XVI. Run-Time Performance for Generating each of the Topic Representations

Topic Representation	Time
TR_1	3s
TR_2	5s
TR_3	23s
TR_4	34s
TR_5	45s
TH_1	192s
TH_2	192s

We did, however, evaluate the accuracy of a separate classifier we developed (using the procedure outlined in Step 2 from Figure 19) to identify four types of discourse relations: CAUSE, CONTRAST, CONDITION, and ELABORATION. We evaluated the classifier's performance on a total of 50 documents taken from DUC documents related to the five topics used in the theme selection evaluation.

Before we could evaluate the performance of our classifier, we had our team of linguists annotate the DUC documents with these four discourse relations.[34] Discourse relations were annotated between automatically-identified discourse units and themes. Based on these measures, we have been able to compute the precision (P), recall (R), and F1-measure for our discourse relation classifier. Results are provided in Table XV.

5.4 Evaluation of System Run-Time Performance

We also measured the time it took for each topic representation to be generated from a collection of texts. All experiments were conducted using a dual-core, 4GB RAM system (see Table XVI).

5.5 Discussion of Empirical Results

The results presented in the previous section suggest that topic representations do play an important role in the creation of informative multi-document summaries. In general, we have found that more articulated topic representations—such as the theme-based representations we introduced in this work—provide for multi-document summaries that are more representative and more coherent than summaries generated using less structured topic representations (such as the "bag-of-terms," and "bag-of-relations"-based approaches introduced in TR_1

[34] The average Kappa across these four discourse relations was 0.6752, ranging from 0.5835 (CONTRAST) to 0.8524 (ELABORATION).

and TR_2, respectively. This may be because sentences selected by more complex themes already create a coherent sequence of ideas which, when taken together, contribute to a more meaningful summary.

These results beg a somewhat obvious question namely: Why should structured topic representations such as TH_1 and TH_2 necessarily outperform flatter representations on an MDS task like the one featured in DUC 2004? What is it, exactly, about these hierarchical topic representations that provides for summaries and with more content better organized than their competitors?

We assume that a topic representation generated from a set of documents D can be thought of as a schematized representation of the ideal content and structure of a multi-document summary of any length generated from D. MDS systems that are leverage topic representations can then be seen as performing the content extraction and ordering necessary to realize these schematic representations in the form of a fluent, natural language text, based on the structure and content of a topic representation. Given this model, we hypothesize that TRs that represent a more articulated model of a natural language text will automatically generate better multi-document summaries than topic representations that may be more impolished lexically, semantically, or pragmatically.

The topic representation methods that performed best in our experiments were the theme-based methods (TH_1 and TH_2), which leveraged a combination of lexical, relational, and discourse structure information in order to generate an MDS. While it is not surprising that these topic representations outperformed the simpler TRs we considered (TR_1 and TR_2), the real value of these methods can be seen as a comparison with the other more articulated TRs such as TR_3 and TR_4, which sought to combine multiple different dimensions of the content of a document collection into a single topic representation. Even though TR_3 featured discourse structure information that presumably could not be captured by extracting terms or relations from single sentences alone, we found that it did not include the necessary lexico-semantic knowledge needed to draw correspondences between semantically similar lexical items or categories of lexical items. Likewise, while TR_4 combined both information from semantic clusters of sentences dealing with the same topic with respect to the information derived from a hidden Markov model-based content model, we found that performance still lagged behind the theme-based representations, which sought to simultaneously model both document structure and content selection.

The results suggest that there may not necessarily be a linear correspondence between topic representation complexity and MDS performance. Our experiments, however, show that topic representations that approximate, (1) the semantic content, (2) the cohesiveness, and (3) the discourse structure of ideal summaries do, in fact, outperform models that approach any of these dimensions in isolation. Based on these results, we believe there is particular merit in creating topic representations which model the ideal content and structure of multi-document summaries in terms of the information available from a collection of documents. This hypothesis is further supported by the discrepancies in performance between the graph-based theme representations and their better-performing counterparts, the linked-list -based theme representations. With graph-based topic representation (TH_1), discourse relations identified between discourse segments are used to reorganize the content of a document collection into a new type of topic representation, which equally considers the organization of individual documents and the salience of extracted information. While we anticipate that this type of topic representation could lead to more coherently organized MDS, it can end up selecting less relevant content for inclusion in a summary in order to preserve the coherence of the output text, this reducing the overall content quality (informativeness) of the multi-document summary. In contrast, a linked-list-based topic representation (TH_2) considers theme-based elements in terms of their observed connectivity to other theme elements recognized in a document collection. In this case, we expect that superior results are observed because content salience remains the primary feature governing the creation of summaries, while the lexicosemantic and discourse-level information is used to organize the content into a coherent summary.

6. CONCLUSIONS

In this article, we investigated the impact that methods for topic representation and structuring can have on the quality of multi-document summaries.

We believe that this work has two main contributions.

First, we have introduced two new topic representations—based on structured sets of topic themes—which we believe provide much more sophisticated models of topic-relevant information as compared to previous topic representation methods. Unlike most traditional forms of topic representation (which are based solely on term counts and shallow forms of lexicosemantic information such as the output of semantic parsers), we have leveraged two different structural representations for topic themes in order to improve the quality of content selection and information ordering for MDS. While previous work has used cohesion and coherence information for MDS, we know of no other work that has incorporated these inter- and intra-sentential relations in order to structure topic representations for MDS.

Second, we believe that our results represent a comprehensive and replicable study, which demonstrates the effectiveness of a structured, theme-based approach to multi-document summarization. We have presented an evaluation of a total of forty MDS methods that use different sentence extraction (EM), sentence compression (CM), and information ordering methods (OM) in order to show how extraction, compression, and ordering for MDS can be improved when topic themes are available as part of the input. We have evaluated each of these summarization methods using a number of standard techniques, including the ROUGE automatic scoring packages and the manual Pyramid evaluation method. The scores obtained in both the automatic ROUGE evaluation and in the manual Pyramid evaluation ranked highest the summaries that used sentence extraction and compression methods based on theme representations that organize information associated with a set of themes as a linked list. The best scores of sentence ordering were obtained by an ordering method based on themes. We believe that these results prove the benefits of using the theme-based topic representation for multi-document summarization.

REFERENCES

BAAYEN, R., PIEPENBROCK, R., AND GULIKERS, L. 1995. *The CELEX Lexical Database* (Release 2) [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania [Distributor], Philadelphia, PA.

BAKER, C. F., FILLMORE, C. J., AND LOWE, J. B. 1998. The Berkeley FrameNet project. In *Proceedings of the Joint Conference of the International Committee on Computation Linguistics and the Association for Computation Linguistics (COLING-ACL'98)*. 86–90.

BARZILAY, R. AND LEE, L. 2004. Catching the drift: probabilistic content models, with applications to generation and summarization. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*. 113–120.

BARZILAY, R., MCKEOWN, K. R., AND ELHADAD, M. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 550–557.

BARZILAY, R., MCKEOWN, K. R., AND ELHADAD, M. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *J. Artif. Intell. Res*. 35–55.

BEJAN, C. A. AND HATHAWAY, C. 2007. Utd-srl: A pipeline architecture for extracting frame semantic structures. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*.

BIRYUKOV, M., ANGHELUTA, R., AND MOENS, M.-F. 2005. Multidocument question answering text summarization using topic signatures. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR'5)*.

CARBONELL, J., GENG, Y., AND GOLDSTEIN, J. 1997. Automated query-relevant summarization and diversity-based reranking. In *Proceedings of the Workshop on AI in Digital Libraries (IJCAI'97)*. 12–19.

CARBONELL, J. G. AND GOLDSTEIN, J. 1998. The Use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, A. Moffat and J. Zobel, Eds., 335–336.

CLARKE, J. AND LAPATA, M. 2006. Models for sentence compression: a comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

COLLINS, M. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, University of Pennsylvania.

DANG, H. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Workshop (DUC'05)*.

DEJONG, G. F. 1982. An overview of the FRUMP system. In *Strategies for Natural Language Processing*, W. G. Lehnert and M. H. Ringle Eds., Lawrence Erlbaum Associates, 149–176.

EULER, T. 2002. Tailoring text using topic words: selection and compression. In *Proceedings of 13th International Workshop on Database and Expert Systems Applications (DEXA'02)*. 215–222.

FELLBAUM, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

GILDEA, D. AND JURAFSKY, D. 2002. Automatic labeling of semantic roles. *Comput. Linguist. 28*, 3, 245–288.

GILDEA, D. AND PALMER, M. 2002. The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL'02)*. 239–246.

GRISHMAN, R. AND SUNDHEIM, B. 1996. Message understanding conference - 6: A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*. 466–471.

HARABAGIU, S. 1997. WordNet-Based Inference of Textual Context, Cohesion and Coherence. Ph.D. thesis, University of Southern California, Los Angeles, CA.

HARABAGIU, S. 2004. Incremental Topic Representations. In *Proceedings of the 20th COLING Conference*.

HARABAGIU, S., HICKL, A., AND LACATUSU, F. 2006. Negation, contrast and contradiction in text processing. In *Proceedings of the Annual Conference of the American Association for Artificial Intelligence (AAAI'06)*.

HARABAGIU, S. AND MAIORANO, S. 2002. Multi-document summarization with GISTexter. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*.

HEARST, M. A. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computat. Ling. 23*, 1, 33–64.

HICKL, A., WILLIAMS, J., BENSLEY, J., ROBERTS, K., RINK, B., AND SHI, Y. 2006. Recognizing textual entailment with LCC's Groundhog System. In *Proceedings of the 2nd PASCAL Challenges Workshop*.

HIRSCHMAN, L., ROBINSON, P., FERRO, L., CHINCHOR, N., BROWN, E., GRISHMAN, R., AND SUNDHEIM, B. 1999. *Hub-4 Event99 General Guidelines and Templettes*. Springer.

HORI, C. AND FURUI, S. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Trans. Inform. Syst. E87-D(1)*, 15–25.

HOVY, E., LIN, C. Y., AND ZHOU, L. 2005. A BE-based multi-document summarizer with sentence compression. In *Proceedings of Multilingual Summarization Evaluation Workshop (ACL'05)*.

HOVY, E., LIN, C.-Y., ZHOU, L., AND FUKUMOTO, J. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

JI, X., XU, W., AND ZHUJING, S. 2006. Document clustering with prior knowledge. In *Proceedings of the 29th Annual International ACM SIGIR Conference*.

KEHLER, A. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI, Stanford, CA.

KNIGHT, K. AND MARCU, D. 2000. Statistics-based summarization—step one: sentence compression. In *Proceedings of the 17th National Conference of the American Association for Artificial Intelligence*. 703–710.

KNOTT, A. AND SANDERS, T. J. M. 1998. The classification of coherence relations and their linguistic markers: an exploration of two languages. *J. Pragmatics 30*, 135–175.

KUDO, T. AND MATSUMOTO, Y. 2003. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. 24–31.

LACATUSU, F., HICKL, A., HARABAGIU, S., AND NEZDA, L. 2004. Lite-GISTexter at *Proceedings of the Document Understanding Conference (DUC'04)*.

LIN, C.-Y. AND HOVY, E. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th Conference of the International Committee on Computational Linguistics (COLING)*.

LIN, C.-Y. AND HOVY, E. 2003. The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL Workshop: Text Summarization (DUC03)*.

MARCU, D. 1998. Improving summarization through rhetorical parsing tuning. In *Proceedings of the Sixth Workshop on Very Large Corpora*. 206–215.

MARCU, D. AND ECHIHABI, A. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*.

MARCU, D. AND GERBER, L. 2001. An Inquiry into the Nature of Multidocument Abstracts, Extracts, and Their Evaluation. In *Proceedings of the Workshop on Automatic Summarization (NAACL'01)*. 1–8.

MCKEOWN, K. R., KLAVANS, J., HATZIVASSILOGLOU, V., BARZILAY, R., AND ESKIN, E. 1999. Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence*. 453–460.

MORRIS, J. AND HIRST, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computat. Ling. 17*, 1, 21–43.

MOSCHITTI, A. AND BEJAN, C. A. 2004. A semantic kernel for predicate argument classification. In *Proceedings of Conference on Computational Natural Language Learning (CoNLL'04)*. 17–24.

NENKOVA, A. AND PASSONNEAU, R. 2004. Evaluating Content Selection in Summarization: the Pyramid Method. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04)*.

NG, V. 2004. Learning noun phrase anaphoricity to improve coreference resolution: issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Asssociation for Computational Linguistics (ACL'04)*.

NICOLAE, C. AND NICOLAE, G. 2006. Bestcut: A graph algorithm for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 275–283.

PALMER, M., GILDEA, D., AND KINGSBURY, P. 2005. The proposition bank: an annotated corpus of semantic roles. *Computat. Ling. 31*, 1, 71–106.

PASSONNEAU, R., NENKOVA, A., MCKEOWN, K., AND SIGELMAN, S. 2005. Applying the Pyramid Method in DUC 2005. In *Proceedings of the Document Understanding Workshop (DUC'05)*.

PRADHAN, S., WARD, W., HACIOGLU, K., MARTIN, J., AND JURAFSKY, D. 2005. Semantic role labeling using different syntactic views. In *Proceedings of the Association for Computational Linguistics 43rd Annual Meeting (ACL'05)*.

RADEV, D. R., JING, H., AND BUDZIKOWSKA, M. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP-NAACL Workshop on Automatic Summarization*.

RILOFF, E. 1996. Automatically generating extraction patterns from untagged text. In *Proceedings of the Conference of the Association for the Advacenmnet of Artificial Intelligence/Innovative Applications of Artificial Intelligence (AAAI/IAAI)*. 1044–1049.

RILOFF, E. AND SCHMELZENBACH, M. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the 16th Workshop on Very Large Corpora*.

SEMEVAL. 2007. Fourth international workshop on semantic evaluations. In *Proceedings of the Association for Computational Linguistics (ACL'07)*.

SENSEVAL-3. 2004. Third international workshop on the evaluation of systems for the semantic analysis of text. In *Proceedings of the Association for Computational Linguistics (ACL'04)*.

SORICUT, R. AND MARCU, D. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

SURDEANU, M. AND TURMO, J. 2005. Semantic role labeling using complete syntactic analysis. In *Proceedings of Conference on Computational Natural Language Learning (CoNLL'05)*.

TURNER, J. AND CHARNIAK, E. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. 290–297.

ZAJIC, D., DORR, B. J., AND SCHWARTZ, R. 2004. BBN/UMD at DUC-2004: Topiary. In *Proceedings of the HLT/NAACL Document Understanding Workshop (DUC'04)*. 112–119.

Received October 2008; revised June 2009; accepted August 2009