Topic Themes for Multi-Document Summarization

Sanda Harabagiu and Finley Lacatusu
Language Computer Corporation
1701 N. Collins Blvd.
Richardson, TX 75080
{sanda, finley}@languagecomputer.com

ABSTRACT

The problem of using topic representations for multi-document summarization (MDS) has received considerable attention recently. In this paper, we describe five different topic representations and introduce a novel representation of topics based on topic themes. We present eight different methods of generating MDS 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 [**INFORMATION STORAGE AND RETRIEVAL**]

General Terms

Algorithms, Performance, Experimentation

Keywords

Summarization, Topic Themes

1. THE PROBLEM

One of the problems of data overload that we are facing today is that there are many documents that cover the same topic. Multi-document summarization (MDS) techniques can address this problem by condensing information found in several documents into a short, readable synopsis, or summary.

Multi-document summaries need to be both informative and coherent. Informativeness is rendered by the methods of selecting the information from documents to incorporate it into the summary. The coherence of the summary is obtained by ordering the information originating in different documents. Much work in summarization dealt with these problems separately. For example, several MDS systems select information for summaries using a cut-and-paste process, whereas [10] proposes that the information that needs to be extracted can be localized through statistical approaches. The ordering of sentences in multi-document summaries was addressed by [3], where three different ordering algorithms were proposed.

In this paper we argue that information selection and ordering in a multi-document summary (MDS) can be based on the structure of the topic covered in the document collection. The topic structure is characterized in terms of topic themes, which are representations of events or states that are reiterated throughout the document collection, and therefore represent repetitive information. The relations that are established between themes determine both the content selection and the ordering of the information. This determines a novel MDS procedure that obtains improved results over previous methods.

Central to the MDS approach proposed in this paper is the notion of topic representation and structure. As first proposed in [11], the topic of a document (or a set of documents) can be represented using a set of terms – known as a topic signature (TS) – that are highly correlated to the topic itself. Under this approach, each topic signature term is assigned an association weight that measures the relatedness of the term to the topic. [6] proposed an extension to this model by considering the relations between topic signature terms. Also a new representation of a topic was proposed in terms of the themes it covers. In a different approach, [1] proposed the use of content models to capture constraints on topic selection and organization for texts in a particular domain.

Our approach represents topics as a structure of themes, which can be considered either linear or graphical. A theme is defined as a cluster of sentences – taken from several different documents – which convey the same semantic information. Sentences corresponding to each theme are extracted using a semantic parser (previously described in [15]) which identifies sentences with common predicates and arguments. Once the themes for a particular document collection are assembled, we identify the set of relations that exist between elements in different themes. These relations are assumed to dictate both (a) the information content to be included in an MDS as well as (b) the order of the themes that are selected.

The remainder of the paper is organized as follows. Section 2 describes previous work on topic representation. In Section 3 we motivate the need for topic themes and propose a themes representation. Section 4 focuses on how the theme structure is used to generate multiple-document summaries, Section 5 presents the experimental results, and Section 6 summarizes the conclusions.

2. TOPIC REPRESENTATION

The representation of topics was the focus of several recent papers. In our experiments we have considered five different topic representations (TRs): (TR_1) representing topics via *topic signatures* (TS_1); (TR_2) representing topics via *enhanced topic sig-*

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

natures (TS_2); (TR_3) representing topics via *thematic signatures* (TS_3); (TR_4) representing topics by modeling *the content structure* of documents; and (TR_5) representing topics as *templates* implemented as a frame with slots and fillers.

TR$_1$. **Topic Representation 1**: *Topic Signatures*. In [11] the topic signature is represented as $TS_1 = \{topic, < (t_1, w_1), ... , (t_n, w_n) >\}$ where the terms t_i are highly correlated to the topic with association weight w_i. The terms are considered to be either stemmed content words, bigrams or trigrams. Term selection and weight association are determined by the use of *likelihood ratio* λ. With the likelihood ratio method, the confidence level for a specific $c = -2log\lambda$ value is found by (a) looking up the χ^2 distribution table, (b) using the value c to select an appropriate cutoff associated weight, and (c) determining the terms selected in the topic signature based on the value c.

To find the candidate topic terms, a set of documents is preclassified into (a) topic relevant texts $\Re$, and (b) topic nonrelevant texts $\tilde{\Re}$. This classification enables the assumption of two hypotheses:
Hypothesis 1 (H_1): $P(\Re|T_i) = p = P(\Re|\tilde{t_i})$ i.e. the relevancy 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 relevancy assuming $p_1 \gg p_2$;
and 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:
$L(H_1) = b(O_{11}; O_{11} + O_{12}, p) \cdot b(O_{21}; O_{21} + O_{22}, p)$
$L(H_2) = b(O_{11}; O_{11} + O_{12}, p_1) \cdot b(O_{21}; O_{21} + O_{22}, p_2)$
where $b(k; n, x)$ represents the binomial distribution[1]. The $-2log\lambda$ value is computed as $-2log\frac{L(H_1)}{L(H_2)}$. Figure 1(a) illustrates the TS_1 topic representation for two different topics T_1 and T_2 evaluated in the DUC multi-document summarization workshop [2].

TR$_2$. **Topic Representation 2**: *Enhanced Topic Signatures*. [6] proposed that topics can be represented by identifying the relevant relations that exist between topic signature terms: $TS_2 = \{topic, < (r_1, w_1), ..., (r_m, w_m) >\}$, where r_i is a binary relation between two topic concepts. Two forms of topic relations are considered: (1) syntax-based relations between the VP and its Subject, Object, or Prepositional Attachments; and (2) C-relations [3] - between events and entities that cannot be identified by syntactic constraints, but belong to the same context. C-relations are motivated by: (a) frequent collocations of certain nouns with the topic verbs or topic nominalizations, and (b) an approximation of the intra-sentential centering, as introduced in [9]. The topic relations are discovered by starting with the topic terms uncovered in TS_1 and selecting a seed syntactic relation between the topic terms. Only nouns and verbs are considered from TS_1. Figure 1(b) illustrates the seed relations, while Figure 1(c) illustrates the enhanced topic representation.

The iterative process of discovering topic relations has four steps:
Step 1: Generate candidate relations: in each document relevant to the seed relation, all syntax-based and C-relations are identified. To discover topic relations we have used a very large corpus of texts: the AQUAINT corpus (LDC Catalog # LDC2002T31) which contains 375 million words corresponding to about 3GB of data. The seed relation becomes a query q that is used by the SMART IR system [4] to generate the set of relevant documents. These documents are processed to identify Verb-Subject, Verb-Object, and Verb-Prepositional Attachment relations. Document processing starts with the identification of named entities. Part of speech (PoS) tags and non-recursive, or basic, noun phrases are identified using the transformation-based learning method reported in [17]. Simple verb phrases (VP) and prepositional phrases (PP) are identified with finite state automata (FSA) grammars. Syntactic relations, such as Verb-Subject, Verb-Object, and Verb-Prepositional Attachment are recognized by another FSA. The C-relations are discovered by creating a salience window for each verb in the document. The NPs of each salience window are extracted and ordered with an ordering relation introduced in [9]. Both syntax-based relations and C-relations are expanded by replacing names with their semantic classes or by replacing words with concepts from a large, hand-crafted ontology of over 200,000 English words.

	T1=PINOCHET TRIAL		*T2=LEONID METEOR SHOWER*	
	Topic Signature TS1		*Topic Signature TS1*	
(a)	pinochet(N)	562	satellite(N)	444
	chile(N)	203	meteor(N)	392
	arrest(V)	104	storm(N)	160
	immunity(N)	100	leonid(N)	119
	argentina(N)	57	comet(N)	87
	garzon(N)	54	shower(N)	69
	dictator(N)	49	earth(N)	65
	london(N)	48	tuttle(N)	59
	crime(N)	46	sky(N)	49
	castellon(N)	43	kelly(N)	46
	request(N)	43	meteorid(N)	43
	Seed Relation for T1		*Seed Relation for T2*	
(b)	arrest – Pinochet		Leonids – shower	
	Topic Signature TS2 for T1		*Topic Signature TS2 for T2*	
(c)	arrest – Pinochet	483	shed – particle	395
	charge – murder	412	predict – shower	321
	react – Chile	216	position – satellites	147
	warrant – judge	114	study – scientists	102
	Topic Signature TS3 for T1		*Topic Signature TS3 for T2*	
(d)	Arrest	483	Meteor	392
	Reaction	216	Shower	321
	Warrant	114	Satellite	147
	Immunity	100	Scientists	102

Figure 1: Topic representations for topics T_1 = *Pinochet Trial* and T_2 = *Leonid Meteor Shower* (a) topic signature TS_1; (b) seed relations for T_1 and T_2; (c) enhanced topic signature TS_2; (d) topic signature TS_3.

Step 2: The candidate topic relations are ranked following a method introduced in [18]. Each relation is ranked based on its *Relevance-Rate* and its *Frequency*. The *Frequency* counts the number of times a relation is identified in the set of relevant documents. *Relevance-Rate = Frequency/Count*, where *Count* measures the number of times an extracted relation is recognized in any document, relevant or not.
Step 3: Select a new topic relation based on the ranking in Step 2.
Step 4: Restart the discovery by using the latest discovered relation for classifying relevant documents. The discovery procedure stops after N=100 iterations or when no new relations are discovered.

TR$_3$. **Topic Representation 3**: *Enhanced Topic Signatures with Themes*. Although topic-relevant terms or relations can be used to capture information about a topic that is repeated throughout a document collection, additional information is needed to produce accurate multi-document summaries [6]. For example, a set of documents focusing on topic T_1=*Arrest of Augusto Pinochet* discusses

[1] $b(k; n, x) = \binom{n}{k} x^k (1-x)^{n-k}$

[2] Document Understanding Conference, http://duc.nist.gov

[3] C-relations can be thought of as context-based relations

not only the arrest itself, but also other themes, e.g. the charges, the extradition request, the international reaction, as well as the reaction of Chilean citizens. Although some of these themes may be general, since they apply to other arrest events (e.g. charges), others may be specific only to the current topic (e.g. reaction of Chilean citizens). A third topic representation that is based on the concept of themes was proposed in [6]: $TS_3 = \{topic, < (Th_1, r_1), ..., (Th_s, r_s) >\}$, where Th_i is one of the themes associated with the topic and r_i is its rank.

The discovery of themes is based on (1) a segmentation of documents produced by the TextTiling algorithm [8], and (2) a method of (i) assigning labels to themes, and (ii) ranking them. [6] considered four cases for theme labeling:

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 theme label is given by the nominalization of the verb, i.e. CHARGES.

Case 2: Several topic relations are recognized in the segment. The label is determined by the relation ranked highest in the topic signature TS_2. for example, if both the relation [*arrest-Pinochet*] and [*charge-murder*] are recognized, only the highest weighted one determines the label, e.g. ARREST.

Case 3: If multiple topics are processed simultaneously, the theme may receive multiple labels.

Case 4: The theme contains topic-relevant terms, but no topic relation. In this case the most relevant noun becomes the label of the theme. For example, if within the paragraph, for TS_1 we encounter the nouns *immunity*, *crime*, and *request*, the theme label will be IMMUNITY, since it has the largest associated weight, as it was illustrated in Figure 1.

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.

Figure 2: A collection of similar sentences.

TR$_4$. Topic Representation 4: *Topics Represented as Content Models*. In addition to topic representations based on TS, we also considered deriving topic themes based on models of the content structure of documents. [1] employs an iterative re-estimation procedure that alternates between (1) creating clusters of text spans with similar word distributions to serve as representatives of within-document topics; 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 topic themes and state transitions capture either (1) orderings within that domain, or (2) the probability of changing from one given topic theme to another. The induction of the topic model is described as:

Step 1. *Initial topic induction,* in which complete-link clustering is used to crate m sentence clusters by measuring sentence similarity with the cosine metric.

Step 2. *The model states and the emission/transition probabilities are determined.* Each cluster corresponds to a state. For each state s_i corresponding to cluster c_i the sentence emission probabilities are estimated using smoothened 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, 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:

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

where m is the number of states and δ_2 is a smoothening 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 2 illustrates samples from a cluster corresponding to topic T_1. The cluster represents the topic representation TR_4.

TEMPLATE Disaster: last week's TORNADOES
Amount Damage: $100 million
Number Dead: 40 / four of the victims
/ a husband, wife, their daughter and her fiancee
Location: Florida / central Florida
Date: last week

(a)

TEXT:
officials in florida have ended the search for a 23–year–old man, bringing the death toll to 40 from last week's tonadoes. funerals are being held across central florida this weekend. four of the victims were buried yesterday, a husband, wife, their daughter and her fiancee. other families spent the day trying to secure belongings from the first heavy rain since the tornadoes. estimates of the damage now exceeds $100 million.

(b)

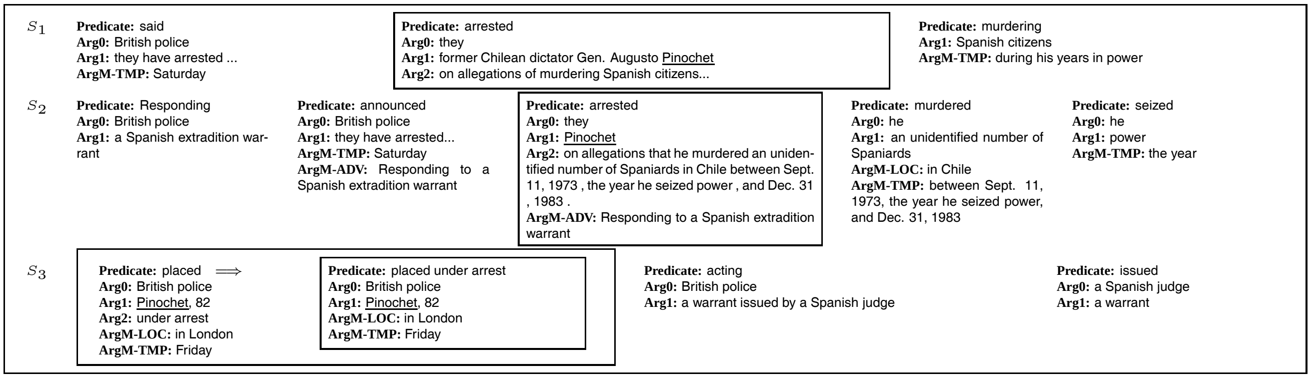
Figure 3: (a) Template representation of topic T_3: "natural disasters"; (b) Text containing information about the topic.

TR$_5$. Topic Representation 5: *Topics Represented as Extraction Templates*. Topics can be represented as a set of inter-related concepts, implemented as a frame having slots and fillers. In Information Extraction, such frames are called *templates* and are populated with information related to the salient facts reported in documents and extracted by the IE systems. For example, if the topic is *"natural disasters"*, Figure 3(a) illustrates a template populated with information extracted from the text illustrated in Figure 3(b).

The idea of representing the topic as a frame-like object was first implemented as underspecified (or "sketchy") scripts, which were used to model a set of pre-defined particular situations, e.g. demonstrations, earthquakes or labor strikes. Since the world contains millions of topics – each which could be described by a script – it is important to be able to generate scripts automatically from corpora. In a first attempt to generate sketchy scripts automatically, [7] proposed using the IS-A and GLOSS lexical relations found in the WordNet[4] lexical database [14] to mine topic relations for topic relevant terms.

The IS-A and GLOSS relations encoded in WordNet participate into lexico-semantic domains between topic-relevant concepts. [7]

[4] Wordnet is a lexical database that encodes the majority of English nouns, verbs, adjectives and adverbs. Each word in WordNet is stored with a set of synonyms known as a *synset*, as well as a definition or *gloss*. In addition to these lexical resources, WordNet incorporates properties of knowledge bases, as it is organized into 24 noun hierarchies and 512 verb hierarchies as well.



S_1

Predicate: said
Arg0: British police
Arg1: they have arrested ...
ArgM-TMP: Saturday

Predicate: arrested
Arg0: they
Arg1: former Chilean dictator Gen. Augusto Pinochet
Arg2: on allegations of murdering Spanish citizens...

Predicate: murdering
Arg1: Spanish citizens
ArgM-TMP: during his years in power

S_2

Predicate: Responding
Arg0: British police
Arg1: a Spanish extradition warrant

Predicate: announced
Arg0: British police
Arg1: they have arrested...
ArgM-TMP: Saturday
ArgM-ADV: Responding to a Spanish extradition warrant

Predicate: arrested
Arg0: they
Arg1: Pinochet
Arg2: 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 .
ArgM-ADV: Responding to a Spanish extradition warrant

Predicate: murdered
Arg0: he
Arg1: an unidentified number of Spaniards
ArgM-LOC: in Chile
ArgM-TMP: between Sept. 11, 1973, the year he seized power, and Dec. 31, 1983

Predicate: seized
Arg0: he
Arg1: power
ArgM-TMP: the year

S_3

Predicate: placed ⟹
Arg0: British police
Arg1: Pinochet, 82
Arg2: under arrest
ArgM-LOC: in London
ArgM-TMP: Friday

Predicate: placed under arrest
Arg0: British police
Arg1: Pinochet, 82
ArgM-LOC: in London
ArgM-TMP: Friday

Predicate: acting
Arg0: British police
Arg1: a warrant issued by a Spanish judge

Predicate: issued
Arg0: a Spanish judge
Arg1: a warrant

Figure 4: The Predicate-Argument structures of the theme illustrated in Figure 2

call these lexico-semantic domains *topical relations* and provide four possible ways of combining the IS-A and GLOSS relations for generating the topical relations: (1) considering a path between a $synset_1$ and any other $synset_2$ in the GLOSS of $synset_1$; (2) considering a path between a $synset_1$ and any $synset_2$ reached through two GLOSS relations; (3) considering a path between a $synset_1$ and any $synset_2$ reached by an IS-A followed by a GLOSS relation; (4) considering a path between a $synset_1$ and a $synset_2$ if there are GLOSS relations from them to a $synset_3$.

The topical relations mined from WordNet have the advantage that they bring forward semantically-connected concepts deemed relevant to the topic. An ad-hoc template generation algorithm was presented in [7]. The algorithm's steps are:

Step 1: Extract all sentences in which one of the concepts traversed by topical relations is present. The concepts from the topical relations are used as seed lexical items for the identification of the template slots.

Step 2: Identify all Subject-Verb-Object (SVO) +Prepositional attachments syntactic structures in which one of the topical concepts is used. For this purpose, we used the phrasal parser implemented for the topic representation TR_2.

Step 3: Add all SVOs in which one referent of a pronoun existing in the SVOs discovered at Step 2 exists. Referents are discovered with the resolution method proposed in [16].

Step 4: Combine the extraction dictionaries with WordNet to classify each noun from the structures identified at Step 2 and Step 3.

Step 5: Generate the semantic profile of the topic. For this reason we compute three values for each semantic class derived at Step 4: (1) *SFreq*: the number of syntactic structures identified in the collection; (2) *CFreq*: the number of times elements from the same semantic class were identified; and (3) *PRel* the probability that the semantic class identifies a relevant slot of the template. Similarly to the method reported in [19], *PRel* = *CFreq*/*SFreq*. To select the template slots the following formula is used:

(CFreq > F1) or ((SFreq > F2) and (PRel > P))

The first test selects roles that come from the semantic categories that are identified with high frequency, under the assumption that this reflects a real association with the topic elaboration in the collection. The second text promotes slots that come from a high percentage of the syntactic structures recognized as containing information relevant to the topic even though their frequency might be low. The values of *F1*, *F2* and *P* vary from one topic to another - we derive them from the requirement that a template should not contain more than 5 slots.

3. THEME REPRESENTATION

Although the documents used to generate a multi-document summary may be relevant to the same general topic, they do not necessarily include the same information. In order to produce exhaustive summaries, MDS systems must be able to identify information that is (1) *common* to multiple documents in the collection, (2) *unique* to a single document in the collection, and (3) *contradictory* to information presented in other documents in the collection. Extracting *all* similar sentences would produce a verbose and repetitive summary, while extracting *some* similar sentences could produce a summary biased towards some sources, as it was noted in [2].

Multi-document summaries based on topic representations similar to the ones presented in Section 2 extract sentences relevant to the topic representations. For example, [10] used the topic representation outlined in (TR1) to extract sentences for MDS based on a *topic signature score* that is equal to the total of signature word scores it contains, normalized by the highest sentence score. When more elaborate topic representations are used, e.g. TR_3, or TR_4, the extractive summarization is based on models of topic themes and shifts between these themes. Results reported in [1] and [6] indicate that this topic representation produces better multi-document summaries. For this reason, we revisit the notion of theme representation and propose herein a more linguistically-motivated definition and representation of themes.

Themes were previously used in the context of MDS [2] to define sets of similar sentences that are detected across a set of documents describing the same topic. Figure 2 illustrates a theme defined through a collection of similar sentences for the topic T_1: *"the Arrest of Augusto Pinochet"*. Each of the sentences listed in Figure 2 communicates about the arrest of Pinochet, which is expressed through the presence of the same predicate, namely the verb "*arrest*", or one of its paraphrases, e.g. the idiom "*place under arrest*", and at least one common argument of the predicate, namely "*Pinochet*". This observation is at the core of our method of representing themes.

Current semantic parsers, similar to the ones described in [5] or [15], are able to recognize all verbal predicates and their arguments similarly to the PropBank [5] annotation [5]. Figure 4 illustrates the predicate-argument structures recognized for the sentences listed in Figure 2. The predicates that were recognized are underlined in Figure 2. The parses were obtained with the method reported in [15][6]. Additionally, the argument list may include functional tags

[5] www.cis.upenn.edu/∼ace

[6] The output of predicate argument parsers generally label arguments sequentially from Arg_0 to Arg_5. Generally, Arg0 would

F_1	**Theme frequency**	The diameter of the cluster corresponding to each theme
F_2	**Predicate signature**	The weight of the predicate in the topic signatureTS_1
F_3	**Argument signatures**	A vector of weights corresponding to the arguments recognized in TS_1
F_4	**Theme relations**	A vector of all the binary relations between the predicate and any argument that is recognized in the enhanced topic signature TS_2
F_5	**Position signature**	A matrix representing the position of the theme in the set of documents
F_6	**Content signature**	A vector corresponding to the content states uncovered by the method reported in [1]. Whenever the predicate of the theme or any of its arguments was present in a content state, the frequency with which they occur is considered in the content signature.
F_7	**Theme position**	A vector with the sentence number in each document where the theme is recognized
F_8	**Theme coverage**	$Theme\_frequency \times \log Inverse\_theme\_frequency$

Table 1: Features used in the selection of candidate themes

from Treebank [7] , e.g. ArgM-DIR indicates a directional, ArgM-LOC indicates a locative, and ArgM-TMP stands for a temporal.

Figure 4 highlights the common predicates between the sentences listed in Figure 2. It is to be noted that for sentence S_3, the predicate "place" was transformed into the idiomatic predicate "place under arrest", which is listed as a synonym of the verb "arrest" in the web resource *dictionary.com* (Merriam-Webster Dictionary of Law). The transformation keeps as Arg1 for "place under arrest" the Arg1 of "place", but Arg2 is no longer recognized. The decision of no longer instantiating the Arg2 for the phrasal predicate "place under arrest" is motivated by the fact that the filler of Arg2 contributed to the recognition of the new predicate. Also, both predicates "place" and "arrest" share the same semantics for the Arg0, Arg1, ArgM-LOC, and ArgM-TMP.

Predicate: ARRESTED (PLACED UNDER ARREST) coverage(15:38)
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

Figure 5: Conceptual representation of a theme.

We generate the theme representation through the following six steps:

Step 1: For every sentence in each document from the collection, the predicate-argument structures are identified. This step also involves the recognition of paraphrases as synonyms or idioms. The resources for synonymy detection are WordNet and *dictionary.com*.

Step 2: All sentences having at least one common predicate with a common argument (e.g. "*arrest*" with argument "*Pinochet*") are clustered together. In this step, the semantic consistency of the other arguments is also checked. For every argument, the consistency is accepted when: (a) the head of the phrase is the same, synonymous, or a name alias; or (b) the argument is a pronoun or a noun phrase that refers to the same entity (e.g. in Figure 4, "*they*" from Arg0 of predicate "*arrest*" in both S_1 and S_2 refers to "*British police*" in S_3); or (c) the arguments are mapped to the same ontological node, when using the ontology employed in TR_2; and (d) the arguments contain the same predicate, having at least one consistent argument. Consistency is sought only between several instantiations of the same argument in the theme. In Figure 4, semantic consistency is sought only for Arg0, Arg1, and Arg2. To find semantic consistency between Arg2 of predicate ARREST in S_1 and Arg2 of the same predicate in S_2 the frame semantics of the noun "allegation" are used. This noun is associated with the frame STATEMENT, for which a FE of "message" is recognized. The message is semantically consistent: in both cases the predicate MURDER is used, for which we have two associated thematic roles: the agent: "he" referring to "Pinochet", and the patient: "Spanish citizens", consistent with "Spaniards", due to mappings to the same concept of the ontology employed in TR_2.

Step 3: Conceptual representations for each cluster are generated. These conceptual representations are similar to the dependency-based representation employed by [2]. The representation consists of (1) the predicate, (2) the semantically consistent argument, and (3) arguments that anchor the predicate in time and location, or describe the cause, manner or effect of the predicate. For example, for the theme illustrated in Figures 2 and 4, the cluster centered around the predicate ARREST would consider arguments Arg0, Arg1, Arg2 (semantically consistent), as well as ArgM-LOC, ArgM-TMP, and ArgM-ADV, as illustrated in Figure 5. Two statistics modeling the coverage of the predicate-argument representation are collected: the number of times the lexeme "arrest" is used as a predicate in the cluster, and the total number of times the same lexeme is recognized throughout the entire collection. For each argument, a list of triplets is recorded as well as a text snippet. The format of the triplets is: (a) the sentence position and document number where the argument was recognized, followed by (b) the word number within the sentence where the argument starts, and (c) the word number where the argument ends.

Step 4: Selection of the candidate themes is made by considering the mappings of the clusters into (1) the topic representation TR_3 and (2) the topic representation TR_4. Both these topic representations correspond to the notions of *theme* and *theme change*. The selection is cast as a binary classification problem that can be solved through inductive methods. We have implemented the classification with binary trees, considering for each candidate theme the features represented in Table 1. For this supervised learning paradigm, the abstracts created by humans are also considered. For training purposes, only themes that are manifested in the human-created abstracts are considered as positive examples. For example, for topic T_1, the themes that were selected were: ARREST, WARRANT, CHARGE, PROTEST, TRIAL, REACT.

Step 5: In a topic, there are meaningful relations between the themes. We have considered two forms of such relations:
(1) Cohesion relations. Often themes co-occur (1) in the same sentence, (2) in successive sentences, or (3) in the same text segment. Themes co-occur in the same sentence because their representative predicates share an argument or one of the predicates belongs to an argument of the other. For example, in Figure 4, both predicates ARREST and MURDER co-occur both in S_1 and S_2. Themes co-occurring in successive sentences are recognized when pairs of

stand for *agent*, Arg1 for *direct object* or *theme*, whereas Arg2 represents *indirect object*, *benefactive*, or *instrument*, but mnemonics tend to be very specific. For example, when retrieving the argument structure for the verb-predicate "arrest", we find Arg0:*police*, Arg1:*criminal*, Arg2: *crime*.

[7] www.cis.upenn.edu/∼treebank

theme-relevant predicates are recognized each in one of the sentences. Cohesive relations between themes belonging to the same segment are identified similarly to the cohesive relations between succeeding sentences. Cohesive relations in the same sentence receive a weight $a = 10$, in successive sentences a weight $b = 5$, and in the same segment a weight $c = 1$.

(2) *Discourse relations*. [13] argues that for some summaries, the structure of discourse helps in selecting better textual units. To study the interaction of discourse relations on our theme representation, we have considered only the CONTRAST and CAUSE-EXPLANATION relations. These two relations were recognized by the same naive Bayes classifiers as the one reported in [13]. Whenever we recognized any of these two discourse relations between a text unit containing one of our selected themes and any other text unit, we would select the relation for inclusion in the theme representation, and later in the summary. The discourse relations are modeled by a three-valued feature. The weights are given by the method presented in [13].

Step 6: The themes are structured into a graph. The nodes are mapped into the conceptual representation of each theme. Figure 5 illustrates such a representation of a theme. The links of the graph correspond to a combination of cohesion and coherence relations. When between two themes we have both a cohesive relation R_1 with weight w_1, and also a coherence relation R_2 with weight w_2, the weight assigned to the link between the themes is equal to $w_1 + w_2$. If only one type of relation exists the link receives the wight of that relation. Unlinked themes are removed from the theme representation.

4. USING TOPIC AND THEME REPRESENTATIONS FOR MDS

Multi-document summarization is performed by (1) extracting sentences that contain the most salient information; (2) compressing the sentences for retaining the most important pieces of information; and (3) ordering the extracted sentences into the final summary. When comparing several MDS methods that use topics or theme representations we have found that most of them contributed mainly to the extraction phase of the MDS process. In total, we have implemented four extraction methods, two ordering methods, and a separate MDS method, based on the theme representations, that performed extraction, compression and ordering simultaneously. Thus we could compare eight MDS methods, as listed in Table 2.

EM$_1$. Extraction Method 1. Similar to [11], sentence extraction is based on topic identification and interpretation, provided by TR_1 in the form of topic signatures TD_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.

EM$_2$. Extraction Method 2. The same procedure 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.

EM$_3$. Extraction Method 3. When TR_3 is employed, we have first selected the segments labeled by the highest scoring theme. To extract sentences, we computed in each segment a sentence topic score based on TS_2 and extracted the sentence with the highest score for each theme.

EM$_4$. Extraction Method 4. Topic representation TR_5 was used in [7] to generate multi-document summaries. In [7], sentences are extracted based on the *importance* of the template partially matching it. 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.

Each Extraction method was used in combination with one of the following two ordering methods for generating MDS:

OM$_1$. Ordering Method 1: Topic representation TR_5 was used in [1] for single-document summarization. Given the content model represented in TR_5, we can learn which themes (represented by the content model's states) should appear in the MDS. As in [1], all the extracted sentences can be tagged by the Viterbi algorithm with a *Viterbi topic label*, or V-*topic* – the name of the state most likely to have generated them. The selection of themes and their order is determined by the probability of generating a summary sentence. This probability is learned by training on abstracts generated by humans. For each document-abstract pair we (1) count the number of sentences that are assigned the same V-*topic* S, and (2) normalize this count by the number of documents containing sentences with V-*topic* S.

OM$_2$. Ordering Method 2: This ordering method, introduced in [3], aims to remove disfluencies from the summary by grouping together topically related themes. It applies only to extraction method EM_3, which has knowledge of themes, since it is based on TR_3. This ordering method defines pairwise relations between themes. Given two themes T_1 and T_2 where T_1 contains sentences $(s_1, s_2, ..., s_n)$ and T_2 contains sentences $(t_1, t_2, ..., t_m)$, we denote by $\#T_1T_2$ the number of pairs of sentences (s_i, t_j) which appear in the same text, and $\#T_1T_2^+$ to be the number of sentence pairs that appear in the same text and are in the same segment. The measure of relatedness between themes T_1 and T_2 is given by the ratio $\#T_1T_2^+/\#T_1T_2$ When the ratio is higher that a predifined threshold [8] the two themes are considered related. A transitive closure of the pairwise relation between themes produces groups of related themes (GRTs). The GRTs are ordered by a chronological ordering (CO) algorithm presented in [3]. In this algorithm, each GRT is assigned a date corresponding to its first publication [9]. The date assignment establishes a partial ordering over the GRTs. When two GRTs have the same date (that is, they are reported for the first time in the same article) we sort them accordingly to the order of presentation in this article. This generates a complete order over the GRTs. Within the GRT, the themes are ordered by applying the CO algorithm to them.

MDS Method	Extraction Method	Ordering Method
MDS_1	EM_1	OM_1
MDS_2	EM_2	OM_1
MDS_3	EM_3	OM_1
MDS_4	EM_4	OM_1
MDS_5	EM_3	OM_2
MDS_6	Theme Representation (Step 4)	OM_1
MDS_7	Theme Representation (Step 4)	OM_2
MDS_8	Theme Representation (Steps 4 and 6)	Traveling Salesman

Table 2: MDS Methods

As illustrated in Table 2, five different combinations of extraction methods and ordering methods lead us to as many MDS methods. Furthermore, the theme representation presented in Section 3 can be used in three more MDS methods. Step 4 of the theme representation selects the themes that need to be extracted. Since the conceptual representation of a theme includes information about the sentences where the themes were described, both ordering methods OM_1 and OM_2 can be used for implementing two additional MDS methods, namely MDS_6 and MDS_7 from Table 2. We considered

[8] We used the same value of the threshold as [3], i.e. $a = 0.6$

[9] Articles released by news agencies are marked with a publication date

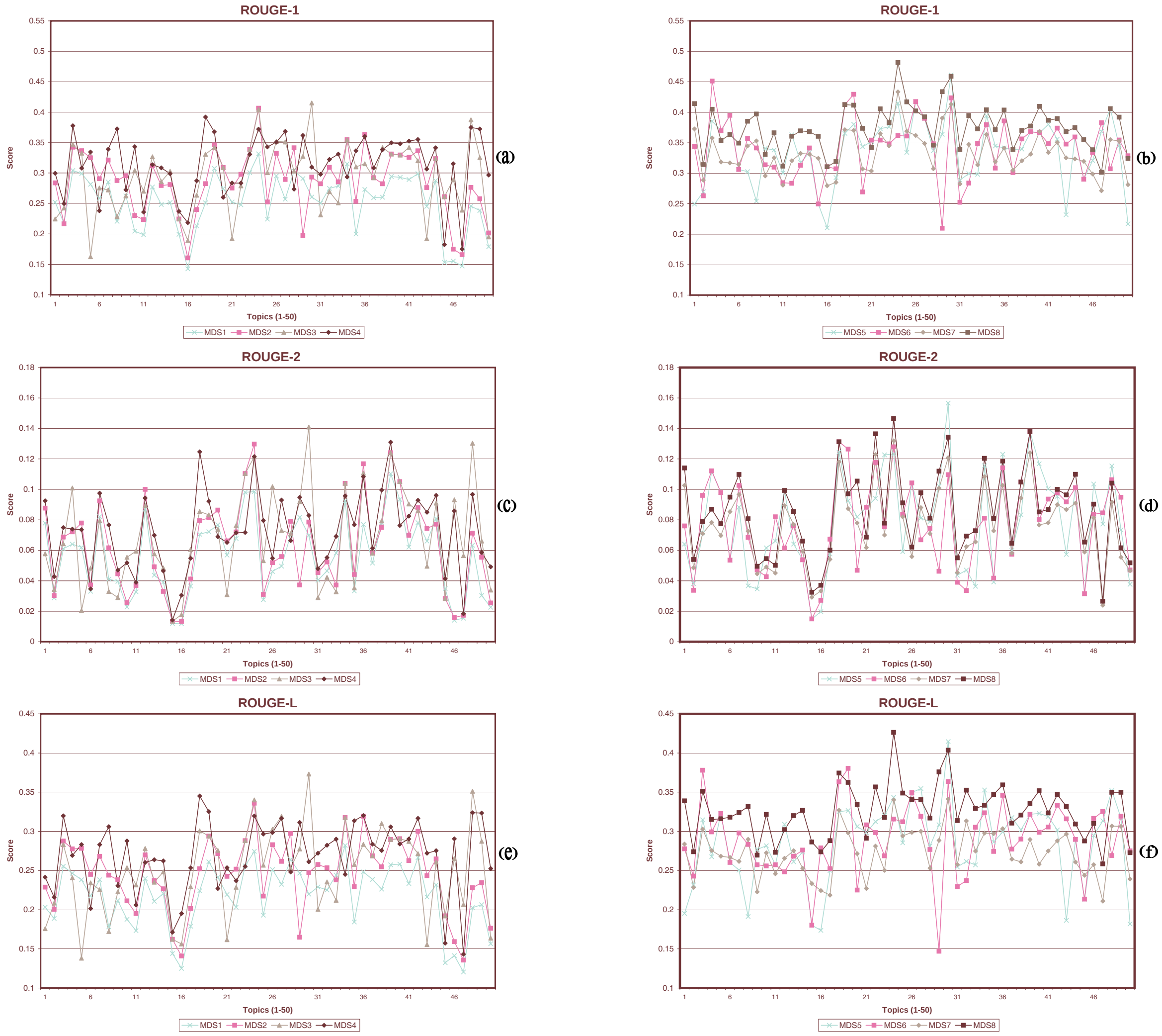


Figure 6: ROUGE Scores for MDS_1 through MDS_8

an eighth MDS method by using the graph representation produced at Step 6 in Section 3. The order between the themes that are produced is obtained by applying the well-known Traveling Salesman Problem on the graph. The Hamiltonian path that is generated is used to select the order of the themes, but the sentences that contain information in the conceptual representation of a theme need to be ordered as well. We do not consider a complete order, but rather select only one sentence that has the highest coverage among the arguments of the theme. When a tie exists, we give priority to the sentence that covers most of the core arguments (Arg0-5) as opposed to the locative or temporal arguments. For example, given the theme illustrated in Figure 5, sentence S_2 is selected, since it covers four arguments (Arg0, Arg1, Arg2, and ArgM-ADV). Sentence S_3 also covers four arguments (Arg0, Arg1, ArgM-LOC and ArgM-TMP), but only two core arguments.

5. EVALUATING MDS

Automatic text summarization has drawn a lot of interest in the natural language processing and information retrieval communities in recent years. Recently, a series of government-sponsored evaluation efforts in text summarization have taken place in both the United States and Japan. These evaluations, known as the Document Understanding Conferences (DUC), have organized yearly evaluations of automatically-produced summaries by comparing the summaries created by 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 [10] – were introduced for both single and multi-document summarization [12]. ROUGE includes four automatic evaluation methods that measure the similarity between summaries: ROUGE-N, ROUGE-L, ROUGE -W, and ROUGE-S. Formally, 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. In this technique, a sequence $R = [r_1, r_2, ...r_n]$ is considered to be a subsequence of another sequence $S = [s_1, s_2, ...s_n]$ if there can be defined a strictly-increasing sequence of indices for S (i.e. $I = [i_1, i_2, ...i_k]$ such that for all $j = 1, 2, ...k$, $s_{ij} = z_j$. The longest common subsequence for

[10] ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation; http://www.isi.edu/∼cyl/ROUGE/

R and S can be defined as the sequence common to both R and S with the greatest length. ROUGE-L is based on the assumption that pairs of summaries with longer LCS scores will be more similar than those summaries with shorter LCS scores. To capture this generalization, if we assume that summary sentences X and Y can be represented as a sequence of words, an LCS-based F-measure can be calculated to estimate the similarity between a reference summary X (of length m) and a candidate summary Y of length n as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m} \; P_{lcs} = \frac{LCS(X,Y)}{n} \; F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

Here, $LCS(X,Y)$ is equal to the length of the LCS of X and Y and β = P_{lcs}/R_{lcs}. LCS can be also used to compute an F-measure for an entire summary, not just a single sentence. The summary-based LCS F-measure can be computed as follows:

$$R_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i,C)}{m} \; P_{lcs} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i,C)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

where u represents a reference summary of m words, C represents a candidate summary of n words, and $LCS_{\cup}(r_i, C)$ is the LCS score of the union LCS between r_i and the candidate summary C.

Figure 6 illustrates ROUGE scores for the 50 topics evaluated in DUC-2004 for each of the eight summarization methods we propose in this paper. Results from three different ROUGE scores are depicted: ROUGE-1 (Figure 6(a) and 6(b)), ROUGE-2 (Figure 6(c) and 6(d)) and ROUGE-L (Figure 6(e) and 6(f)). Although performance does vary from topic to topic regardless of the summarization system employed, the graphs above show that the best results were obtained when MDS were generated using method MDS$_8$, followed closely by MDS$_7$, and MDS$_6$, as shown in Figure 7.

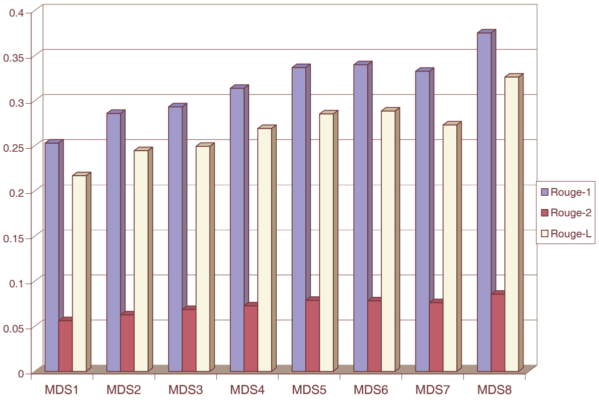


Figure 7: Summary of the scores in Figure 6.

6. CONCLUSIONS

In this paper we investigated five topic representations that were used before in MDS and proposed a new representation based on topic themes. We have presented a total of six new MDS methods that use different information extraction and information ordering methods. We have shown how both extraction and ordering for MDS can be improved when topic themes are available as part of the input. Finally, we have evaluated these summarization methods using the ROUGE scoring package and the topics and document collections featured in the DUC 2004 evaluation.

The theme representations proposed in this paper are based on the shallow semantic information provided by semantic parsers, a source of linguistic information much more sophisticated than those employed by previous thematic representations. Additionally, we have represented themes in a graph-like structure (determined by both coherence relations and cohesion relations) that improve the quality of ordering information for MDS. Although the idea of using cohesion and coherence for summarization is not new in of itself, it has not previously been applied to thematic representations based on predicate-argument structures.

In future work, we plan to extend thematic representations in order to incorporate additional semantic information, to recognize theme paraphrases, and to represent themes using macro-predicates that correspond to natural classes of predicates. Additionally, we plan to expand the number of coherence relations that we recognize between themes. Finally, we intend to incorporate compression methods into MDS and to customize a compression method that is based on topic themes.

7. REFERENCES

[1] R. Barzilay and L. Lee. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. In *HLT-NAACL 2004: Proceedings of the Main Conference*, pages 113–120, 2004.

[2] R. Barzilay, K. R. McKeown, and M. Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA, June 16–20 1999.

[3] R. Barzilay, K. R. McKeown, and M. Elhadad. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. In *Journal of Artificial Intelligence Research*, pages 35–55, July 2002.

[4] C. Buckley, M. Mitra, J. Walz, and C. Cardie. SMART High Precision: TREC 7. In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, pages 285–298, 1998.

[5] D. Gildea and M. Palmer. The necessity of syntactic parsing for predicate argument recognition. In *Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL-02)*, pages 239–246, Philadelphia, PA, 2002.

[6] S. Harabagiu. Incremental Topic Representations. In *Proceedings of the 20th COLING Conference*, Geneva, Switzerland, 2004.

[7] S. Harabagiu and S. Maiorano. Multi-Document Summarization with GISTexter. In *Proceedings of the Third LREC Conference 2002 (LREC 2002)*, Canary Islands, Spain, June 2002.

[8] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[9] M. Kameyama. Recognizing referential links: An information extraction perspective. In *Proceedings of the ACL/EACL '97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pages 46–53, 1997.

[10] C.-Y. Lin and E. Hovy. Identifying Topics by Position. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 283–290. Association for Computational Linguistics, March 31 - April 3 1997.

[11] C.-Y. Lin and E. Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th COLING Conference*, Saarbrücken, Germany, 2000.

[12] C.-Y. Lin and E. Hovy. The potential and limitations of automatic sentence extraction for summarization. In D. Radev and S. Teufel, editors, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[13] D. Marcu and A. Echihabi. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, July 2002.

[14] G. A. Miller. WordNet: a lexical database for English. *Communications of the Association for Computing Machinery*, 38(11):39–41, 1995.

[15] A. Moschitti and C. A. Bejan. A semantic kernel for predicate argument classification. In *Proceedings of CoNLL-2004*, pages 17–24. Boston, MA, USA, 2004.

[16] V. Ng and C. Cardie. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 4Oth Meeting of the Association for Computational Linguistics*, 2002.

[17] G. Ngai and R. Florian. Transformation-Based Learning in the Fast Lane. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 40–47, 2001.

[18] E. Riloff. Automatically generating extraction patterns from untagged text. In *AAAI/IAAI, Vol. 2*, pages 1044–1049, 1996.

[19] E. Riloff and M. Schmelzenbach. An Empirical Approach to Conceptual Case Frame Acquisition. In *Proceedings of the Sixteenth Workshop on Very Large Corpora*, 1998.