Incremental Topic Representations

Sanda Harabagiu
Language Computer Corporation
Richardson, Texas 75080, USA
sanda@languagecomputer.com

Abstract

We consider the problem of modeling information about the topic discussed in a text. We describe in this paper two incremental enhancements of the topic signatures introduced by (Lin and Hovy, 2000). The first enhancement considers topic representations in terms of relevant topic relations instead of relevant terms. The second enhancement is based on ranking the topic themes. Topic representations are integrated in two NLP applications: Information Extraction and Multi-Document Summarization. Our experiments show that incorporating the two enhanced representations in both applications produces substantial improvements over previously-proposed topic representations that can be acquired automatically.

1 The Problem

The topic of a text plays an important role in NLP applications. The problem is that researchers have used several topic representations, sometimes in an unsystematic way. We argue that topic representations should be based on information that can be acquired automatically from a topic-relevant corpus. Topic signatures, as introduced in (Lin and Hovy, 2000) provide such an example. In this paper we present two incremental enhancements to topic signatures and show how they can be used with good results in Information Extraction (IE) and Multi-Document Summarization (MDS). Section 2 presents the notion of topic signature and its enhancement to topic-relevant relations. Section 3 details the automatic procedure of acquiring the new topic representation. Section 4 presents the second enhancement, namely the topic themes. Section 5 discusses the application of these topic representations to IE and MDS. Section 6 summarizes the conclusions.

2 Topic Signatures

Most documents can be characterized as a sequence of sub-topical discussions that occur in the context of one or several main topic discussions[1]. To determine the boundaries of each subtopic, the TEXTTILING approach reported in (Hearst, 1997) can be used, relying on lexical cohesion relations. However, this method does not indicate any relations between a subtopic and any of the main topics from the document. In contrast, when a set of documents about the same topic is provided, a topic characterization can be acquired automatically. (Lin and Hovy, 2000) propose a method for characterizing topics through a lexically determined *topic signature* TS defined as $\{topic, <(t_1, w_1), ...(t_n, w_n)>\}$, where each of the terms t_i is highly correlated with the *topic* with an association weight w_i. Lin and Hovy define the terms as either stemmed content words, bigrams or trigrams. The selection of the terms as well as the assignment of the association weight is determined by the use of the likelihood ratio.

But topics are not characterized only by terms, there are also relations between topic concepts that need to be identified. By assuming that the largest majority of these relations take place between verbs/nominalizations and other nouns, the topic representations can be produced in two iterations. In the first iteration only nouns and verbs are considered as terms in the topic signature TS_1 generated by the method reported in (Lin and Hovy, 2000). In the second iteration, the topic signature TS_2 is defined as $\{topic, <(r_1, w_1), ...(r_m, w_m)>\}$, where r_i is a binary relation between two topic concepts. We start with a single seed topic relation r_s, and then we discover additional topic relations with the methodology detailed in Section 3.3. The associated weight of the relations is determined by the frequency with which it is recognized in the collection of documents relevant to the topic.

To determine the seed relation r_s we perform a four-step procedure on the topic signature TS_1:
Step 1: Filter out topic outliers. Terms from TS_1 that have an extreme variation from the mean weight are filtered out. In our experiment, we have considered nine different document collections, each representative for a different topic. We used the training documents from the Hub-4 Event-99 evaluations (Hirschman et al., 1999). The topics are: (T1) *NATURAL DISASTERS*, (T2) *DEATHS*, (T3) *BOMBINGS*, (T4) *MARKET CHANGES*, (T5) *COURT CASES*, (T6) *ILLNESS OUTBREAKS*; (T7) *MEDICAL RESEARCH*, (T8) *ELECTIONS* and (T9) *MOVEMENT OF PEOPLE*. Figure 1 illustrates the first 15 terms of topic signatures TS_1 obtained for the topics T3 and T4 as well as the outliers that are filtered out.

[1] The technique for multi-paragraph segmentation reported in (Hearst, 1997) is based on this observation.

Step 2: Morphological expansion. Terms from TS_1 are grouped into a set of nouns N^{TS} and a set of verbs V^{TS}. For each nominalization from N^{TS}, information from CELEX (CELEX , 1998) is used to determine V^m, the verb from which it was morphologically derived. Unless V^m and the nominalization are homonyms, the nominalization is eliminated from N^{TS}, while the pair $<V^m,\{N^m\}>$ is added to the set V^{TS}. The set $\{N^m\}$ represents all the nominalizations of V^m available from CELEX. An Example of verb-nominalization pair for the topic T3 is $<$*"explode"*,{*"explosion"*}$>$.

	TOPIC=BOMBINGS Topic Signature TS1		TOPIC=MARKET CHANGES Topic Signature TS1	
outliers	say(V)	1.544e+05	market(N)	1.669e+05
	have(V)	1.420e+05	government(N)	1.482e+05
	be(V)	1.336e+05	company(N)	1.214e+05
	bombing(N)	54223	index(N)	85047
	people(N)	47479	investor(N)	81626
	kill(V)	32064	be(V)	78427
	attack(N)	29932	say(V)	75760
	official	27781	rise(V)	71826
	government(N)	22975	analyst(N)	59883
	explosion(N)	12253	change(V)	51975
	police(N)	17503	growth(N)	42947
	explode(V)	16755	expect(V)	41652
	peace(N)	16536	raise(V)	38232
	hamas(N)	13666	bond(N)	36159
	come(V)	13380	economy(N)	35512

Figure 1: Topic signatures and outliers.

Step 3: Semantic Normalization. Many of the terms from N^{TS} are named entities or semantic concepts that can be normalized semantically. For this purpose, we retrieve all the sentences in which any of the nouns from N^{TS} occurs in the document collection and apply a Named Entity Recognizer (NER) that was trained to identify names from 15 different categories [2]. Whenever a noun N^i from N^{TS} is tagged by the NER, it is replaced by the pair $<NC,\{N_i\}>$. The name class NC enables the normalization of all nouns of the same name class N_a, N_b, ... into a single pair $<NC,\{N_a, N_b, ...\}>$. Similarly, we perform a conceptual unification of all the other nouns. For this purpose, we use a hand-crafted ontology that encodes 150 concepts that subsume 22,000 words. Thus every noun from N^{TS} that is not a name belongs to some conceptual pair $<CC,\{N_x, N_y, ...\}>$, where CC is the concept that subsumes N_x and N_y. If such subsumptions are not identified in the ontology, a dummy concept C_0 is considered to subsume all un-mapped nouns. The processing from steps 2 and 3 create an intermediary representation of the topic signature TS^{int}, defined as $\{$*topic,* $<V^{TS},N^{TS}>\}$. Each element from V^{TS} is a <Verb-{Nominalizations}> pair, whereas each element from N^{TS} is a pair consisting of a semantic class (a name class or an ontological concept) and a list of nouns from TS_1 that are tagged/mapped in that semantic class. Figure 2 illustrates the TS^{int} for the topic T3=*BOMBINGS*.

Verbs – Nominalizations	*Semantic Class – Nouns*
<explode,{explosion}>	<GROUP,{people,government}>
<kill,{}>	<ORGANIZATION,{hamas}>
<attack,{attack}>	<NATIONALITY,{palestinian,isreali}>
<injure,{injury}>	<ARTIFACT,{bomb,building}>
<try,{trial}>	<C0,{peace,security,time}>

Figure 2: Intermediary topic signature.

Step 4: Selection of topic seeds. For every verb V_i which represents the first element of a pair from V^{TS} and for every semantic class C_j which represents the first element of a pair from N^{TS} create a possible relation $r_{i,j}$=[V_i-C_j], unless C_j is the dummy class C_0. In the latter case, the relations will take the form $r_{i,k}$=[V_i-N_k], where N_k is any of the nouns listed in the dummy class C_0. For every relation $r_{i,j}$ compute $R_{i,j}$, which counts the number of times verb V_i or any of its nominalizations collocates with any of the nouns associated with C_j in N^{TS}. The collocation window consists of the sentence containing V_i and its immediately preceding and succeeding sentences. The relation $r_{a,b}$ with the largest $R_{a,b}$ across all relations becomes the seed relation. When $r_{a,b}$ is of type [V_i-C_j], and C_j is not a name class, $r_{a,b}$ is replaced by $r_{a,s}$=[V_i-N_S], where noun N_s has the highest contribution to the magnitude of $R_{a,b}$ among all nouns from C_j. Figure 3 lists the seeds we obtained for each of the nine topics from Event-99.

Topic	*Seed Relation*
T1 = NATURAL DISASTERS	[hit – tornado]
T2 = DEATHS	[kill – PERSON]
T3 = BOMBINGS	[explode – bomb]
T4 = MARKET CHANGES	[raise – QUANTITY]
T5 = COURT CASES	[accuse – crime]
T6 = ILLNESS OUTBREAKS	[spread – infection]
T7 = MEDICAL RESEARCH	[discover – cure]
T8 = ELECTIONS	[campaign – PERSON]
T9 = MOVEMENT OF PEOPLE	[fly – LOCATION]

Figure 3: Seed Relations.

3 Topic Relations

3.1 Trigger Words

Two forms of topic relations are considered: (1) syntax-based relations between the VP and its *Subject*, *Object*, or *Prepositional Attachment*[3]; and (2) *C-relations* which represent relations between events and entities that cannot be identified by syntactic constraints. 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 (Kameyama, 1997).

For each topic relation R_t=[V_i-N_j] we assume that events are recognized by a small set of *trigger words*,

[2] 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

[3] When nominalizations are used in topic relations, we deal only with prepositional attachments between two noun phrases.

typically verbs or nominalizations lexicalizing the events. We discover trigger words for V_i only when the relation R_t has the syntax type S_t (*Verb-Subject*) or (*Verb-Object*). The trigger words are acquired by:
[1] retrieving all the other nouns N_k (relevant entities) from the corpus for which [V_i-N_k] has the syntax type S_t; and
[2] finding all the other verbs V_l for which [V_l-N_k] has the syntax type S_t.[4] The set of verbs V_l represent the trigger words for the relation R_t.

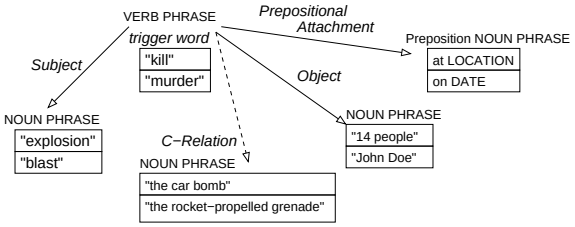

Figure 4: Topic relations.

Figure 4 illustrates both forms of topic relations. The Figure shows also instances of trigger words and of relevant entities. It is to be noted that sets of topic relations that share the same VP can be grouped into *extraction patterns*. This is equivalent to the observation from (Yangarber et al., 2000) stating that extraction patterns can be decomposed in a set of binary relations.

3.2 Motivating Example

We motivate the need to consider additional binary extraction relations with the example illustrated in Figure 5. The text from Figure 5 belongs to a document relevant for topic T3=*BOMBINGS*. The only syntax-based relations that can be identified in this text are the two verb-object relations. They both relate trigger verbs ("kill" and "wound") to entities that represent victims. However, the text also contains information about the type of bomb: "a truck crammed with explosives" and about the location: "downtown Colombo". As indicated in Figure 5, relations between the trigger words and these two relevant entities are C-relations.

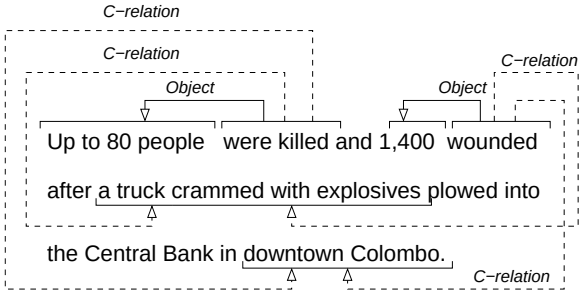

Figure 5: Example of topic relations.

[4]Trigger words for nominalization N^V of verb V are obtained by using the morphological derivations of the trigger words for V.

3.3 The Model of Discovering Topic Relations

Our model of discovering extraction relations employs a very large corpus of texts. We used the AQUAINT corpus (LDC Catalog #LDC2002T31), which contains 375 million words correlating to about 3GB of data. The idea[5] is that starting with a seed set of extraction relations we could separate the corpus into: (1) a set of relevant documents containing the seed relations, and (2) a complementary set of non-relevant documents. Instead of considering the entire AQUAINT corpus we used only documents retrieved when formulating a query q that characterizes the scenario domain. The Information Retrieval (IR) system we employ is SMART (Buckley et al., 1998). For example, for the "bombing" domain, we used a single-keyword query:{ *"explode"*}[6] and retrieved 2000 documents.

The trigger words for the new topic relation are also derived. For example, in the case of the Topic T3 the derived trigger words for the seed relations are *"explode"* and *"detonate"*. The corresponding relevant entities for the BOMB_*Word* consists of 3 words: "bomb", "grenade"and "mine". The trigger words, the corpus D_q and the seed relation are the only inputs to our procedure that discovers both syntax-based and C-relations from the documents. The discovery procedure has the following steps:

Step 1: *GENERATE CANDIDATE RELATIONS*.

□ Syntax-based relations:

a. From each document from D_q we extract all Verb-Subject, Verb-Object, Verb-Prepositional Attachment relations. For this purpose we used a document parser that is based mainly on finite state technology. Document processing starts with the identification of named entities. Part-of-speech (POS) tags and non-recursive, or basic, noun phrases (NPB) are identified using the TBL method reported in (Ngai and Florian, 2001). 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.

b. Each syntax-based relation is expanded by considering three possibilities:

(i) Replace each word with its root form, e.g. the verb "wounded" with "wound" and the noun "trucks" with "truck".

(ii) Replace the word with any of the concepts that subsume it in a hand-crafted, general ontology, e.g. "truck" may be replaced by VEHICLE, ARTIFACT, or OBJECT.

(iii) Replace each name with its corresponding named entity class, e.g. "Los Angeles" with LO-

[5]This idea was first reported in (Yangarber et al., 2000).

[6]The query q is generated by considering the verb or nominalization from the most-recent added topic relation.

CATION and "Bank of America" with ORGANIZATION. Figure 6 illustrates the expansion of relations.

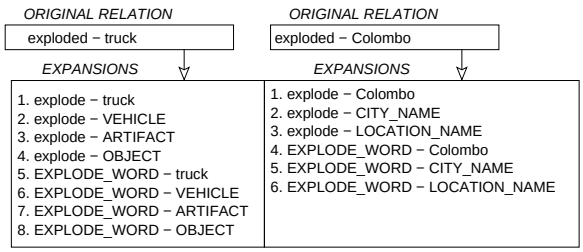

Figure 6: Expansions of two relations.

□ Salience-based C-relations:

a. Additional topic relations may be discovered within a salience window for each verb. The window is created by considering K sentences preceding and succeeding the sentence containing the verb. In our experiments we set $K = 2$.

b. The NPs of each salience window are extracted and ordered. The basic underlying hypothesis is that C-relations between a verb and an entity from its domain are similar to the anaphoric relations between entities in texts. Therefore, as illustrated in Figure 7, six possible text spans, TS1-6, can be defined. The prominence of entities related to the anchor can be approximated by a left-to-right ordering. Entities are first retrieved from text span T_i before being retrieved from T_{i+1}. The same approximation was introduced by (Kameyama, 1997) for resolving coreference relations.

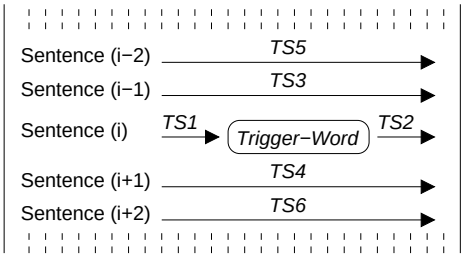

Figure 7: Ordering salient entities.

c. Candidate extraction relations are generated in each salience window. First, [*Trigger-Verb* $\rightarrow NP_i$] relations are created and expanded for each candidate entity. The expansions are done similarly as for syntax-based relations. However, when considering one expansion for [*Trigger-Verb* $\rightarrow NP_j$] the expansion is allowed only if it was not already introduced by any expansion for any [*Trigger-Verb* $\rightarrow NP_k$], with $k < j$. For syntax-based relations, repetitive expansions do not exist. This is the rationale for disabling repetitions of C-relation expansions.

Step 2: *RANK CANDIDATE RELATIONS*. Following the method introduced in (Riloff, 1996) each relation is ranked based on its *Relevance-Rate* and its *Frequency*. 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 non-relevant. Additionally, we maintain only relations with $\beta < Count/MaxCount < \gamma$, where $MaxCount$ indicates the total number of instances for the most common relation, to avoid noise or uninformative relations.[7]

Step 3: *SELECT A NEW TOPIC RELATION*. The ranking from Step 2 determines an order between all candidate extraction relations. Only the first relation is selected and added to the set of discovered relations. The relations and its ranks constitute the new topic signature TS_2. Initially, TS_2 the seed relation.

Step 4: *RESTART THE DISCOVERY*. The new set of discovered relations is used to re-classify the documents from D_q into relevant and non-relevant. A new iteration resumes by jumping to Step 2, where the relations are ranked again, based on the new set of relevant documents determined by the query derived from the verb/nominalization of the most recently added relation. The discovery procedure stops after N = 100 iterations, or when no new relations are discovered.

Syntax−based Topic Relations

1. VP(EXPLODE−word)−Subject−> NP(ARTIFACT | OBJECT)
2. VP(EXPLODE−word)−Object−> NP(BOMB−word)
3. VP(DAMAGE−word)−Object−> NP(ARTIFACT | VEHICLE | BOMBING)
4. VP(INJURE−word)−Object−> NP(NUMBER)
5. VP(EXPLODE−word)−Prep_Attach{in | outise}−> NP(ARTIFACT)
6. VP(EXPLODE−word)−Object−> NP(DATE | TIME)
7. VP(CLAIM−word)−Object−> NP("responsability")
8. VP(INJURE−word)−Object−> NP(HUMAN | LIVING | PERSON−Name)
9. VP(CLAIM−word)−Subject−> NP(HUMAN | LIVING | PERSON−Name)
10. VP(KILL−word)−Object−> NP(NUMBER)

(a)

C−Relations

1'. VP(INJURE−word)−−> NP(BOMB−word)
2'. VP(KILL−word)−−> NP(BOMB−word)
3'. VP(EVENT)−Object−> NP(BOMB−word)
4'. VP(EXPLODE−word)−−> NP(ARTIFACT | OBJECT)
5'. VP(EXPLODE−word)−−> NP(MALE | HUMAN | LIVING | PERSON−Name)
6'. VP(EXPLODE−word)−−> NP(CITY−Name | COUNTRY−Name)
7'. VP(EXPLODE−word)−−> NP(ORGANIZATION−Name)
8'. VP(EXPLODE−word)−−> NP(DATE | TIME)
9'. VP(INJURE−word)−−> NP(HUMAN | LIVING)
10'. VP(KILL−word)−−> NP(NUMBER)

(b)

Figure 8: Topic relations for T3: (a) top 10 syntax-based relations; (b) top 10 C-relations.

3.4 Examples of Discovered Relations

Figure 8(a) lists the top 10 syntax-based relations discovered whereas Figure 8(b) lists the top 10 C-relations discovered. From Figure 8(a) we find that the system discovered that ARTIFACTS or OBJECTS determine explosions (relations 1 and 5); which are also determined by BOMBS (relation 2). Different

[7] We used $\alpha = 0.7$, $\beta = 0.01$, and $\gamma = 0.4$.

numbers of people or living beings are injured or killed (relations 4, 8, and 10). Someone claims responsibility (relations 7 and 9).

From Figure 8(b) we find that C-relations characterize additional relevant information, for example: injuries and killings are related to bombs (relations 1' and 2'), which adds up to the fact that people are the ones killed (relations 4', 8', and 10'). Explosions occur at LOCATIONS and on DATES (relations 6' and 8'). The perpetrators are HUMANS, sometimes MALES and they might belong to or target an ORGANIZATION(relation 7'). Throughout Figure 8 the notation VERB|NOUN-*word* refers to the trigger words to the relevant entities associated with the VERB|NOUN.

4 Topic Themes

There are NLP applications for which topic representation as a collection of relevant relations is not sufficient. This is because topic relations capture only the most characteristic and repetitive information about a topic. We argue that additional information is of interest, especially for such applications as multi-document summarization. For example, a document focusing on topic T3=*BOMBINGS* discusses also other themes, e.g. arrests, suspects, security measures, bomb delivery and detonation methods. Some of these themes may be as well central to other topics, e.g. arrests and indictments are central to topic (T5) *COURT CASES*. However, some themes may not be covered by any of documents in the collection. For some NLP applications such as multi-document summarization, information about which themes are more characteristic for a topic in a collection is important. For this purpose, we considered a third representation of topic signatures. TS_3 is defined as $\{topic, <(Th_1, r_1), ...(Th_s, r_s)>\}$, where Th_i represents one of the themes associated with the topic and r_i is its rank.

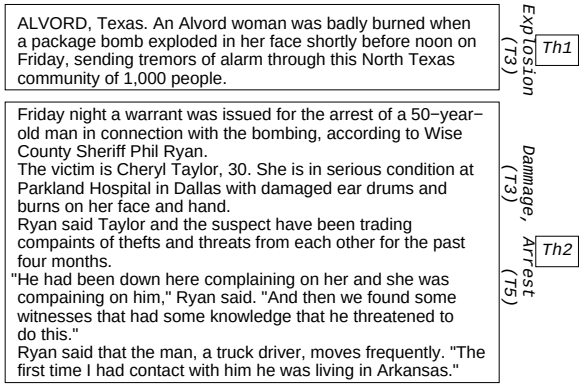

Figure 9: Topic themes.

Figure 9 illustrates the segmentation of a document discussing topic T3=*BOMBINGS* into topic themes. The segmentation was produced by the TEXTTILING algorithm (Hearst, 1997) when term similarity was adjusted to consider identity for topic words belonging to the same set of trigger words or compatible concepts. The themes of topics are labeled automatically. There are four different cases for label assignment of a theme for topic T:

Case 1: A single topic relation pertaining to the topic T is recognized throughout the theme segment. One of the trigger words that facilitates the recognition of the relation becomes the theme label. For the example from Figure 9, theme Th_1 is labeled as *"Explosion"* because the relation [*explode-bomb*] was identified and the trigger word *"explode"* was more frequently responsible for the recognition of this relation in the topic collection than was *"detonate"*.

Case 2: Several topic relations are recognized in the same theme. The label is determined by the topic relation which is ranked highest in the topic signature $TS_2(T)$.

Case 3: Topic relations pertaining to other topics are recognized as well. The theme receives additional labels, determined by the new topics recognized. For the example illustrated in Figure 9, the second theme has two labels, one pertaining to topic T3=*BOMBINGS*, the other pertaining to topic (T5)=*COURT CASES*.

Case 4: The theme contains relevant concepts for the topic T but no topic relation of T is recognized. In this case the UNKNOWN label is assigned.

The ranking of the topic themes for topic T considers that (1) relations pertaining to any other topic $T' \neq T$ contribute less to the ranking than relations from the signature of T; and (2) whenever no relation from T is identified in a theme, but only relevant concepts exists, the ranking of that theme is lower than that of themes containing such relations. Consequently we computed the rank of a theme as $R(th_i) = \sum w(r_j^T) + k_1 \times \sum w(r_k^{T'})$ whenever relations r_j^T pertaining to the topic T are identified in th_i and relations $r_k^{T'}$ of any other topic are also identified; k_1 represent the ratio between the largest weight of any relation $r_k^{T'}$ and the smallest weight of any relation r_j^T.

5 Application to Information Extraction and Summarization

IE systems have the purpose of extracting domain-specific information from natural language texts. The extracted information is organized as database entries for subsequent retrieval and processing. Each entry in the database is defined by a *templette*, enabling the definition of a particular class of events of interest to the topic. For example, the templette for topic T3=*BOMBINGS* is represented by the following list of slots:

Slot Name	Slot Description
BOMB	The description or type of bomb.
PERPETRATOR	The alleged, suspected, claimed, or known perpetrator.
DAMAGE	A description of the physical damage to objects other than persons.
INJURY	Identifiers of persons injured.
DEAD	Identifiers of persons killed.
LOCATION	The most specific event location.
DATE	Description or indication of the time of the bombing.

TOPICS	Syntax-Based Relations			Syntax-Based Relations + C-Relations			Handcrafted Relations		
	P	R	F1	P	R	F1	P	R	F1
NATURAL DISASTERS	77.9%	60.4%	68.1%	65.3%	88.9%	75.3%	70.5%	74.6%	72.5%
DEATHS	63.9%	65.5%	60%	51.5%	80.78%	62.9%	78.2%	44.8%	57%
BOMBINGS	66.2%	40.8%	50.5%	61.3%	58.2%	59.7%	76.9%	45.2%	56.8%
MARKET CHANGES	78.4%	85.9%	82%	80.3%	84.4%	82.3%	80.3%	83.5%	81.9%
COURT CASES	55.7%	37%	44.5%	40.3%	73.67%	52.1%	76.2%	35.3%	48.3%
ILLNESS OUTBREAKS	72.4%	59.46%	65.3%	70.3%	79%	74.4%	71.5%	75.4%	73.4%
MEDICAL RESEARCH	78.4%	75.2%	76.8%	77.1%	82%	79.5%	78.2%	76%	77%
ELECTIONS	68.8%	73.1%	70.9%	60.2%	97%	74.3%	73.5%	69.4%	71.4%
MOVEMENT OF PEOPLE	65.3%	64.5%	64.9%	64.8%	89.5%	75.2%	77.2%	63%	69.4%
Micro-average	69.6%	62.4%	64.7%	63.4%	81.4%	**70.6%**	75.8%	63%	**67.5%**

Table 1: Results for the identification of topic relations.

For each of the nine topics we had templette definitions available and moreover, we had access to their slots filled with correct information from the training data of the Event-99 evaluations. This information enabled us to create a gold standard annotation for each topic. For each slot filler, we have annotated (1) its source in the document and (2) a topic relations that could enable its identification. Sometimes, the slot filler originated in several places in the document. In this case, a topic relations was manually annotated for each instance of the text snippet. The annotations were used to evaluate the procedure of discovering topic relations. We compared the discovered relations against the gold standard. The results are listed in Table 1. The precision P counts the number of times any of the identified relations correspond to a gold-standard annotation. The recall R measures how many of the gold-standard annotations were matched by any relation. We also used $F1 = \frac{2P \times R}{P+R}$ to measure the performance of identifying topic relations. We have considered three possibilities: (1) only syntax-based relations; (2) the combination of syntax-based relations with C-relations; and (3) relations that were handcrafted for the IE systems we employed (hidden citation)[8]. The $F1$-score for the combination of relations is higher than the $F1$-score of the handcrafted relations, mainly because of the higher recall with which these relations are identified. Although less precise than the hand-crafted relations or the syntax-based relations alone, the combination of relations manages to identify more topic-relevant information. Since C-relations are responsible for an average increase 3.11% in $F1$-score, we were interested in the percentage C-relations in the topic signatures. They range from 18% for (T9) *MOVEMENT OF PEOPLE* to 34% for T3=*BOMBINGS*.

By adding in average 13.37 new relations we can discover in general 3.11% more topic-relevant information. This data-driven method of generating topic signatures enhances the recall of topic information, in contrast with the topic-relevant information produced by experts, which emphasizes the precision of the identified data. A much higher improvement was obtained when the coreference module of the IE system was employed. When pronominal coreference was solved, syntax-based relations were recognized with an average $F1$-score of 88.3%. Pronominal coreference determines the recongition of syntax-based relations and C-relations with an average $F1$-score of 95.4%. In the same conditions handcrafted relations were identified with an $F1$-score of 90.3%.

Topic-relevant information is also important for Multi-Document Summarization (MDS). Therefore we integrated topic relations and topic themes in the MDS system (Harabagiu and Maiorano, 2002) which represents topics with information available from WordNet(Fellbaum, 1998). Topic relations are discovered from the definitions of WordNet synsets with a methodology which was reported in (Harabagiu and Maiorano, 2002). MDS is performed in three stages, which involve: (1) the identification of topic-relevant information and extraction of sentences that contain it; (2) sentence compression; and (3) ordering sentences originating in different documents. Our topic representation interacts with both the first stage and the third stage.

Sentence extraction is based on the recognition of topic relations in the documents. Sentences that contain at least two topic relations were marked-up. Additionally, we used the entity and event coreference module of the IE system to identify the reference to the same event. For each set of sentences describing the same event, we selected only the one that scored highest when we considered the weights of the topic relations it matched. The ranking of the themed provides with an ordering of the extracted sentences regardless of which document they originate from.

We evaluated the multi-document summaries on the 59 document sets used in DUC-2002. Each multi-document summary generated was compared with a gold-standard summary created by humans. To compute the quantitative measures of overlap between the system-generated summaries and the gold-standard summary, the human-created summary was segmented by hand into *model units* (MUs), which are informational units that express one self-contained fact. MUs are sometimes sentence clauses, sometimes entire clauses. In contrast, the summaries generated by the summarization systems were automatically segmented into *peer units* (PUs) - which

[8]These relations correspond to the topic representation of the knowledge engineers that have developed the IE system.

are always sentences. Precision is calculated as the number of PUs matching some MU divided by the number of PUs in the automatic summary. Recall is calculated as the number of MUs that is completely covered by some PUs over to total number of MUs. We compared the results of the MDS when using the topic signatures TS_3 with the results obtained with the topic representation based on information from WordNet:

Topic Representation	P	R	F1
WordNet-based (in DUC)	20.6%	20.7%	20.65%
TS_3	32.4%	35.8%	34.02%

The $F1$-score increase of 13.37% indicates that the topic representation TS_3 produces more informative multi-document summaries than those generated based on information from WordNet. This score does not measure the coherence of the summaries, which is determined by the theme rankings. To measure coherence, similarly to the DUC evaluations, we counted the number of sentences which where in the wrong place because they indicate a strange time sequence, suggest a wrong cause-effect relationship or just don't fit topically with neighboring sentences. The coherence score is obtained by the fraction of misplaced sentences over the number of sentences in each summary. For the WordNet-based topic representation we obtained a coherence score of 0.19, whereas for the topic representation TS_3 we obtained 0.11, which indicates more coherent summaries.

6 Conclusions

In this paper we present two new topic representations which capture (1) the most relevant relations for a topic, and thus the concepts that participate in the relations; and (2) the sub-topics, or the interactions between a topic and different themes. These two representations combine in a unique way (1) a method for acquiring topic signatures as relevant words; (2) a method for acquiring topic-relevant relations; and (3) a method for multi-paragraph topic segmentation. The resulting representations have been integrated in an IE system and in a MDS system, producing improvements of up to 3.11% for IE, when no coreference resolution was available, 5.1% when pronouns were resolved in IE and 13.37% for summarization. In the future we plan to use these topic representations for answering complex questions in the context of a predefined topic.

7 Acknowledgements

This work was supported by an ARDA grant. I would like to thank Mihai Surdeanu, John Lehmann, Paul Aarseth and Luke Nezda for helping with the implementation of the methods described herein.

References

C. Buckley, M. Mitra, J. Walz and C. Cardie. 1998. SMART High Precision: TREC7. In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*:285-298.

CELELX. 1998. www.ldc.upenn.edu. Consortium for Lexical Resources, University of Pennsylvania.

C. Fellbaum. 1998. WordNet: An Electronic Lexical Database and Some of its Applications. *MIT Press.*

Sanda Harabagiu and Steven Maiorano. 2002. Multi-Document Summarization with GISTexter. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002).*

M. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23:33-64.

Lynette Hirschman, Patricia Robinson, Lisa Ferro, Nancy Chinchor, Erica Brown, Ralph Grishman, Beth Sundheim 1999. Hub-4 Event99 General Guidelines and Templettes.

Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000).*

Megumi Kameyama. 1997. Recognizing Referential Links: An Information Extraction Perspective. In *Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts (ACL-97/EACL-97).*

Grace Ngai and Radu Florian. 2001. Transformation-Based Learning in The Fast Lane. In *Proceedings of the North American Association for Computational Linguistics (NAACL 2001)*:40-47.

Ellen Riloff. 1996. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*:1044-1049.

Roman Yangarber, Ralph Grishman, Pasi Tapainen and Silja Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*: 940-946.