Multi-Document Summarization with GISTEXTER

Sanda M. Harabagiu*, Steven J. Maiorano†

*Language Computer Corporation
Dallas TX 75206 USA
sanda@languagecomputer.com

†University of Sheffield
Sheffield S1 4DP UK
stevemai@mac.com

Abstract

This paper presents the architecture and the multidocument summarization techniques implemented in the GISTEXTER system. The paper presents an algorithm for producing incremental multi-document summaries if extraction templates of good quality are available. An empirical method of generating ad-hoc templates that can be populated with information extracted from texts by automatically acquired extraction patterns is also presented. The results of GISTEXTER in the DUC-2001 evaluations account for the advantages of using the techniques presented in this paper.

1. Introduction

One of the major problems faced when searching information from vast on-line sources stems from the fact that the same topic is covered in multiple documents, each adding a new perspective. Current indexing and retrieval methods do not aim at bringing forward the different perspectives, but rather at exploiting the redundancy of information in documents of similar content. A user may be potentially helped to see at a glance both similarities and differences in the information content when provided with Multi-Document Summarization (MSD) techniques. MDS can be viewed as either as (1) an extension of single-document summarization of a collection of documents covering the same topic; or (2) an application of Information Extraction (IE) since it uses information extracted from the documents to generate a summary that takes into account several different perspectives of the extracted information.

Single-documents summaries are produced with the goal of presenting the most important content in a condensed way. Two possible techniques can be used to create summaries: (1) sentence extraction, followed by sentence compression; or (2) identification of the relevant information followed by generation of a textual summary from the facts that need to be included. In (Mani and Maybury 1999) a representative set of papers focusing on the first technique are collected whereas (Radev and McKeown 1998) and (Lin and Hovy 2000) report on implementations of the second technique. Multi-document summaries produced by extending the first technique present the problem of ordering the sentences extracted from different documents in order to obtain a coherent summary. This problem was addressed in (Barzilay et al.2001) and a strategy of combining cohesive indicators along with chronological ordering was reported. The extension of single-document summarization techniques based on topic identification for dealing with multiple documents involves the notion of *topic representation.*

Topics can be represented as a set of inter-related concepts, implemented as a frame having slots and fillers.

TEMPLATE
Doc_NR: CNN19980301.1000.0329
Event: <Natural_Disaster−CNN19980301.1000.0329−1>
Comment: Prototypical
<Natural_Disaster−CNN19980301.1000.0329−1> :=
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)

TOPIC SIGNATURE: victim, damage, estimate, flood, tornado, dead, week

(c)

Figure 1: (a) Template representation of the "natural disasters" topic; (b) Text containing information about the topic; (c) Topic signature for "natural disasters".

In the Information Extraction technology, 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 1 illustrates a template populated with information extracted from the text illustrated in Figure 1(b). An alternative representation of a topic was proposed in (Lin and Hovy 2000), with the goal of modeling the minimum amount of knowledge required to effectively identify concepts related to a topic. This representation, called *topic signature*, associates a target concept (i.e. the topic) with a vector of related terms (i.e. the signature). Each $term_i$ from the signature has an associated weight w_i. (Lin and Hovy 2000) report on an automatic method of signature term extraction and weight estimation. Figure 1(c) illustrates the signature terms for the natural disasters topic, ob-

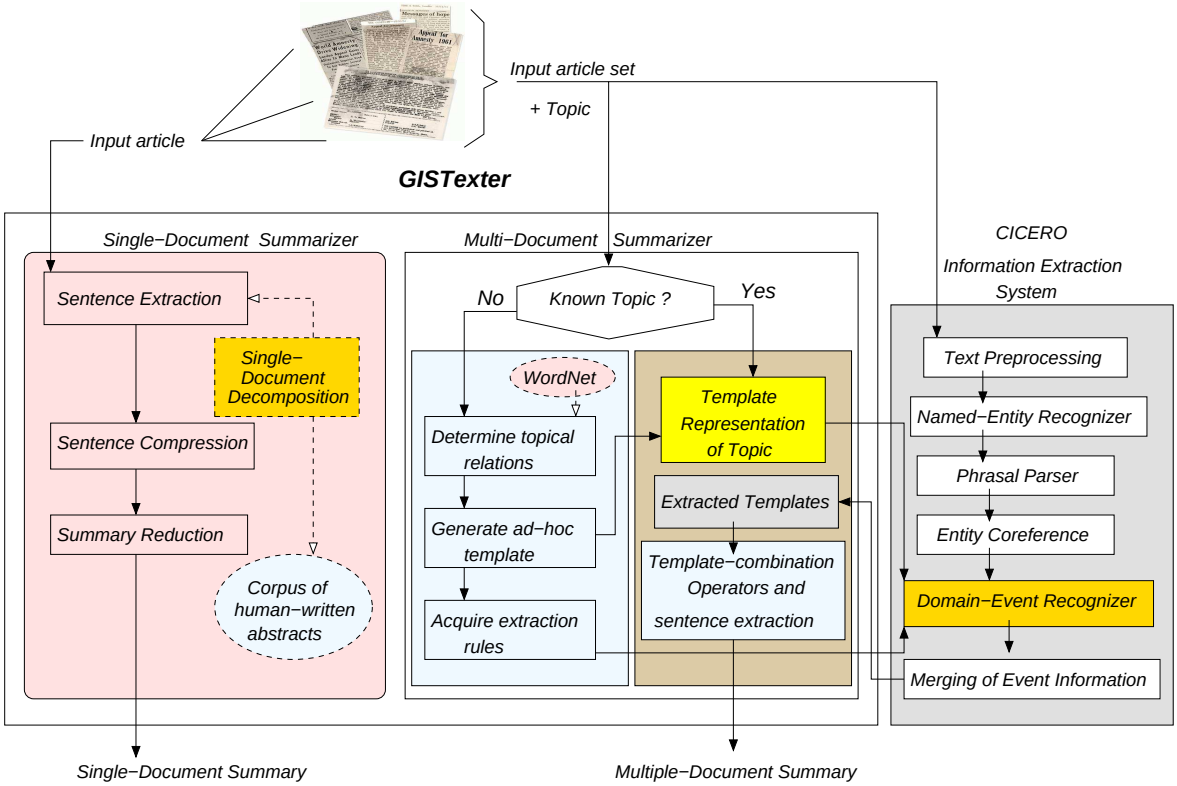

Figure 2: Architecture of GISTEXTER

tained with the method reported in (Lin and Hovy 2000).

When topic signatures are known, they can be used to improve the quality of sentence extraction. Moreover, unlike templates, topic signatures can be extracted automatically by using statistical methods. However, this topic representation does not address the issue of possible similarities and differences between the salient facts. Our experiments show that such knowledge is paramount to producing coherent multi-document summaries. A way of deriving possible relations between salient facts is by relying on templates extracted from the texts. This idea was implemented in SUMMONS for the terrorism domain and the results reported in (Radev and McKeown 1998) indicate that it generates informative and coherent multi-document summaries that cover different perspectives of each topic. The drawback of this method is that it requires (1) a manually generated template; (2) an IE system that extracts information for the specific topic and its template representation; and (3) empirical methods for finding all possible relations between extracted templates.

In this paper we present a third alternative, that combines the advantages of both topic representations. Whenever we have a template representation for a given topic and we have trained an IE system to work for that domain, we extract salient information and populate templates, enabling the generation of multi-document summaries informing about all the relations between salient facts. When new topics are presented, we generate an ad-hoc template which combines statistical information with topical relations mined from the WordNet lexical database. To populate the ad-hoc templates we acquire extraction rules and derive possible relations between the salient facts that were extracted. To be able to evaluate this summarization technique we could either use the human assesses results from DUC-2001 or compared our method to multi-document summarization techniques that rely on sentences extraction and compression. In the latter case we could use the single-document summarization module implemented in GISTEXTER and apply to its output (1) topic signature information and (2) sentence ordering techniques. In this paper we preferred to present only the results originating the the DUC-2001 evaluations.

The rest of the paper is organized as follows. Section 2 presents the architecture of GISTEXTER, our single-document and multi-document summarization system. Section 3 details the IE-based multi-document summarization whereas Section 4 presents the technique for deriving ad-hoc templates. Section 5 reports and discusses the experimental results and Section 6 summarizes the conclusions.

2. The architecture of GISTEXTER

GISTEXTER is a summarization system implemented for the evaluations of the Document Understanding Conferences (DUCs)[1]. The architecture of the system is shown in Figure 2. Input to the system is either a single document or a collection of documents sharing the same topic. When a summary of a single document is sought, GISTEXTER first extracts the key sentences, similarly to most single-document summarizers. The *sentence extraction* function

[1] See http://www-nlpir.nist.gov/projects/duc/

is learned, using the technique of *single-document decomposition.* This technique analyzes the features of human-written abstracts of single documents. In the second stage, to further filter out un-necessary information, the extracted sentences are compressed. In the final stage a *summary reduction* is performed, to trim the whole summary to the length of 100 words. Figure 3 illustrates a single-document summary produced by GISTEXTER.

SQUADS of workers fanned out across storm − battered Louisiana yesterday to begin a massive rebuilding effort after Hurricane Andrew had flattened whole districts, killing two people and injuring dozens more.
The government estimated it would cost Dollars 20bn − Dollars 30bn to tidy and rebuild in Florida.
Louisiana state officials said they had no overall count of storm − related injuries but initial estimates reckoned fewer than 100.
Most of the storm 's fury was spent against sparsely populated farming communities and swampland in the state, sparing it the widespread destruction caused in Florida, where 15 people died.

Figure 3: Single-document summary produced by GISTEXTER.

When multi-document summaries need to be created, the processing takes additionally into account the topic of the document set. Sometimes the topic is well-known and may be already implemented in Information Extraction (IE) systems. In this case an IE system identifies all the information that needs to be used in the multi-document summary. Other times the topic is completely new, and the summary is generated by modeling the topic in an ad-hoc manner.

GISTEXTER produces multi-document summaries by relying on the output of the CICERO IE system[2]. CICERO, as reported in (Surdeanu and Harabagiu 2002) produces unsurpassed quality of extraction because it combines the role of linguistic extraction patterns with coreference knowledge. For multi-document summarization, this means that the templates generated by CICERO are easily mapped into text snippets from the texts, in which pronouns and other anaphoric expressions are resolved. These text snippets can be used to generate coherent, informative multi-document summaries.

To extract information from a set of documents, CICERO needs to have a template representation of the topic. The template slots are filled whenever textual information relevant for the topic is identified. To recognize each topic-relevant event and entity, CICERO first pre-processes the text, by tokenizing the article and recognizing the part-of-speech and attributes of each word against a rich dictionary structure. Next, all names from the article are categorized by a named entity recognizer which taggs *Red Cross* as an *Organization* and *Florida* as a *Location.* A phrasal parser brackets all noun and verb phrases, to enable the recognition of linguistic patterns that relate to the topic. Since anaphoric expressions are often used, before matching the text against linguistic patterns, coreference resolution takes place.

Linguistic patterns are matched to identify the topic-relevant information. For example, for the topic of "natural disasters", the rule [*Casualty-expression* {*to*|*from*} *$Number* {*from*|*because-of*} *Disaster-word*] is matched against the snippet "the death toll to 40 from last week's tornado" in the text from Figure 1(b). Other extraction patterns are matched against the text and populate the rest of the template illustrated in Figure 1(a). CICERO extracts all the templates from the article collection and keeps mappings from the template slots the the text snippets containing information that fills the slots. These text snippets are indicators of the summary content.Figure 4 illustrates the 50-word long, the 100-word long and the 200-word long multi-document summaries generated by GISTEXTER for a collection of articles dealing with "natural disasters".

Hurricane Andrew , August 1992 , is claimed to be the costliest natural disaster in US history. It hit the Bahamas , Forida and along the Gulf of Mexico from Alabama to eastern Texas. 15 people died in Florida, four in Bahamas, two in Louisiana. US insurers expected to pay out Dollars 7.3bn .

(a)

Hurricane Andrew , which hit in August 1992 , was one of the fiercest in the US in decades and is claimed to be the costliest natural disaster in US history . It hit the Bahamas , Forida and headed west along the Gulf of Mexico from Alabama to eastern Texas. 15 people died in Florida , four deaths had been reported in the Bahamas , two people had died in Louisiana . There were estimates that the storm caused more than Dollars 20bn of damage in Florida and Louisiana . The government estimated it would cost Dollars 20bn − Dollars 30bn to tidy and rebuild in Florida . US insurers expected to pay out an estimated Dollars 7.3bn .

(b)

Hurricane Andrew , which hit in August 1992 , was one of the fiercest in the US in decades and is claimed to be the costliest natural disaster in US history . It hit the Bahamas , Forida and headed west along the Gulf of Mexico from Alabama to eastern Texas. Andrew , ripped roofs off houses, smashed cars and trucks , snapped power lines and uprooted trees , making billions of dollars of property damage in southern Florida . Associated tornadoes devastated Laplace , 20 miles west of New Orleans in Louisiana . 15 people died in Florida , four deaths had been reported in the Bahamas , two people had died in Louisiana . There were estimates that the storm caused more than Dollars 20bn of damage in Florida and Louisiana . The government estimated it would cost Dollars 20bn − Dollars 30bn to tidy and rebuild in Florida , and to care for residents displaced by the storm . Some 2m people remained without electricity . In Louisiana it inflicted severe damage on rural communities , severe damage in small coastal centres such as Morgan City , Franklin and New 75 people had been injured . US insurers expected to pay out an estimated Dollars 7.3bn . On the Florida losses alone , Hurricane Andrew became the most costly insured catastrophe in the US .

(c)

Figure 4: Multiple-document summary produced by GISTEXTER: (a) 50-word summary; (b) 100-word summary; (c) 200-word summary.

Whenever the topic of the collection of documents has not been previously encoded in the CICERO IE system and no template representation of the topic exists, we need to perform some additional processing to gist the missing information. Thus we need to generate in an ad-hoc manner: (1) the template and (2) the extraction rules that enable CICERO to identify the relevant information. To this end, we have developed a methodology for generating an ad-hoc template based on the topical relations that can be identified from WordNet (Miller 1995). When the template is known, several possible methods of acquiring extraction rules can be applied, e.g. the methods reported in (Yangarber and et al.2000) (Riloff and Jones 1999) or (Harabagiu and Maioarano 2000). For GISTEXTER, we applied the techniques reported in (Harabagiu and Maioarano 2000).

With an ad-hoc template available, CICERO's domain-event recognizer acts in the same way as for topics that are encoded in the IE system. Moreover, entity coreference takes place for new topics also, since the coreference methods implemented in CICERO are topic-independent. The

[2]CICERO is an ARDA-sponsored on-going project that studies the effects of incorporating world knowledge into IE systems. CICERO is being developed at Language Computer Corporation.

quality of the extraction is not be as good as in the case of previously studied topics because additional semantic knowledge is required to correctly merge incomplete templates. Nevertheless, for multi-document summarization, the extraction quality for ad-hoc templates is reasonable, as it determines acceptably coherent summaries. Example of multiple-document summaries produced by GISTEXTER for a new topic, namely the *"mad cow disease"*, are illustrated in Figure 5.

The mad cow disease eats holes in the brains of its victims. The incurable disease eats holes in the brains of its victims ; The suspicion is that the disease was transmitted through cattle feed . Another fear is that BSE could cause illness in humans .

(a)

Mad cow disease was diagnosed only in 1986 . The incurable disease eats holes in the brains of its victims ; There is a human variant of spongiform encephalopathy , known as Creutzfeldt − Jakob disease . Because of the disease , the U.S. Department of Agriculture 's Animal and Plant Health Inspection Service banned imports of cattle , embryos and bull semen from Great Britain in July (1990) . Similar embargoes have been been imposed by Australia , Finland , Israel , Sweden , West Germany and according to the agriculture ministry , and the European Community has proposed a ban on exports of British cattle older than 6 months .

(b)

Mad cow disease was diagnosed only in 1986 . The incurable disease eats holes in the brains of its victims ; in late stages a sick animal may act skittish or stagger drunkenly . The suspicion is that the disease was transmitted through cattle feed , which used to contain sheep by − products as a protein supplement . There is a human variant of spongiform encephalopathy , known as Creutzfeldt − Jakob disease . Both BSE and CJD are caused by mysterious particles of infectious protein called prions . Because of the disease , the U.S. Department of Agriculture's Animal and Plant Health Inspection Service banned imports of cattle , embryos and bull semen from Great Britain in July (1990) , said Margaret Webb , a USDA spokeswoman in Washington . Similar embargoes have imposed by Australia , Finland , Israel , Sweden , West Germany and New Zealand , according to the agriculture ministry , and the European Community has proposed a ban on exports of British cattle older than 6 months . THE CONDITION has been found in Britain 's sparsely−scattered antelope population , the government has admitted (1992) It has also been found in domestic cats and their larger relations , the cheetah and the puma , in eland and nyala , and in the gemsbok and the Arabian oryx .

(c)

Figure 5: Multiple-document summary produced by GISTEXTER for the "mad-cow disease" topic: (a) 50-word summary; (b) 100-word summary; (c) 200-word summary.

3. Information Extraction-based Multi-Document Summarization

Information Extraction (IE) is a technology that targets the identification of topic-related information in free text and translates it into database entries. Typically, IE systems extract around 10% if a document textual content (cf. (Hobbs and et al.1997)). This represents a compression ratio that qualifies extraction templates for multi-document summarization. This observation was previously employed in the design of the architecture of the SUMMONS multi-document summarization system (Radev and McKeown 1998). In SUMMONS, summarization is viewed as a two-tiered process: (a) *conceptual* and (b) *linguistic* summarization. Conceptual summarization deals with content selection whereas linguistic summarization is concerned with linguistic realization of the content.

To perform conceptual summarization, SUMMONS uses the templates produced by IE to apply a set of *content planning operators* on them for combining the extracted information. These operators, fully detailed in (Radev and McKeown 1998) detect *change of perspective*, *contradiction*, *information addition* or *refinement*. The application of each operator is decided by a set of heuristics, specially crafted for each topic and for each given corpus. The resulting combined templates are then translated into *functional descriptions* (FDs), which are conceptual representations of the template meanings. FDs are used by the linguistic component of SUMMONS that relies on a lexicon and a grammar of English to realize the conceptual representation into a sentence. The linguistic component consists of a lexical choser, which determines the high-level sentence structure of each sentence and the words that realize each semantic role. SUMMONS incorporates the FUF/SURGE (Elhadad 1993) sentence generator.

In GITEXTER we decided to use IE templates for multi-document summarization in a different way. First we considered not only the populated templates alone, but also the mapping into the text snippets that are the source of their slot fillers. Second, since coreference information is also used to fill slots, we keep pointers to the coreference chains that contain any entity that fills a template slot. Thus for each Template T_i having the slots TS_i^1, TS_i^2, ..., TS_i^n we keep two additional forms of information: (1) the text snippet $TextS_i^j$ that matched one of the extraction rules, and thus enabled the filling of a slot TS_i^j; and (2) all the entities from the text that corefer with the information filling any slot $TextS_i^j$. Figure 6 illustrates a snapshot of populated templates and their mappings. The Figure illustrates some coreference chains as well. Both text snippet information and coreference information is made available by the CICERO IE system.

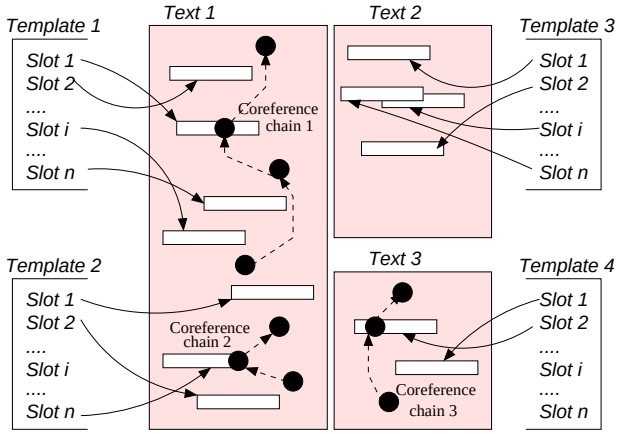

Figure 6: Mappings between extracted templates and text snippets. Whenever a relevant text snippet contains an anaphor, pointers to all other entities with which it corefers are kept in a coreference chain.

To generate multi-document summaries we use two observations: (1) the order in which relevant text snippets appear in the original articles accounts for the coherence of the documents; and (2) to be comprehensible, summaries need to include sentences or sentence fragments that contain the antecedents of each anaphoric expression from relevant text snippets. Since all articles contain information about a given topic, it is very likely that a large percentage of the templates share the same filler for one of the slots. In the case of the "natural disasters" topic, this filler

was " *hurricane Andrew*". We call this filler the *dominant event* of the collection. Additionally, we are interested in the templates extracting information about other events that may be compared with the dominant event in the collection. Thus templates are classified into four different sets: (a) $Templates_1$ - templates about the dominant event that originate in documents that contain relevant information about related events; (b) $Templates_2$ - other templates about the dominant event; (c) $Templates_3$ - templates about non-dominant events that originate in articles that contain information about the dominant event; and (d) $Templates_4$ - other templates.

To generate a multi-document summary of length L GISTEXTER extracts sentences from the document set in four different increments. The rationale for choosing four increments is based on the four different summary lengths imposed by the DUC evaluations, e.g. 50-word, 100-word, 200-word and 400-word long summaries. Since it is not know apriori how many templates are extracted nor what is the cardinality of each $Templates_i$ set, for each summary increment we perform at least one comparison with the target length L to determine if the resulting summary needs to be reduced or not. The IE-based multi-document summary is produced by the following algorithm:

Algorithm IE-based MD-Summarization (L)

Step 1: Select the most representative templates. To this end, for each template T_i from $Templates_j$, with $1 \leq j \leq 4$, for each slot TS_i^j we count the frequency with which the same filler was used to fill the same slot of any other template. The *importance* of T_i is measured as the sum of all frequency counts of all its slots. This measure generates an order on each of the four sets of templates. Whenever there are ties, we give preference to the template that has the largest number of mapped text snippets traversed by coreference chains. Template T_0 is the most important template from $Templates_1$. If $Templates_1$ is null, the same operation is performed on $Templates_2$.

Step 2: Summary-increment 1.

Select sentences containing the text snippets mapped from T_0 in the order in which they appear in the text from where T_0 is selected. If anaphoric expressions occur in any of these sentences, include sentences containing their antecedents in the same order as in the original article.

if length(summary) $> L$ generate appositions for dates and locations and drops the corresponding sentences.

if length(summary) $> L$ drop coordinated phrases that do not contain any of the mapped text snippets.

while length(summary) $> L$ drop the last sentence.

Step 3: Summary-increment 2.

For each slot from T_0 that has other fillers in some other template from $Templates_1$ or $Templates_2$, add the sentence containing the corresponding mapped text snippet immediately after the sentence mapped by template T_0 for the same slot. If anaphoric expressions occur in any of these sentences, include sentences containing their antecedents in the same order as in the original article. Continue this process until either (1) the length of the summary is larger than $L-1$ or until there are no more sentences to be added.

Step 4: Summary-increment 3.

Add sentences mapped by the most important template from $Templates_3$. Repeat the process as at Step 2 until length L is reached or no more sentences can be added.

Step 5: Summary-increment 4.

Add sentences mapped by the most important template from $Templates_4$. Repeat the process as at Step 2 until length L is reached or no more sentences can be added.

Figure 7 illustrates the inter-leaving of extracted sentences that each summary increment produces in the resulting multi-document summarization.

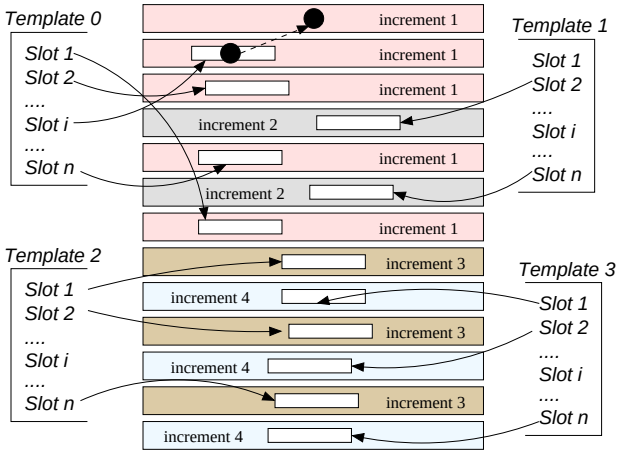

Figure 7: Multi-document summarization produced by four different summary increments.

4. Ad-hoc Extraction for Multi-Document Summarization

Whenever the topic of a document collection is not encoded in an IE system, the Algorithm presented in Section 3. cannot be applied. Two main sources of information are missing: (1) the topic template-representation; and (2) the mappings between template slots and text snippets. In (Harabagiu and Maioarano 2000) we have shown that if the template representation of a topic is known, linguistic patterns that identify the mappings of the template slots into text snippets can be acquired automatically. In this paper, we focus on the mechanism of generating the template representation of the topic.

The idea of representing the topic as a frame-like object was first advocated in the late 70's by DeJong (DeJong 1982), who developed a system called FRUMP (Fast Reading Understanding and Memory Program) to skim newspaper stories and extract the main details. The topic representation used in FRUMP is the *sketchy script*, which model a set of pre-defined particular situations, e.g. demonstrations, earthquakes or labor strikes. Since the world contains millions of topics, it is important to be able to generate sketchy script automatically from corpora. In addition some of the current large-scale lexico-semantic knowledge bases may be used to contribute information for the generation of the topic templates. In our methodology, we have employed WordNet (Miller 1995), the lexical database that encodes a majority of the English nouns, verbs, adjectives and adverbs.

4.1. Extracting Topical Relations from WordNet

WordNet is both a thesaurus and a dictionary. It is a thesaurus because each word is encoded along with its synonyms in a synonyms set called *synset*, representing a *lexical concept*. WordNet is a dictionary because each synset is defined by a gloss. Moreover, WordNet is a knowledge base because it is organized in 24 noun hierarchies and 512 verb hierarchies. Additionally WordNet encodes three meronym relations (e.g. HAS-PART, HAS-STUFF and HAS-MEMBER) between nouns and two causality relations (e.g. ENTAILMENT and CAUSE-TO) between verbs. However, there are no direct relations between the concepts used in any of the template representation of the topics encoded in the CICERO IE system. Nevertheless we noticed that chains of lexico-semantic relations can be mined from WordNet to account for the connection between any pair of template concepts of known topics. To illustrate how such chains of relations can be mined, we first consider two of the relations already encoded in WordNet and then show how additional relations can be uncovered as lexico-semantic chains between two concepts pertaining to the same topic. We call these lexico-semantic chains *topical relations*.

The sources of topical relations

In WordNet, a synset is defined in three ways. First it is defined by the common meaning of the words forming the synset. This definition relies on psycholinguistic principles, based on the human ability to disambiguate a word if several synonyms are presented. Second, the synset is defined by the attributes it inherits from its super-concepts. Third, a glossed definition is provided to each synonym. A GLOSS relation connects a synonym to its definition. We believe that glosses are good sources for topical relations, since they bring forward concepts related to the defined synset. We consider four different ways of using the glosses as sources for topical relations:

1. We extend the GLOSS relation to connect the defined synset not only to a textual definition but to each content word from the gloss, and thus to the synset it represents. For example, the gloss of synset {*bovine spongiform encephalitis, BSE, mad cow disease*} is *(fatal disease of cattle that affects the central nervous system; causes staggering and agitation)*. A GLOSS relation exists between the defined synset and *fatal, disease, cattle, affect, central nervous system, staggering* and *agitation.*

2. Each concept from a gloss has its own definition, and thus by combining the GLOSS relations, we connect the defined synset to the defining concepts of each concept from its own gloss.

3. The hypernym of a synset has also a gloss, thus a synset can be connected to the concepts from the gloss of its hypernym. Similarly to the IS-A relations, other WordNet lexico-semantic relations can be followed to reach a new synset and have access to the concepts used in its gloss. Such relations may include HAS-MEMBER, HAS-PART or ENTAILS and CAUSE-TO. Lexical relations based on morphological derivations, if available may be used too[3]. Morphological relations include the NOMINALIZATION relations, known to be useful in IE.

4. A synset can be used itself to define other concepts, therefore connections exist between each concept and all concepts it helps define.

Figure 8 illustrates the four possible sources of topical relations based on two of the WordNet relations, namely GLOSS and IS-A.

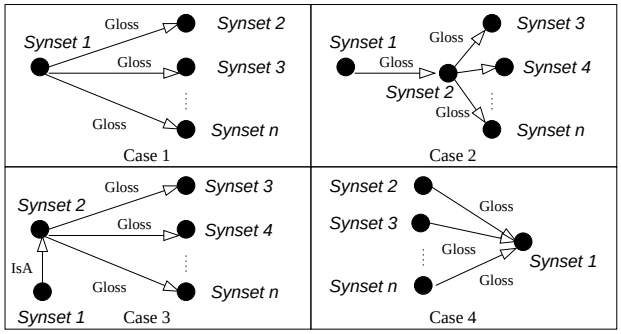

Figure 8: Four sources of topical relations.

Topical relations as Paths between WordNet Synsets

Two principles guide the uncovering of topical relations. First we believe that redundant connections rule out 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. Second, the shorter the paths, the stronger their validity. Consequently, we rule out paths of length larger than 4. This entails the fact that each topic may be represented by at least five synsets.

Figure 9 shows the topical relations produced by the paths originating at the WordNet synset {*mad cow disease*} and traversing concepts like {*mental illness*}, {*agitation*} or {*brain, mind*}. It is to be noted that each concept may be reached by at least two different paths of relations.

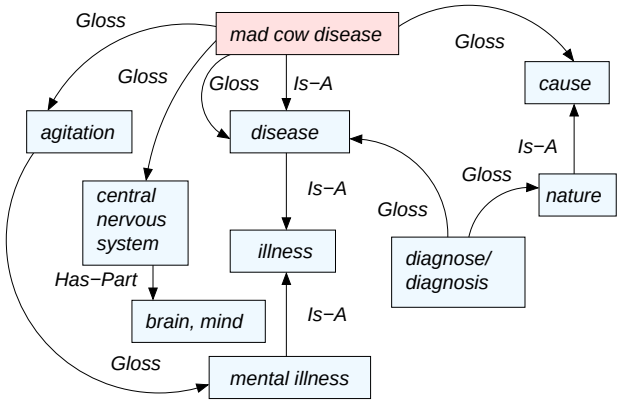

Figure 9: Topical relations for the "mad cow disease" topic.

[3]WordNet 2 already encodes derivational morphology.

4.2. Ad-hoc Templates

A template representation of a topic can be viewed as a list of semantic roles, each role being a slot that is filled by information extracted from text. The topical relations mined from WordNet have the advantage that they bring forward semantically-connected concepts deemed relevant to the topic. However these concepts cannot be mapped directly into a list of slots. First, WordNet was not devised with the IE application in mind - it is a general resource of English lexico-semantic knowledge. Because of this, some concepts relevant to a given topic may not be encoded in WordNet. Second, several WordNet concepts traversed by topical relations may be categorized under the same semantic role. Third, some semantic roles may be encoded in WordNet at a very abstract level, and thus they may never be reached by topical relations. Fourth, some of the semantic roles derived from topical relations may never be filled, since there is no corresponding information in the texts. To address all these issues, we have developed a corpus-based technique for creating ad-hoc lists of semantic roles for the template representation of the collection topic. Our algorithm for ad-hoc template generation was inspired by the empirical approach for conceptual case frame acquisition presented in (Riloff and Schmelzenbach 1998).

Algorithm Ad-hoc Template Generation

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 a seed lexical items used for the identification of the template slots.

Step 2: Identify all Subject-Verb-Object (SVO) +Prepositional attachments syntatic structures in which one of the topical concepts is used. For this purpose, we used the phrasal parser implemented in CICERO as well as all the syntactic variants of the SVO syntactic structures used to implement extraction patterns.

Step 3: Apply the IE coreference resolution module and consider all the syntactic SVO structures involving all coreferring expression of any of the nouns used in the syntactic structures discovered at Step 2.

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 (Riloff and Schmelzenbach 1998), **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 because of the semantic categories that are identified with high frequency, under the assumption that this reflect 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.

5. Evaluation

We participated with GISTEXTER in the DUC-2001 multi-document summarization involving 30 document sets. For each test data set the multi-document summary generated by our system was compared with a gold-standard summary created by humans. For each data set, the author of the gold-standard summary assessed the degree of matching between the model summary and the summaries generated by the systems evaluated in DUC-2001. Three qualitative measures were used to compare the systems as a whole. These measures are *grammaticality*, *cohesion* and *organization*. Each of these measures were scored on a scale between 0 and 4.

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 by assessors into *model units* (MUs), which are informational units that should express one self-contained fact in the ideal case. 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 are always sentences. Subsequently, the assessor located the PU(s) that covered the content of each MU, if any, and assigned an estimate of the degree of matching, between 1 and 4.

From the assessor judgments, an extensive analysis was reported in (McKeown et al.2001). In this paper we report only on the results of GISTEXTER. For evaluating the content, we consider only the *Precision* and *Recall* measures. Precision is calculated as the number of PUs matching some MU divided by the number of PUs in the peer summary, considering all summaries automatically generated for the same collection. Our system scored a precision of 50.76%. As reported in (McKeown et al.2001), this estimate of the precision is conservative, since the number of PUs that are considered correct can be increased by considering information about the PUs not assigned to MUs. However, this information is not available, since the data on PUs not assigned to MUs is qualitative in nature (e.g. "some", "most") rather than a count. In (McKeown et al.2001) it is proposed to use a more accurate measure by using weights reflecting the degree of matching between PUs and MUs. Unfortunately, this data was not available after the evaluation. However an accurate analysis of the recall was possible. The recall is defined as the number of MUs matched on or above a threshold t divided by the total number of MUs in the gold-standard summary. A very strict recall measure is granted when $t = 4$ whereas a lenient recall measure considers even MUs with "little" content covered for $t = 1$.

We present here also a method proposed in (McKeown et al.2001) for combining the four recall measures corresponding to the four possible values of t. Instead of treating the degree of matching as an ordinal value, it was proposed

t	Recall	Ranking between all systems
1	35.53	3
2	28.82	3
3	15.03	1
4	7.42	1
Average	88.41	1

Table 1: Recall of summary content obtained by GISTEXTER in DUC-2001

to consider it a ratio, i.e. assume that a value of matching of 2 is twice as good as a matching of 1 and half as good as a matching of 4. Under this assumption, the degree of matching over MUs can be averaged and the recall is considered as an average of the degree of matching. Table 1 shows the values of the recall obtained by GISTEXTER depending on the threshold t as well as the average recall. The Table also shows that for recall GISTEXTER was scored as the best system among all 12 participating multi-document summarization systems.

To evaluate the style of the summary, the assessments were based on the *grammaticality*, *organization* and *coherence/cohesion* of the output of GISTEXTER. Table 2 shows the scores obtained by our system as well as the relative ranking between the automated multi-document summarization systems that participated in DUC-2001.

Criterion	Value	Ranking between all systems
Grammaticality	3.5086	8
Cohesion	2.3362	1
Organization	2.6121	1

Table 2: The evaluation of the summary style for GISTEXTER in DUC-2001

6. Conclusions

In this paper we have shown that multi-document summarization of good quality can be obtained if extraction templates populated by a high performance IE systems are available. We have presented an IE-based multi-document summarization procedure that incrementally adds information to create summaries of variable size. The decision of using incremental additions of sentences from multiple documents based on their mapping from the template slots produced very good results for coherence and organization in the DUC-2001 evaluations.

Additionally, in this paper we have presented a method for creating ad-hoc templates for new topics that made possible the good recall results we obtained with GISTEXTER in DUC-2001. The scores on the style of the summary showed that the grammaticality of the summaries produced by our system were not very good. We are currently working on a generation module that would realize the structure of the populated templates.

7. References

R. Barzilay, N. Elhedad and K. McKeown. Sentence ordering in multidocument summarization. In *Proceedings of Conference on Human Language Technology (HLT-2001)*, 2001.

G. DeJong. An overview of the FRUMP system. In *Stategies for natural language processing*, W. Lehnert and M. Ringle Eds., pages 149-176, Lawrence Erlbaum Associates, 1982.

M. Elhadad. Using argumentation to control lexical choice: A unifi cation-based implementation. PhD Thesis, Computer Science Department, Columbia University, 1993.

S. Harabagiu and S. Maiorano. Acquisition of Linguistic Patterns for Knowledge-Based Information Extraction. In *Proceedings of Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000.

D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, M. Stickel and M. Tyson. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In *Finite State Language Processing*, edited by Emmanuel Roche and Yves Schabes, MIT Press, 1997.

C.Y. Lin and E. Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrcken, Germany, July 31- August 4, 2000.

I. Mani and M. Maybury, eds. Advances in Automatic Text Summarization. MIT Press, 1999.

K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman and S. Teufel. Columbia Multi-Document Summarization: Approach and Evaluation. In *Workshop Notes for the DUC-2001 Summarization*, pages 43–64, September 2001.

George A. Miller. WordNet: a lexical database for English. In *Communications of the ACM*, Vol.38, No.11:39–41, 1995.

Fernando Pereira and Rebecca Wright. Finite-state approximation of phrase-structure grammars. In *Finite State Language Processing*, edited by Emmanuel Roche and Yves Schabes, MIT Press, 149–179, 1997.

D. Radev and K. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469-500, September 1998.

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

E. Riloff and R. Jones. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, pp. 474-479, 1999.

M. Surdeanu and S. Harabagiu. Infrastructure for Open-Domain Information Extraction. In *Proceedings of Conference on Human Language Technology (HLT-2002)*, 2002.

R. Yangarber, R. Grishman, P. Tapanainen and S. Huttunen Automatic Acquisition of Domain Knowledge for Information Extraction. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbrcken, Germany, 2000.