Applying Natural Language Generation to Indicative Summarization

Min-Yen Kan and **Kathleen R. McKeown**
Department of Computer Science
Columbia University
New York, NY 10027, USA
{min,kathy}@cs.columbia.edu

Judith L. Klavans
Columbia University
Center for Research on Information Access
New York, NY, 10027
klavans@cs.columbia.edu

Abstract

The task of creating indicative summaries that help a searcher decide whether to read a particular document is a difficult task. This paper examines the indicative summarization task from a generation perspective, by first analyzing its required content via published guidelines and corpus analysis. We show how these summaries can be factored into a set of document features, and how an implemented content planner uses the topicality document feature to create indicative multidocument query-based summaries.

1 Introduction

Automatic summarization techniques have mostly neglected the *indicative* summary, which characterizes what the documents are about. This is in contrast to the informative summary, which serves as a surrogate for the document. Indicative multidocument summaries are an important way of helping a user discriminate between several documents returned by a search engine.

Traditional summarization systems are primarily based on text extraction techniques. For an indicative summary, which typically describes the topics and structural features of the summarized documents, these approaches can produce summaries that are too specific. In this paper, we propose a natural language generation (NLG) model for the automatic creation of indicative multidocument summaries. Our model is based on the values of high-level document features, such as its distribution of topics and media types.

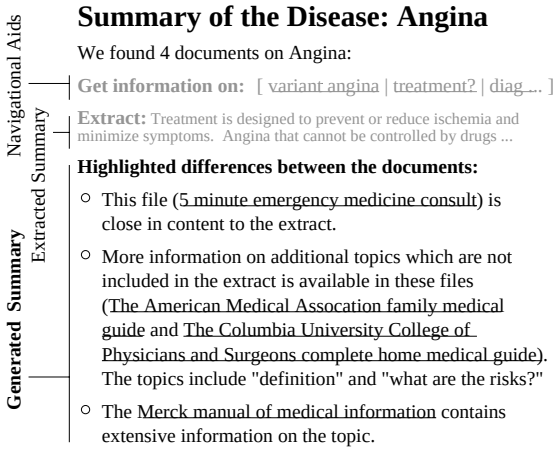

Figure 1: A CENTRIFUSER summary on the healthcare topic of "Angina". The generated indicative summary in the bottom half categorizes documents by their difference in topic distribution.

Specifically, we focus on the problem of content planning in indicative multidocument summary generation. We address the problem of "what to say" in Section 2, by examining what document features are important for indicative summaries, starting from a single document context and generalizing to a multidocument, query-based context. This yields two rules-of-thumb for guiding content calculation: 1) reporting differences from the norm and 2) reporting information relevent to the query.

We have implemented these rules as part of the content planning module of our CENTRIFUSER summarization system. The summarizer's architecture follows the consensus NLG architecture (Reiter, 1994), including the stages of content calculation and content planning. We follow the generation of a sample indicative multidocument query-based summary, shown in the bottom half

of Figure 1, focusing on these two stages in the remainder of the paper.

2 Document features as potential summary content

Information about topics and structure of the document may be based on higher-level document features. Such information typically does not occur as strings in the document text. Our approach, therefore, is to identify and extract the document features that are relevant for indicative summaries. These features form the potential content for the generated summary and can be represented at a semantic level in much the same way as input to a typical language generator is represented. In this section, we discuss the analysis we did to identify features of individual and sets of multiple documents that are relevant to indicative summaries and show how feature selection is influenced by the user query.

2.1 Features of individual documents

Document features can be divided into two simple categories: a) those which can be calculated from the document body (e.g. topical structure (Hearst, 1993) or readability using Flesch-Kincaid or SMOG (McLaughlin, 1969) scores), and b) "metadata" features that may not be contained in the source article at all (e.g. author name, media format, or intended audience). To decide which of these document features are important for indicative summarization, we examined the problem from two points of view. From a top-down perspective, we examined prescriptive guidelines for summarization and indexing. We analyzed a corpus of indicative summaries for the alternative bottom-up perspective.

Prescriptive Guidelines. Book catalogues index a number of different document features in order to provide enhanced search access. The United States MARC format (2000), provides index codes for document-derived features, such as for a document's table of contents. It provides a larger amount of index codes for metadata document features such as fields for unusual format, size, and special media. ANSI's standard on descriptions for book jackets (1979) asks that publishers mention unusual formats, binding styles, or whether a book targets a specific audience.

Descriptive Analysis. Naturally indicative summaries can also be found in library catalogs, since the goal is to help the user find what they need. We extracted a corpus of single document summaries of publications in the domain of consumer healthcare, from a local library. The corpus contained 82 summaries, averaging a short 2.4 sentences per summary. We manually identified several document features used in the summaries and characterized their percentage appearance in the corpus, presented in Table 1.

Document Feature	% appearance in corpus
Document-derived features	
Topicality (e.g. "Topics include symptoms, ...")	100%
Content Types (e.g. "figures and tables")	37%
Internal Structure (e.g. "is organized into three parts")	17%
Readability (e.g. "in plain English")	18%
Special Content (e.g. "Offers 12 credit hours")	7%
Conclusions	3%
Metadata features	
Title	32%
Revised/Edition	28%
Author/Editor	21%
Purpose	18%
Audience	17%
Background/Lead	11%
Source (e.g. "based on a report")	8%
Media Type (e.g. "Spans 2 CDROMs")	5%

Table 1: Distribution of document features in library catalog summaries of consumer healthcare publications.

Our study reports results for a specific domain, but we feel that some general conclusions can be drawn. Document-derived features are most important (i.e., most frequently occuring) in these single document summaries, with direct assessment of the topics being the most salient. Metadata features such as the intended audience, and the publication information (e.g. edition) information are also often provided (91% of summaries have at least one metadata feature when

they are independently distributed).

2.2 Generalizing to multiple documents

We could not find a corpus of indicative multidocument summaries to analyze, so we only examine prescriptive guidelines for multidocument summarization.

The Open Directory Project's (an open source Yahoo!-like directory) editor's guidelines (2000) states that category pages that list many different websites should "make clear what makes a site different from the rest". "the rest" here can mean several things, such as "rest of the documents in the set to be summarized" or "the rest of the documents in the collection". We render this as the following rule-of-thumb 1:

1. for a multidocument summary, a content planner should report differences in the document that deviate from the norm for the document's type.

This suggests that the content planner has an idea of what values of a document feature are considered normal. Values that are significantly different from the norm could be evidence for a user to select or avoid the document; hence, they should be reported. For example, consider the document-derived feature, *length*: if a document in the set to be summarized is of significantly short length, this fact should be brought to the user's attention.

We determine a document feature's norm value(s) based on all similar documents in the corpus collection. For example, if all the documents in the summary set are shorter than normal, this is also a fact that may be significant to report to the user. The norms need to be calculated from only documents of similar type (i.e. documents of the same domain and genre) so that we can model different value thresholds for different kinds of documents. In this way, we can discriminate between "long" for consumer healthcare articles (over 10 pages) versus "long" for mystery novels (over 800 pages).

2.3 Generalizing to interactive queries

If we want to augment a search engine's ranked list with an indicative multidocument summary, we must also handle queries. The search engine ranked list does this often by highlighting query terms and/or by providing the context around a query term. Generalizing this behavior to handling multiple documents, we arrive at rule-of-thumb 2.

2. for a query-based summary, a content planner should highlight differences that are relevant to the query.

This suggests that the query can be used to prioritize which differences are salient enough to report to the user. The query may be relevant only to a portion of a document; differences outside of that portion are not relevant. This mostly affects document-derived document features, such as topicality. For example, in the consumer healthcare domain, a summary in response to a query on treatments of a particular disease may not want to highlight differences in the documents if they occur in the symptoms section.

3 Introduction to CENTRIFUSER

CENTRIFUSER is the indicative multi-document summarization system that we have developed to operate on domain- and genre-specific documents. We are currently studying consumer healthcare articles using it. The system produces a summary of multiple documents based on a query, producing both an extract of similar sentences (see Hatzivassiliglou et al. (2001)) as well as generating text to represent differences. We focus here only on the content planning engine for the indicative, difference reporting portion. Figure 2 shows the architecture of the system.

We designed CENTRIFUSER's input based on the requirements from our analysis; document features are extracted from the input texts and serve as the potential content for the generated summary. CENTRIFUSER uses a plan to select summary content, which was developed based on our analysis and the resulting previous rules.

Our current work focuses on the document feature which most influences summary content and form, *topicality*. It is also the most significant and useful document feature. We have found that discussion of topics is the most important part of the indicative summary. Thus, the text plan is built around the topicality document feature and other features are embedded as needed. Our discussion

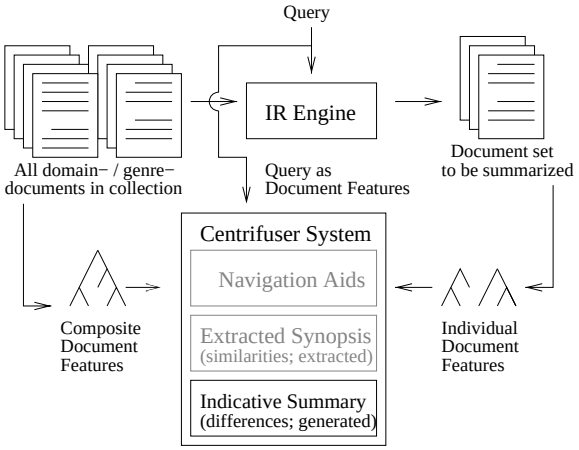

Figure 2: CENTRIFUSER architecture.

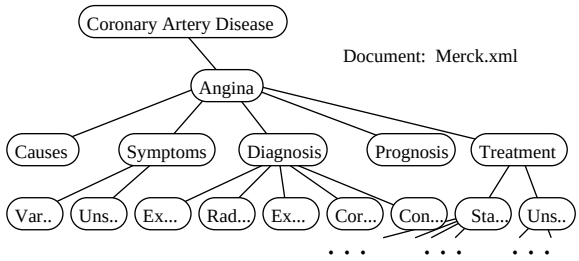

Figure 3: A topic tree for an article about coronary artery disease from *The Merck manual of medical information*, constructed automatically from its section headers.

now focuses on how the topicality document feature is used in the system.

In the next sections we detail the three stages that CENTRIFUSER follows to generate the summary: content calculation, planning and realization. In the first, potential summary content is computed by determining input topics present in the document set. For each topic, the system assesses its relevance to the query and its prototypicality given knowledge about the topics covered in the domain. More specifically, each document is converted to a tree of topics and each of the topics is assigned a topic type according to its relationship to the query and to its normative value. In the second stage, our content planner uses a text plan to select information for inclusion in the summary. In this stage, CENTRIFUSER determines which of seven document types each document belongs to, based on the relevance of its topics to the query and their prototypicality. The plan generates a separate description for the documents in each document type, as in the sample summary in Figure 1, where three document categories was instantiated. In the final stage, the resulting description is lexicalized to produce the summary.

4 Computing potential content: topicality as topic trees

In CENTRIFUSER, the topicality document feature for individual documents is represented by a tree data structure. Figure 3 gives an example *document topic tree* for a single consumer healthcare article. Each document in the collection is represented by such a tree, which breaks each document's topic into subtopics.

We build these document topic trees automatically for structured documents using a simple approach that utilizes section headers, which suffices for our current domain and genre. Other methods such as layout identification (Hu et al., 1999) and text segmentation / rhetorical parsing (Yaari, 1999; Kan et al., 1998; Marcu, 1997) can serve as the basis for constructing such trees in both structured and unstructured documents, respectively.

4.1 Normative topicality as composite topic trees

As stated in rule 1, the summarizer needs normative values calculated for each document feature to properly compute differences between documents.

The *composite topic tree* embodies this paradigm. It is a data structure that compiles knowledge about all possible topics and their structure in articles of the same intersection of domain and genre, (i.e., rule 1's notion of "document type"). Figure 4 shows a partial view of such a tree constructed for consumer healthcare articles.

The composite topic tree carries topic information for all articles of a particular domain and genre combination. It encodes each topic's relative typicality, its prototypical position within an article, as well as variant lexical forms that it may be expressed as (e.g. alternate headers). For instance, in the composite topic tree in Figure 4, the topic "Symptoms" is very typical (.95 out of 1),

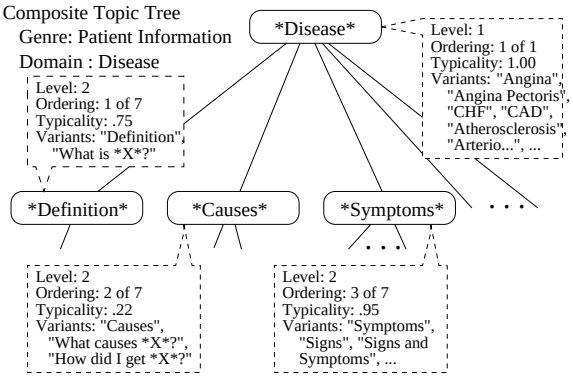

Figure 4: A sample composite topic tree for consumer health information for diseases.

may be expressed as the variant "Signs" and usually comes after other its sibling topics ("Definition" and "Cause").

Compiling composite topic trees from sample documents is a non-trivial task which can be done automatically given document topic trees. Within our project, we developed techniques that align multiple document topic trees using similarity metrics, and then merge the similar topics (Kan et al., 2001), resulting in a composite topic tree.

5 Content Planning

NLG systems traditionally have three components: content planning, sentence planning and linguistic realization. We will examine how the system generates the summary shown earlier in Figure 1 by stepping through each of these three steps.

During content planning, the system decides what information to convey based on the calculated information from the previous stage. Within the context of indicative multidocument summarization, it is important to show the differences between the documents (rule 1) and their relationship to the query (rule 2). One way to do so is to classify documents according to their topics' prototypicality and relevance to the query. Figure 5 gives the different document categories we use to capture these notions and the order in which information about a category should be presented in a summary.

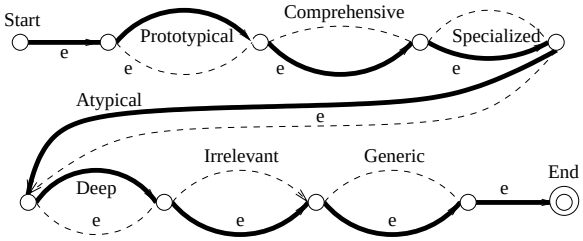

Figure 5: Indicative summary content plan, solid edges indicate moves in the sample summary.

5.1 Document categories

Each of the document categories in the content plan in Figure 5 describes documents that are similar in their distribution of information with respect to the topical norm (rule 1) and to the query (rule 2). We explain these *document categories* found in the text plan below. The examples in the list below pertain to a general query of "Angina" (a heart disorder) in the same domain of consumer healthcare.

1. Prototypical - contains information that one would typically expect to find in an on-topic document of the domain and genre. An example would be a reference work, such as *The AMA Guide to Angina*.

2. Comprehensive - covers most of the typical content but may also contain other added topics. An example could be a chapter of a medical text on angina.

3. Specialized - are more narrow in scope than the previous two categories, treating only a few normal topics relevant to the query. A specialized example might be a drug therapy guide for angina.

4. Atypical - contains high amounts of rare topics, such as documents that relate to other genres or domains, or which discuss special topics. If the topic "Prognosis" is rare, then a document about life expectancy of angina patients would be an example.

5. Deep - are often barely connected with the query topic but have much underlying information about a particular subtopic of the query. An example is a document on "Surgical treatments of Angina".

6. Irrelevant - contains mostly information not relevant to the query. The document may be very broad, covering mostly unrelated materials. A

document about all cardiovascular diseases may be considered irrelevant.

7. Generic - don't display tendencies towards any particular distribution of information.

5.2 Topic types

Each of these document categories is different because they have an underlying difference in their distribution of information. CENTRIFUSER achieves this classification by examining the distribution of *topic types* within a document. CENTRIFUSER types each individual topic in the individual document topic trees as one of four possibilities: *typical*, *rare*, *irrelevant* and *intricate*. Assigning topic types to each topic is done by operationalizing our two content planning rules.

To apply rule 2, we map the text query to the single most similar topic in each document topic tree (currently done by string similarity between the query text and the topic's possible lexical forms). This single topic node – the query node – establishes a relevant scope of topics. The relevant scope defines three regions in the individual topic tree, shown in Figure 6: topics that are relevant to the query, ones that are too *intricate*, and ones that are *irrelevant* with respect to the query. Irrelevant topics are not subordinate to the query node, representing topics that are too broad or beyond the scope of the query. Intricate topics are too detailed; they are topics beyond k hops down from the query node.

Each individual document's ratio of topics in these three regions thus defines its relationship to the query: a document with mostly information on treatment would have a high ratio of relevant to other topics if given a treatment query; but the same document given a query on symptoms would have a much lower ratio.

To apply rule 1, we need to know whether a particular topic "deviates from the norm" or not. We interpret this as whether or not the topic normally occurs in similar documents – exactly the information encoded in the composite topic tree's typicality score. As each topic in the document topic trees is an instance of a node in the composite topic tree, each topic can inherit its composite node's typicality score. We assign nodes in the relevant region (as defined by rule 2), with labels based on their typicality. For convenience, we set

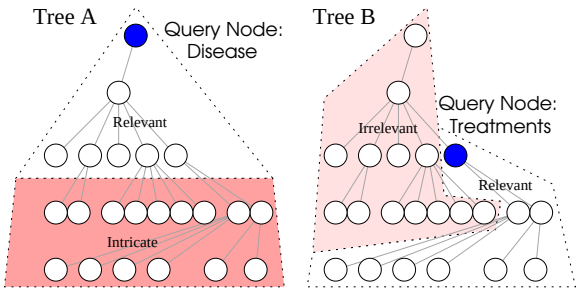

Figure 6: The three topic regions as defined by the query, for k = 2 (k being the *intricate* beam depth.

a typicality threshold α, above which a topic is considered *typical* and below which we consider it *rare*.

At this point each topic in a document is labeled as one of the four topic types. The distribution of these four types determines each document's document category. Table 2 gives the distribution parameters which allow CENTRIFUSER to classify the documents.

Document Category	Topic Distribution
1. Prototypical	$>$ 50+% *typical* and $>$ 50+% all possible *typical*
2. Comprehensive	$>$ 50+% all possible *typical*
3. Specialized	$>$ 50+% *typical*
4. Atypical	$>$ 50+% *rare*
5. Deep	$>$ 50+% *intricate*
6. Irrelevant	$>$ 50+% *irrelevant*
7. Generic	n/a

Table 2: Classification rules for document categories.

Document categories add a layer of abstraction over the topic types that allow us to reason about documents. These document labels still obey our content planning rules 1 and 2: since the assignment of a document category to a document is conditional on its distribution of its topics among the topic types, a document's category may shift if the query or its norm is changed.

In CENTRIFUSER, the text planning phase is implicitly performed by the classification of the summary document set into the document categories. If a document category has at least one document attributed to it, it has content to be conveyed. If the document category does not have any documents attributed to it, there is no information to convey to the user concerning the par-

ticular category.

An instantiated document category conveys a couple of messages. A description of the document type as well as the elements attributed to it constitutes the minimal amount of information to convey. Optional information such as details on the instances, sample topics or other unusual document features, can be expressed as well.

$$\begin{bmatrix} relation : setElements \\ args : \begin{bmatrix} docCat : atypical \\ element : AMAguide \\ element : CUguide \end{bmatrix} \end{bmatrix}$$
$$\begin{bmatrix} relation : description \\ args : [\ docCat : atypical\ ] \end{bmatrix}$$
$$\begin{bmatrix} relation : hasTopics \\ args : \begin{bmatrix} docCat : atypical \\ topic : definition \\ topic : whataretherisks? \end{bmatrix} \end{bmatrix}$$

Figure 7: Messages instantiated for the *atypical* document category for the summary in Figure 1.

The text planner must also order the selected messages into a coherent plan for subsequent realization. For our summary, this is a problem on two levels: deciding the ordering between the document category descriptions and deciding the ordering of the individual messages within the document category. In CENTRIFUSER, the discourse plans for both of these levels are fixed. Let us first discuss the inter-category plan.

Inter-category. We order the document category descriptions based on the ordering expressed in Table 2. The reason for this order is partially reflected by the category's relevance to the user query (rule 2). Document categories like *prototypical* whose salient feature is their high ratio of relevant topics, are considered more important than document categories that are defined by their ratio of intricate or irrelevant topics (e.g. *deep*).

This precendence rule decides the ordering for the last few document types (deep $\rightarrow$ irrelevant $\rightarrow$ generic). For the remaining document types, defined by their high ratio of typical and rare topics, we use an additional constraint of ordering document types that are closer to the article type norm before others. This orders the remaining beginning topics (prototypical $\rightarrow$ comprehensive $\rightarrow$ specialized $\rightarrow$ atypical). The reason for this is that CENTRIFUSER, along with reporting salient differences by using NLG, also reports an multi-document extract based on similarities. As similarities are drawn mostly from common topics – that is, *typical* ones – typical topics are regarded as more important than rare ones.

Figure 5 shows the resulting inter-category discourse plan. As stated in the text planning phase, if no documents are associated with a particular document category, it will be skipped, reflected in the figure by the ϵ moves. Our sample summary summary contains prototypical (first bullet), atypical (second) and deep (third) document categories, and as such activates the solid edges in the figure.

Intra-category. Ordering the messages within a category follows a simple rule. Obligatory information is expressed first, while optional information is expressed afterwards. Thus the document category's constituents and its description always come first, and information about sample topics or other unusual document features come afterwards, shown in Figure 8. The result is a partial ordering (as the order of the messages in the obligatory information has not been fixed) that is linearized later.

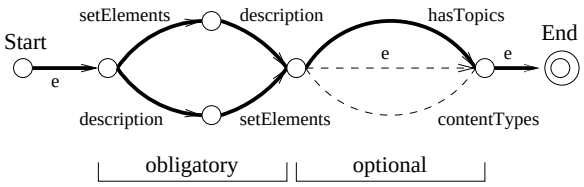

Figure 8: Intra-category discourse plan, solid edges indicate moves in the atypical document category. The final choice on which obligatory structure to use is decided later during realization.

6 Sentence Planning and Lexical Choice

In the final step, the discourse plan is realized as text. First, the sentence planner groups messages into sentences and generates referring expressions for entities. Lexical choice also happens at this stage. In our generation task, the grouping task is minimal; the separate categories are semantically distinct and need to be realized separately (e.g., in the sample, each category is a separate list item). The obligatory information of the description of the category as well as the members of the category are combined into a single sentence, and op-

tional information (if realized) constitute another sentence.

6.1 Generating Referring Expressions

One concern for generating referring expressions is constraining the size of the sentence. This is an issue when constructing referring expressions to sets of documents matching a document type. For example, if a particular document category has more than five documents, listing the names of each individual document is not felicitous. In these cases, an exemplar file is picked and used to demonstrate the document type. Resulting text is often of the form: "There are 23 documents (such as the *AMA Guide to Angina*) that have detailed information on a particular subtopic of angina."

Another concern in the generation of referring expressions is when the optional information only applies to a subset of the documents of the category. In these cases, the generator will reorder the elements of the document category in such a way to make the subsequent referring expression more compact (e.g. "The first five documents contain figures and tables as well" versus the more voluminous "The first, third, fifth and the seventh documents contain figures and tables as well").

```
(S1/description+setElements
  (V1 :value ''be available'')
  (NP1/atypical :value
     ''more information on additional
       topics which are not included
       in the extract'')
  (NP2/setElements :value
     ''files (The AMA guide and
       CU Guide)''))
(S2/hasTopics
  (V1 :value ''include'')
  (NP1/atypicalTopics :value ''topics'')
  (NP2/topicList :value
     ''definition and
       what are the risks?''))
```

Figure 9: Sentence plan for the atypical document category.

6.2 Lexical Choice

Lexical choice in CENTRIFUSER is performed at the phrase level; entire phrases can be chosen all at once, akin to template based generation. Currently, a path is randomly chosen to select a lexicalization. In the sample summary, the atypical document category's (i.e. the second bullet item) description of "more information on additional topics ..." was chosen as the description message among other phrasal alternatives. The sentence plan for this description is shown in Figure 9.

For certain document categories, a good description can involve information outside of the generated portion of the summary. For instance, Figure 1's prototypical document category could be described as being "an reference document about angina". But as a prototypical document shares common topics among other documents, it is actually well represented by an extract composed of the similarities across document sets. Similarity extraction is done in another module of CENTRIFUSER (the greyed out portion in the figure), and as such we also can use a phrasal description that directly references its results (e.g., in the actual description used for the prototypical document category in Figure 1).

6.3 Linguistic Realization

Linguistic realization takes the sentence plan and produces actual text by solving the remaining morphology and syntactic problems. CENTRIFUSER currently chooses a valid syntactic pattern at random, in the same manner as lexical choice. Morphological and other agreement constraints are minor enough in our framework and are handled by set rules.

7 Current status and future work

CENTRIFUSER is fully implemented; it produces the sample summary in Figure 1. We have concentrated on implementing the most commonly occuring document feature, topicality, and have additionally incorporated three other document features into our framework (document-derived *Content Types* and *Special Content* and the *Title* metadata).

Future work will include extending our document feature analysis to model context (to model adding features only when appropriate), as well as incorporating additional document features. We are also exploring the use of stochastic corpus modeling (Langkilde, 2000; Bangalore and Rambow, 2000) to replace our template-based realizer with a probabilistic one that can produce felici-

tous sentence patterns based on contextual analysis.

8 Conclusion

We have presented a model for indicative multidocument summarization based on natural language generation. In our model, summary content is based on document features describing topic and structure instead of extracted text. Given these features, a generation model uses a text plan, derived from analysis of naturally occurring indicative summaries plus guidelines for summarization, to guide the system in describing document topics as typical, rare, intricate, or relevant to the user query. We showed how the topicality document feature can be derived from the set of input documents and represented as a topic tree for each document along with a merged composite topic for all documents in the collection against which prototypicality and query relevance can be computed. Our ongoing work is examining how to automatically learn the text plans along with the tactics needed to realize each piece of the instantiated plan as a sentence.

References

ANSI. 1979. American national standard for describing books in advertisements, catalogs, promotional materials and book jackets. New York, USA. ANSI Z39.13-1979.

Srinivas Bangalore and Owen Rambow. 2000. Exploiting a probabilistic hierarchical model for generation. In *Proc. of the 18th Intl. Conf. on Computational Linguistics (COLING 2000), Saarbrucken, Germany*.

Vasileios Hatzivassiliglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown. 2001. Simfinder: A flexible custering tools for summarization. In *Human Language Technologies 2001*.

Marti Hearst. 1993. Text tiling: A quantitative approach to discourse segmentation. Technical report, University of California, Berkeley, Sequoia.

Jianying Hu, Ramanujan Kashi, and Gordon Wilfong. 1999. Document image layout comparison and classification. In *Proc. of the Intl. Conf. on Document Analysis and Recognition (ICDAR)*.

Min-Yen Kan, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear segmentation and segment relevence. In *WVLC6*, pages 197–205, Montréal, Québec, Canada, August. ACL.

Min-Yen Kan, Judith L. Klavans, and Kathleen R. McKeown. 2001. Synthesizing composite topic structure trees for multiple domain specific documents. Technical Report CUCS-003-01, Columbia University.

Irene Langkilde. 2000. Forest-based statistical sentence generation. In *6th Applied Natural Language Processing Conference (ANLP'2000)*, pages 170–177, Seattle, Washington, USA.

Library of Congress. 2000. Marc 21 format for classification data : including guidelines for content designation. Washington, D.C., USA. ISN 0660179903.

Daniel Marcu. 1997. The rhetorical parsing of natural language texts. In *Proceedings of 35th ACL and 8th EACL*, pages 96–103, Madrid, Spain.

Harry McLaughlin. 1969. SMOG grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

ODP. 2000. Open Directory Project guidelines. http://dmoz.org/guidelines.html, November.

Ehud Reiter. 1994. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible? In *Proc of the Seventh International Workshop on Natural Language Generation (INLGW-1994)*, pages 163–170, Kennebunkport, Maine, USA.

Yaakov Yaari. 1999. *The Texplorer*. Ph.D. thesis, Bar Ilan University, Israel, April.