Next-Generation Summarization: Contrastive, Focused, and Update Summaries

René Witte
Fakultät für Informatik
Universität Karlsruhe (TH)
Karlsruhe, Germany
witte@ipd.uka.de

Sabine Bergler
Department of Computer Science
Concordia University
Montréal, Canada
bergler@cs.concordia.ca

Abstract

Classical multi-document summaries focus on the common topics of a document set and omit distinctive themes particular to a single document—thereby often suppressing precisely that kind of information a user might need for a specific task. This can be avoided through advanced multi-document summaries that take a user's context and history into account, by delivering focused, contrastive, or update summaries. To facilitate the generation of these different summaries, we propose to generate all types from a single data structure, *topic clusters*, which provide for an abstract representation of a set of documents. Evaluations carried out on five years' worth of data from the DUC summarization competition prove the feasibility of this approach.

1 Introduction

As a much-noticed study attested, *information overload harms concentration more than marijuana.*[1] Internet search engines continue to deliver more and more information to users, when in fact they would rather have less [7]. One approach for mitigating information overload is to *compress* the information delivered from information retrieval (IR) engines through automatic summarization: Instead of displaying a list of relevant documents with keyword-specific highlights, a system can deliver a multi-document summary containing the most important information.

In recent years, extensive experiments on multi-document summarization has been carried out within the Document Understanding Conference (DUC) competition[2] sponsored by the U.S. NIST. In DUC, system developers participate in experiments based on common tasks and data, which allows a comparison of different approaches using various evaluation metrics.

In general, the purpose of a multi-summary is not to serve as a replacement for the real texts, rather, they aim to help a reader to find relevant *topics*. In combination with short (keyword-style) summaries for an individual document, a human reader should be able to quickly determine: *(a)* whether the set itself contains information on a relevant topic and *(b)* which of the individual text(s) should be read for an in-depth understanding of the topic.

However, this approach is not the most efficient one when more information is available concerning a user's *context* and *history*: Did he read some of the documents in the set before? Then he might only be interested in *updates*, in new information. Is he working on a specific task? Then he primarily needs information pertaining to the task at hand, not a general summary. These scenarios have been addressed in DUC with the introduction of focused and update summaries. In addition, within this paper we propose a third kind, *contrastive* summaries. These are designed for a differential analysis of a document set, showing first the *commonalities* of all texts and additionally the topics that are *unique* to each individual document.

In practice, a user (a.k.a. "knowledge worker") might need all of these (and other) kinds of summaries while performing knowledge-intensive tasks, ideally embedded within a dynamic, semantic desktop environment that allows for changing the displayed content on-the-fly. Here, the concerns of language system engineers become important due to the growing number of required features. Developing, testing, and deploying individual summarization systems for each of these kinds of summary is not feasible. Thus, we propose a different approach: the generation of an abstracting data structure we call *Topic Clusters*, from which all of these summaries (and some additional) can immediately be generated.

Our research is significant for several reasons: (1) We revive the almost abandoned field of contrastive summarization with a contemporary application focus and a simple, practical approach for generating them; (2) The fact that we investigate automatic summarization for actual deployment within a user's semantic desktop, deriving requirements that go beyond purely NLP issues by addressing software engineering concerns; and (3) We abstract from the generation of a single type of multi-document summary to arrive at a general data structure that can be used for computing *all* of them.

2 Summarization Tasks

As described above, a single kind of summary is not sufficient to adequately cover the information needs of a user performing a particular task. In this section, we motivate and define summaries that go beyond the classical, generic multi-document type.

2.1 Contrastive Summaries

Consider a user performing an analysis of a document set, e.g., on the top 50 list of hits delivered by an IR engine for a specific query. To avoid reading all of them, he instructs his semantic desktop to produce a multi-document summary of the whole set, as well as

[1] "Info-overload harms concentration more than marijuana." *New Scientist*, April 30, 2005, p. 6. http://www.newscientist.com/channel/being-human/mg18624973.400.

[2] DUC, http://duc.nist.gov

short (ten words, keyword-style) single-document summaries of each text. If he is only interested in the most important topics of the set, this combination will help to detect those, as well as provide cues regarding a good candidate document to read in full. However, if an analysis requires finding the *differences* across the documents in a set, this technique will not work: Since both the multi-document summaries and the individual summaries have to focus on the most important and ubiquitous elements of the texts within the given space constraints, all *distinguishing* information is usually suppressed.

For example, one document set from the DUC 2004 competition contains texts regarding *Hurricane Mitch*. A topic-detecting summary generation algorithm would therefore generate or extract sentences about this natural disaster. Likewise, creating a very short per-document summary results in a similar task: find the most important topic(s) of each text. For the document cluster on Hurricane Mitch, the keywords `Hurricane` or `Central America` are extracted for every text, thus suppressing its distinguishing sub-topics (e.g., concerning EU relief efforts, military rescue operations, or the pope's appeal for aid).

The idea of homing in first on a cluster of multiple documents by their common topic, and then on the particular document of greatest interest within this cluster based on its distinctive topics leads us to propose *contrastive summaries*. We define a contrastive summary of multiple documents as a summary that indicates the common topics of all the articles as well as unique topics of each contained article.

While this idea is not entirely new, none of the current systems makes use of contrastive information. As Mani points out [5], "while similarity across documents is relatively well-understood, differences are not." We believe this is partly due to the lack of a suitable algorithm that can be easily implemented and works robustly even on large document sets (DUC requires summarization of 25–50 documents/set).

2.2 Focused Summaries

So far, we addressed the summarization of documents without additional, user-specific information. But in real life, nobody really wants to spent hours on *Google* searching for potentially relevant information. What users need is useful information pertaining to their task at hand, like writing a report, an email, or a research paper. Shouldn't a system be able to sense a user's current *context*, search for relevant information by itself, and present a summary thereof? Coupled with current information retrieval techniques and intelligent information system architectures [10], a new generation of language-aware information systems could proactively deliver the information users need, instead of requiring them to spend their limited time searching for them.

This leads to the idea of a *focused* summary, which only contains information relevant to the user's current context. This kind of summary essentially ignores information that does not contribute to the user's current task—a very useful property when trying to reduce the information overload.

Within the DUC competition, the context is modeled as a set of open-ended questions.[3] Being able to generate focused summaries has important practical applications for next-generation semantic desktop environments.

2.3 Update Summaries

The last kind of next-generation summary we address in this paper are *updates*. Here, the assumption is that a user has already read a number of documents on a certain topic and is only interested in new information that has not been covered before. A typical application scenario are newswire analysts that have to deal with multiple instances of the same or similar stories, as it is evolving over time.

Note that this kind of summary can be combined with both generic, focused, and contrastive summaries. In fact, the DUC 2007 competition defined the update task as a combination of generic updates with a context question, i.e., *focused update summaries*.

3 Topic Clusters

To generate contrastive, focused, and update summaries, we introduce a generic data structure that abstracts from individual tokens in a document collection: *topic clusters*. In the next subsection, we motivate this idea, followed by brief description of our approach for topic cluster generation.

3.1 Requirements

The target of our research is the individual user facing information overload caused by modern Internet/Intranet (IR) search engines: Rather than displaying a large list of documents with only keyword excerpts, we propose to condense the information contained in the result set through automatic summarization. A user should be able to switch between different kinds of summaries in a dynamic fashion, depending on his current work context and tasks.

From these observations, we can derive three main requirements for a data structure for summary generation:

Requirement #1: Domain-Independence. *The algorithm should work independently of an application domain.*

This follows directly from the intended application within a semantic desktop, where the summarization component acts a user's agent when interpreting results from Internet/Intranet searches.

Requirement #2: Flexibility. *The data structure needs to be flexible enough to generate all required kinds of summaries: single- vs. multi-document, general vs. focused, contrastive and update.*

The first reason for this requirement is that a user needs to be able to dynamically switch between different summary views for a given document collection. Moreover, developing, implementing, an maintaining multiple algorithms would be prohibitively expensive from a software engineering perspective.

Requirement #3: Efficiency. *The data structure must be abstract enough for summary generation, while at the same time simple enough to be computed in a speedy and robust fashion.*

This requirement ensures the ineligibility of highly sophisticated proposals that are not possible to implement in contemporary desktop environments.

[3] An example for a DUC 2005 focused summary context is: *"What countries are or have been involved in land or water boundary disputes with each other over oil resources or exploration? How have disputes been resolved, or towards what kind of resolution are the countries moving? What other factors affect the disputes?"*

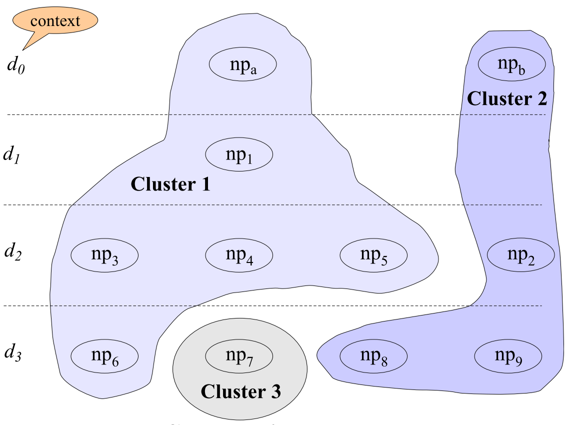

Fig. 1: *Topic Clusters for a three-document set d_1–d_3, plus an additional context document d_0 for focused summarization*

3.2 Topic Cluster Definition

We now present our approach satisfying the requirements stated in the previous subsection. As we demonstrate below, it helps to abstract from the individual words within a document collection when generating summaries of various kinds. As a useful level of abstraction we use the notion of a *topic*, a particular theme within a set of documents.

We can now define topic *clusters* as an abstract representation of the topics occurring within a document collection. A single topic cluster represents the set of all entities in a document set pertaining to the topic.[4] All topic clusters together (i.e., a topic cluster set) represent the entirety of themes in a document set. Note that each topic cluster has a certain *size*, i.e., the number of contained entities, and spans a certain subset of documents, namely those containing the entities making up the cluster.

An example for topic clusters, generated on a hypothetical four-document set, is shown in Fig. 1. Here, an algorithm detected three topic clusters, using noun phrases (NPs) as the underlying entities. To show a possible implementation for generating topic clusters, we briefly present our algorithm in the next subsection.

3.3 Generating Topics Clusters

Our approach for generating topic clusters relies on a fuzzy set theory-based clustering algorithm working on coreference chains. The algorithm is described in detail in [12]. Within this paper, we only present a brief summary of the main steps. The input is a set of documents, as it could have been generated from a search engine. With this set, we perform three steps to obtain a topic cluster, which can then be used for summary generation as shown in Section 4.

Noun Phrases. The first step in our approach is the generation of noun phrase (NP) chunks for each document. This can be easily achieved with off-the-shelf part-of-speech taggers and transducer-based NP chunkers, which are available for most languages.

Coreference Chains. In a second step, we produce inter- and intra-document coreference chains from the generated NPs. Our approach relies on a fuzzy-set based approach [11], but in principle any coreference algorithm can be deployed for this step.

[4] As a useful entity size we empirically determined base noun phrases (NPs).

Fuzzy Clustering. The final step takes the computed inter- and intra-document chains and clusters them with a fuzzy algorithm [12]. The output of this algorithm is a set of NP clusters. Each cluster has a certain size (number of contained NPs) and spans a certain number of documents. Thus, the end result directly corresponds to the topic cluster data structure.

4 Summary Generation

In this section, we show how the various multi-document summaries defined in Section 2 can be generated based on topic clusters.

4.1 Generic Summaries

We begin by discussing the generation of generic multi-document summaries. Although not part of our list of next-generation summaries, this already illustrates the main points of summary generation based on topic clusters and also provides the foundation for the generation of advanced summary types.

The aim in generic multi-document summarization is to identify the most salient (shared) topics within a collection of documents. The summary, typically sentences (or sentence parts) extracted from the documents, should reflect as many common topics as space permits. This kind of summary can be immediately generated from the topic cluster data structure: Topics are identified by clusters, so by extracting those clusters that span all documents (or a sufficiently large subset thereof), a summarizer can obtain the common themes of all documents. In order to rank the topics by relevance, a summarizer can evaluate the *size* of each cluster: the larger a cluster, the more important the topic contained within.

Based on these ideas, we can define a strategy that generates multi-document summaries by selecting (at least) one candidate noun phrase from each topic cluster, in decreasing order of importance (topic cluster size), until a prescribed size limit has been reached or all topics are exhausted. The candidate NPs, in turn, can be used to select the sentences they appear in as a candidate text extract. Fig. 3 (top section) shows an example for a (roughly) 100-word summary generated with this strategy from DUC 2004 data [8].

Extractive sentence-based summarization typically involves additional techniques, i.e., replacing dangling pronominal references, eliminating duplicate noun phrases, or removing relative clauses. However, within the scope of this paper we are not concerned with this kind of post-processing, which is widely discussed in the literature (see e.g. [5] and the DUC proceedings).

4.2 Contrastive Summaries

Contrastive summaries consist of two parts: a summary of the *common themes* across all documents, and document-specific *contrastive themes*. The first part is identical to generic multi-document summarization as described in the previous subsection.

The topic cluster provides all information required for contrast detection: topics that span all documents (or a configurable percentage, e.g., $> 90\%$) are common topics, as they are used for generic multi-document summaries. Topics covered only in a single document (or, again, in a subset, say, $< 5\%$) indicate unique, distinguishing topics. For example, in Fig. 1, Cluster 3 would be such a (single-element) cluster representing a distinguishing topic for document d_3.

Common clusters	
Hurricane Mitch in Central America (31) – Honduras (21) – the country's central coast (15) – last week's storm (12)	

Distinctive clusters	
D_1	Gen. Mario Hung Pacheco – the shelves of some stores and some gasoline stations – mayor of Utila – a hurricane warning – the northwest Caribbean for five days
D_2	the western Caribbean on Wednesday – 165 kms – Honduras with 120 – west at only 2 mph – a resident of Guanaja Island
D_3	the center – emergency measures on the Caribbean coast of the Yucatan Peninsula – a boat – hotels – The storm's power
D_4	the storm's death toll in the region to 357 – 231 people have been confirmed dead
D_5	floods – the Guatemalan border – a state of emergency – 50 kph – late Sunday
D_6	area – the slopes of the Casita volcano in northern Nicaragua – Sunday night – a 32-square mile – addition
D_7	homes – The greatest losses – affiliate in San Miguel province – a statement – the EU
D_8	the audience – all public and private institutions and all men – the pope – a gift – six Russian cosmonauts
D_9	access to places – other countries – the recovery effort – More help – at least 300 children at the shelter for diarrhea, conjunctivitis and bacterial infections
D_{10}	Taiwan – aid and pledges of assistance – Residents – Cuba's offer – the saddest thing

Fig. 2: *Topic cluster results for a set of ten documents on "Hurricane Mitch"*

By sorting these distinguishing clusters by their size, we can obtain a ranked list of topics that are *the most important for a document but not mentioned in any other documents*. Like before, we can select one or several candidate NPs from each cluster (for instance, based on their length or their position within the document) and use those NPs to select sentences for the final output. Of course, for a given document set a topic cluster algorithm might not detect any distinctive clusters. Based on our experiments, this typically happens for very short articles (2–5 sentences), or very homogenous document sets.

As a real-world example, consider the topic cluster generated from the DUC 2004 multi-document set `d30002`. This set contains ten documents, all on the "Hurricane Mitch" topic, each with slightly different information about the same natural disaster. After running our clustering algorithm, we obtained the topics shown in Fig. 2. The *common clusters* section shows the four biggest clusters (with their respective size in brackets), i.e., the most important topics spanning all documents, each identified by a candidate NP.[5] The *distinctive clusters* section shows the five biggest isolated topic clusters for each document with one selected noun phrase each.

How to present such contrastive summaries to the user is highly dependent on the integration within a desktop environment. In addition to the common topic summary as shown above, we currently give the per-document keywords as shown in Fig. 2, which can be expanded to view a sentence extract, like in Fig. 3.

4.3 Focused Summaries

The next type of summary we address here are *focused* summaries, which are not concerned with summarizing a document (set), but rather with collecting information on an explicit interest expressed through context information, like a *user profile*. Focused summaries have been evaluated on a large scale starting with Task 5 in DUC 2004 [8]; in DUC 2005 and 2006, it was the only task (DUC 2007 added the update task).

Topic clusters also allow to generate focused summaries, simply by including the context information as another, distinct document d_0 when computing the topic cluster data structure. Then, all topics that overlap with document d_0 also contain information relevant to the context. All other clusters, even if they are bigger, are discarded for this kind of summary, i.e., we *slice* the topic clusters with the context entities. As before, elements within the clusters have to be further ranked, extracted, and post-processed to create the final summary. Fig. 1 shows an example for this idea: both Cluster 1 and Cluster 2 overlap with the context

[5] Here, we simply used the longest NP, however, a targeted summarizer might apply additional strategies.

Common Topic Summary
The Honduran president closed schools and public offices on the coast Monday and ordered all air force planes and helicopters to evacuate people from the Islas de la Bahia, a string of small islands off the country's central coast. National police spokesman Ivan Mejia said the Coco, Segovia and Cruta rivers all overflowed their banks Monday along Honduras' eastern coast. The European Union on Tuesday approved 6.4 million European currency units (dlrs 7.7 million) in aid for thousands of victims of the devastation caused by Hurricane Mitch in Central America. The greatest losses were reported in Honduras, where an estimated 5,000 people died and 600,000 people – 10 percent of the population – were forced to flee their homes after last week's storm.

Distinctive Topics Summaries
D_1 : The head of the Honduran armed forces, Gen. Mario Hung Pacheco, said 5,000 soldiers were standing by to help victims of the storm, but he warned the military could not reach everyone.
D_2 : Hurricane Mitch paused in its whirl through the western Caribbean on Wednesday to punish Honduras with 120-mph (205-kph) winds, topping trees, sweeping away bridges, flooding neighborhoods and killing at least 32 people.
D_3 : Hurricane-force winds whirled up to 30 miles (50 kilometers) from the center, with rain-laden tropical storm winds extending well beyond that.
D_4 : At least 231 people have been confirmed dead in Honduras from former-hurricane Mitch, bringing the storm's death toll in the region to 357, the National Emergency Commission said Saturday.
D_5 : El Salvador – where 140 people died in flash floods – a state of emergency Saturday, as did Guatemala, where 21 people died when floods swept away their homes.
D_6 : Nicaraguan Vice President Enrique Bolanos said Sunday night that between 1,000 and 1,500 people were buried in a 32-square mile (82.88 square-kilometer) area below the slopes of the Casita volcano in northern Nicaragua.
D_7 : EU spokesman Pietro Petrucci said the funds will be used to provide basic care such as medicine, food, water sanitation and blankets to thousands of people whose homes were destroyed by torrential rains and mudslides.
D_8 : Among those attending the audience were six Russian cosmonauts taking a special course in Italy.
D_9 : Aid groups and governments have called for other countries to send medicine, water, canned food, roofing materials and equipment to help deliver supplies.
D_{10} : Taiwan said today it will donate dlrs 2.6 million in relief to Honduras, Nicaragua, El Salvador and Guatemala.

Fig. 3: *Topic cluster-generated contrastive multi-document summary*

(d_0), hence they contain information relevant for the user's focus. Cluster 3 does not overlap with the context and is therefore discarded (even if it were bigger than the other clusters).

To demonstrate that our approach works cross-lingual and cross-domain, we give an example for a focused summary generated from a German encyclopedia on building architecture. Here, we integrated the summarization component into a semantic desktop for building historians and architects [13]. Fig. 5 shows an excerpt of the result to the question "Welche Art von Putz bietet Schutz vor Witterung?" *(Which kind of plaster would be suitable to protect brickwork against weather influences?).* The first sentence of the automatic summarization means: *The joint filling is important for the resistance of the brickwork, especially for those parts exposed to the weather, as well as the quality of the bricks.* Within a concrete application scenario described in [13], this provides helpful information for an architect, who can now continue to access information in the encyclopedia concerning brick quality using the provided cross-links.

"Summarize Microsoft's antitrust problems, including its alleged illegal behavior and antitrust proceedings against the company."

Docsubset D0711C-A

Ever since ibm selected Microsoft to supply the operating system for the first ibm pc in 1981, ibm has been out-hustled and outsmarted by Microsoft in the pc business – not the sort of behavior the nation's antitrust laws were meant to prohibit. Antitrust experts note, it is certainly legal for Microsoft to offer a distributor – as the ibm pc business was – incentives to aggressively distribute its product. If accepting those incentives means featuring Microsoft's Windows over os/2 – made by another unit of ibm – that is mainly a business decision for ibm to make, they say.

Docsubset D0711C-B

The case is assigned to Jackson, who seeks to expedite the proceedings by setting an early trial date. Microsoft denies all charges of wrongdoing. To Microsoft in 1997, is a harsh critic of the company. Meanwhile, Jackson's ruling could have an impact on other cases already pending against Microsoft. And Microsoft didn't get to be Microsoft by shrinking from the battlefield. Tuesday's courtroom statements – uncompromising attacks from each side – only seemed to underline the gap between the government and Microsoft. In the government's view, Microsoft is trying to change the subject with such arguments.

Docsubset D0711C-C

The company said it would appeal. Klein said the ruling would bring positive change. He said the ruling showed how no company is above the law. The document was part of the public record from an older antitrust case against Microsoft, the Justice Department's first suit against the company. The Justice Department and 17 state attorneys general proposed to break Microsoft into two companies. Local press reports said that attorneys for the Justice Department and the 19 states that successfully sued Microsoft for antitrust violations are considering ways to break up the company as a method to curb anticompetitive practices.

Fig. 4: *Topic cluster-generated update summary for the DUC 2007 data set D0718D (context shown on top)*

"Welche Art von Putz bietet Schutz vor Witterung?"

Ist das Dichten der Fugen für die Erhaltung der Mauerwerke, namentlich an den der Witterung ausgesetzten Stellen, von Wichtigkeit, so ist es nicht minder die Beschaffenheit der Steine selbst. Bei der früher allgemein üblichen Art der gleichzeitigen Ausführung von Verblendung und Hintermauerung war allerdings mannigfach Gelegenheit zur Beschmutzung und Beschädigung der Verblendsteine geboten. Will man einen dauerhaften Putz erzielen, so gilt für alle Arten von Mauerwerk die Regel, da die zu putzenden Flächen frei von Staub sein müssen, da dieser trennend zwischen Mauer und Putz wirken und das feste Anhaften des letzteren verhindern würde. . . .

Fig. 5: *Excerpt from a focused summary generated based on a question (shown on top) from a German encyclopedia on architecture*

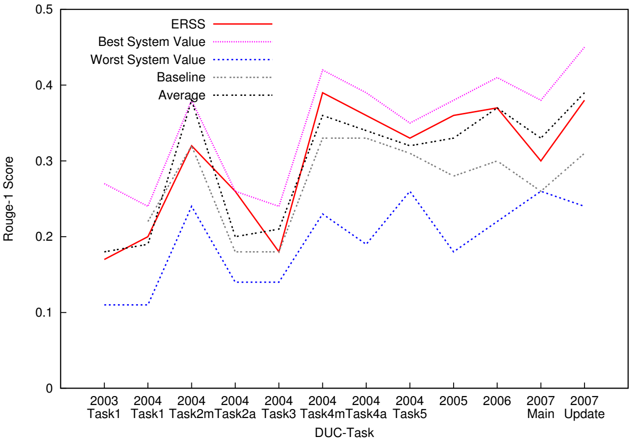

Fig. 6: *Summarization based on topic clusters: ERSS performance on the DUC data from 2003–2007*

4.4 Update Summaries

To generate update summaries, we first generate the topic clusters based on the context and the current set of documents (including all previous documents, i.e., not just the new ones). For the first subset within an update cluster, summary generation is identical to a standard (main task) focused summary, as presented above. For each subsequent update subset, we re-generate the topic cluster, by adding the new documents to the current set.

To generate focused update summaries for the extended document sets, we again select sentences based on a ranking scheme: (1) The highest rank is given to sentences from clusters that overlap with the context (i.e., cover topics from the questions) but do not contain any elements from documents of a previous update (i.e., these are topical information *only* addressed in a new document). (2) A medium rank is given to sentences from clusters that overlap with the context and appear in the newly added (updated) set of documents (i.e., new information addressing a topic that has been addressed before). And (3) the lowest rank is given to all remaining sentences from clusters that overlap with the context (i.e., answer a question from the context).

In Fig. 1, Cluster 2 is an example for a highly ranked cluster after adding d_2, because it overlaps with the context (d_0) and does not contain elements from a previous update (d_1). Thus, the sentences picked from d_2 will contain information regarding the focus question that has not been addressed in a previous document (subset), here, d_1. Note that generic update summaries (without a focus question) can be generated in the same fashion, by simply omitting the context slicing step.

Fig. 4 shows an example for an update summary generated from DUC 2007 data. Compared to a non-update summary of the same set (not shown here due to space constraints), the update summary clearly shows the development of the topic through time—before the trial, during the trial, and its aftermath.

5 Evaluation

We evaluated our ideas with an implementation based on the fuzzy coreference cluster algorithm [12] for generating topic graphs, using the data of the DUC competitions from 2003–2007. These involved yearly changing tasks, including single-document (DUC 2003, Task 1 and DUC 2004, Task 1) and multi-document summaries (DUC 2004, Task 2), short (keyword) vs. long (sentence) summaries, generic (DUC 2003–2004) and focused summaries (DUC 2004, Task 5; DUC 2005–2006; DUC 2007 Main) as well as cross-lingual (DUC 2004, Tasks 3, 4) and update summaries (DUC 2007 Update). We generated the summaries for all of these different tasks with a single system (ERSS), based on the topic cluster as the only data structure.

We use the same evaluation method as in DUC, namely ROUGE[6] [4], to allow a direct comparison of our results with all other systems participating in DUC. Fig. 6 summarizes the results, comparing our system ERSS with the best, worst, average, and baseline system for each year and task.[7] For the detailed results from each year, we refer to reader to our DUC papers [14].

Overall, we can see that the topic graph algorithm performs very competitively with state-of-the-

[6] In this evaluation, we use the ROUGE-1 score only, to allow a comparison for all years and tasks.

[7] Note that the DUC competition so far included no contrastive summarization task, hence this kind of summary is not included in the evaluation.

art multi-document summarization systems. An analysis of the generated summaries showed that the biggest factor negatively impacting ERSS' scores is the current lack of any post-processing (removing dangling references, cleaning up redundancies, etc.).

6 Related Work

Clustering approaches have long been applied to document analysis (see e.g. [1] for an overview), including summarization (e.g., [9]), but our work differs in that we cluster entities (NPs) rather than individual (TF∗IDF-weighted) words.

With respect to contrastive summaries, a motivation related most closely to ours is given by [6] (with previous work in 1997), who also attempt to find both *similarities* and *differences* among related documents. However, Mani [5, p. 188] describes this approach as "rather complex" and "recommended only for pairs of documents," whereas we are concerned with finding contrasts in large document sets (up to 50 for the DUC 2005 data). Also, [6] are not concerned with what we call "contrastive summaries" (as in Fig. 3) but rather present their results in form of sentence extracts aligned between a document pair—which clearly does not help at all in reducing information overload.

In [15], the authors define the problem of "comparative text mining" (CTM) for a given text collection as "(1) discovering the different common themes across all the collections; (2) for each discovered theme, characterize what is in common among all the collections and what is unique to each collection." They also apply a clustering strategy based on a cross-collection mixture model, but using only simple word-level statistics, which we believe is much less useful for creating summaries than our entity-based clustering approach.[8]

The research area of *change summarization* is concerned with tracking a single document (or a document collection) over time and extracting new/fading topics. [2] evaluate such changes, providing the result in form of web page ranking lists.

7 Conclusions and Future Work

In this paper, we investigated several types of multi-document summaries and their generation using a single abstracting data structure, topic clusters.

In particular, we revisited the notion of contrastive summaries, which show, at the same time, both topics *common* to all documents, as well as their *distinctive* information. Although this kind of summary has already been proposed ten years ago by Mani et al. and also alluded to in many other places (e.g., [3]), contrastive summaries are still virtually unknown. We believe this is partly due to the lack of a simple, robust, flexible algorithm, which allows to create this kind of summary from a given document collection. Contrastive summaries are in our view an important contribution to multi-document summarization, especially for less homogeneous collections where the individual documents contain different information only loosely coupled by a common topic. For these collections, a summary of the commonalities does not enable an information seeker to select a relevant document from the collection, and individual summaries are also not guaranteed to highlight the *differences* between the individual documents.

[8] A typical example cluster in [15] is the topic list *"port, jack, ports, will, your, warm, keep, down"*.

From a language engineering perspective, we essentially decoupled the generation of summaries from the generation of the topic cluster data structure. This allows for both, using different algorithms to compute the graph while keeping the summarization engine intact, as well as using the same data structure for generating multiple kinds of summaries. The evaluation we performed on multiple tasks over five years of data from the DUC competition show that this approach is feasible and delivers competitive performance.

More work is needed in determining efficient ways of integrating automatically created summaries in modern desktop environments. For example, a suitable, dynamic web interface could display topics in a hierarchical fashion, which would allow a user to "see" content that appears in a subset, but not in all documents. Summaries could be incrementally expanded, from keyword sets, like in Fig. 2, to complete summaries, with a single click, allowing a user to navigate from highly compressed views over summaries to the complete document. Creating summaries for dynamically changing document collections—like a newswire stream—can enhance the awareness of newly appearing topics (distinctive clusters) and fading topics. As [7] points out, *"What we find changes who we become."*

References

[1] M. W. Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval.* Springer, 2003.

[2] A. Jatowt, K. K. Bun, and M. Ishizuka. Change Summarization in Web Collections. In *Innovations in Applied Artificial Intelligence: 17th Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, LNCS, pages 653–662, 2004.

[3] M.-Y. Kan, K. R. McKeown, and J. L. Klavans. Domain-specific informative and indicative summarization for information retrieval. In *Proc. of the Document Understanding Conference*, New Orleans, U.S.A., 2001.

[4] C.-Y. Lin. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25–26 2004.

[5] I. Mani. *Automatic Summarization.* John Benjamins B.V., 2001.

[6] I. Mani and E. Bloedorn. Summarizing Similarities and Differences Among Related Documents. *Inf. Retr.*, 1(1-2):35–67, 1999.

[7] P. Morville. *Ambient Findability.* O'Reilly, 2005.

[8] NIST, editor. *DUC 2004 Workshop on Text Summarization*, Boston Park Plaza Hotel and Towers, Boston, USA, May 6–7 2004. http://duc.nist.gov/pubs.html#2004.

[9] D. R. Radev, H. Jing, M. Styś, and D. Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938, 2004.

[10] R. Witte. An Integration Architecture for User-Centric Document Creation, Retrieval, and Analysis. In *Proceedings of the VLDB Workshop on Information Integration on the Web (IIWeb'04)*, pages 141–144, Toronto, Canada, August 30 2004. http://rene-witte.net/downloads/witte_iiweb04.pdf.

[11] R. Witte and S. Bergler. Fuzzy Coreference Resolution for Summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, pages 43–50, Venice, Italy, June 23–24 2003. Università Ca' Foscari.

[12] R. Witte and S. Bergler. Fuzzy clustering for topic analysis and summarization of document collections. In Z. Kobti and D. Wu, editors, *Proc. of the 20th Canadian Conference on Artificial Intelligence (Canadian A.I. 2007)*, LNAI 4509, pages 476–488, Montréal, Québec, Canada, May 28–30 2007. Springer.

[13] R. Witte, P. Gerlach, M. Joachim, T. Kappler, R. Krestel, and P. Perera. Engineering a Semantic Desktop for Building Historians and Architects. In *Proc. of the Semantic Desktop Workshop at the ISWC*, volume 175 of *CEUR Workshop Proceedings*, pages 138–152, Galway, Ireland, November 6 2005. http://CEUR-WS.org/Vol-175/34_witte_engineeringsd_final.pdf.

[14] R. Witte, R. Krestel, and S. Bergler. Generating Update Summaries for DUC 2007. In *Proc. of Document Understanding Workshop (DUC)*, Rochester, NY, USA, April 26–27 2007.

[15] C. Zhai, A. Velivelli, and B. Yu. A Cross-Collection Mixture Model for Comparative Text Mining. In *Proc. of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*, pages 743–748. ACM Press, 2004.