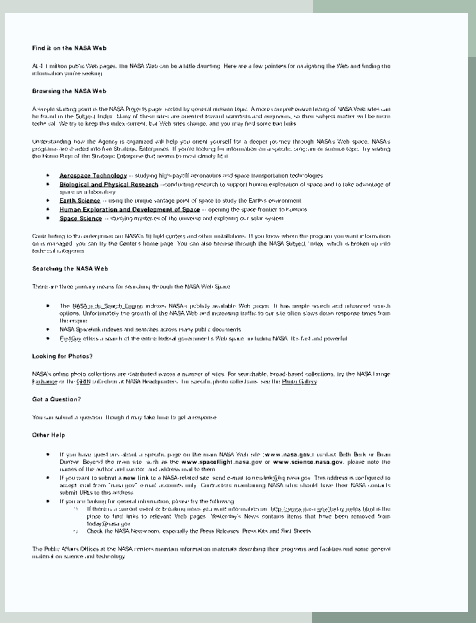
Structured & Unstructured Document Summarization

Hassan Alam, Aman Kumar, Mikako Nakamura, Fuad Rahman, Yuliya Tarnikova, and Che Wilcox

Structured Document



Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting. Here are a few pointers for navigating the Web and finding the information you're seeking. A more comprehensive listing of NASA Web sites can be found in the Subject Index. We try to keep this index current, but Web sites change, and you may find some bad links. NASA's programs are divided into five Strategic Enterprises. NASA's online photo collections are distributed across a number of sites. Contractors maintaining NASA sites should have their NASA contacts submit URLs to this address. Contractors maintaining NASA sites should have their NASA contacts submit URLs to this address.

Flat Summary

This is presented by combining the textual summary with only the title (if any) of the document. This is a quick way of converting a structured document to a non-structured summary form. This works best when the source document has a flat structure. As expected, the document structure is lost, but the generated summary is coherent and meaningful.

Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting. A simple starting point is the NASA Projects page, sorted by general mission topic. A more comprehensive listing of NASA Web sites can be found in the Subject Index.

There are three primary means for searching through the NASA Web Space:
- The NASA-wide Search Engine indexes NASA's publicly available Web pages.
- NASA Spacelink indexes and searches across many public documents.
- FirstGov offers a search of the entire federal government's Web space, including NASA.

NASA's online photo collections are distributed across a number of sites. You can submit a question, though it may take time to get a response. If you have questions about a specific page on the main NASA Web site (www.nasa.gov) contact Beth Beck or Brian Dunbar.

Distributed Flat Summary

In distributed flat summary, each section is given its fair share of representation, calculated by associating the summary length of each section to the corresponding content weight. This is a quick way of converting a structured document to a flat summary form and works best when the source document is structured with uneven content distribution.

Find it on the NASA Web
At 4.1 million public Web pages, the NASA Web can be a little daunting.

Browsing the NASA Web
A simple starting point is the NASA Projects page, sorted by general mission topic.
A more comprehensive listing of NASA Web sites can be found in the Subject Index.

Searching the NASA Web
There are three primary means for searching through the NASA Web Space:
- The NASA-wide Search Engine indexes NASA's publicly available Web pages.
- NASA Spacelink indexes and searches across many public documents.
- FirstGov offers a search of the entire federal government's Web space, including NASA.

Looking for Photos?
NASA's online photo collections are distributed across a number of sites.

Got a Question?
You can submit a question, though it may take time to get a response.

Other Help
If you have questions about a specific page on the main NASA Web site (www.nasa.gov,) contact Beth Beck or Brian Dunbar.

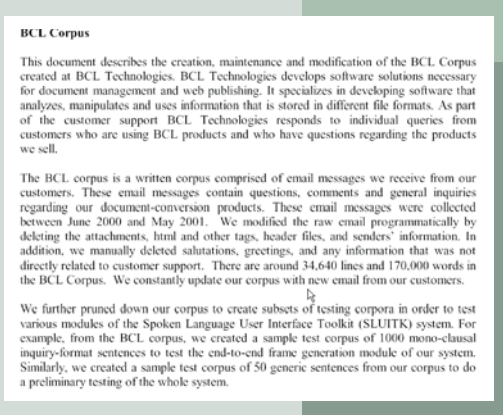
Structured Summary

This is presented by combining the textual summary with overall structure of the document. This preserves the structure of the original document and super-imposes the summary on that structure. This works best when the source document has a well-defined hierarchical structure, the content is evenly distributed and the composition is focused on a small number of themes.

Unstructured Document

Lexical Chains

Lexical chains have been used to create summaries of a document. Cohesion is a way of connecting different parts of text into a single theme. In other words, this is a list of semantically related words, constructed by the use of co-reference, ellipses and conjunctions. This aims to identify the relationship between words that tend to co-occur in the same lexical context.

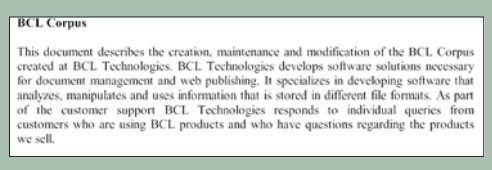


BCL Corpus

This document describes the creation, maintenance and modification of the BCL Corpus created at BCL Technologies. BCL Technologies develops software solutions necessary for document management and web publishing. It specializes in developing software that analyzes, manipulates and uses information that is stored in different file formats. As part of the customer support BCL Technologies responds to individual queries from customers who are using BCL products and who have questions regarding the products we sell.

The BCL corpus is a written corpus comprised of email messages we receive from our customers. These email messages contain questions, comments and general inquiries regarding our document-conversion products. These email messages were collected between June 2000 and May 2001. We modified the raw email programmatically by deleting the attachments, html and other tags, header files, and senders' information. In addition, we manually deleted salutations, greetings, and any information that was not directly related to customer support. There are around 34,640 lines and 170,000 words in the BCL Corpus. We constantly update our corpus with new email from our customers.

We further pruned down our corpus to create subsets of testing corpora in order to test various modules of the Spoken Language User Interface Toolkit (SLUITK) system. For example, from the BCL corpus, we created a sample test corpus of 1000 mono-clausal inquiry-format sentences to test the end-to-end frame generation module of our system. Similarly, we created a sample test corpus of 50 generic sentences from our corpus to do a preliminary testing of the whole system.



BCL Corpus

This document describes the creation, maintenance and modification of the BCL Corpus created at BCL Technologies. BCL Technologies develops software solutions necessary for document management and web publishing. It specializes in developing software that analyzes, manipulates and uses information that is stored in different file formats. As part of the customer support BCL Technologies responds to individual queries from customers who are using BCL products and who have questions regarding the products we sell.

Flat Summary

This is presented by combining the textual summary with only the title (if any) of the document.

www.bcltechnologies.com

BCL Technologies Inc. 990 Linden Dr., Suite #203, Santa Clara CA 95050, USA