Context-based URL Summarization

Yves Petinot
ypetinot@cs.columbia.edu
Department of Computer Science, Columbia University
503 Computer Science Building
1214 Amsterdam Avenue
New York, New York, 10027

Research Advisor: Prof. Kathleen McKeown

This work is an attempt to solve the problem of automatically producing static (as opposed to query-dependent), human-readable, summaries for Web URLs. We propose a novel approach for context-based summarization where contextual data is used primarily to guide the summarization process. More specifically we use contextual data to retrieve descriptive content that is similar to the expected target summary and use text-to-text summarization to produce the final summary. We describe the implementation of our prototype and report on our experimental results which show a meaningful improvement over a state-of-the-art extractive context-based summarizer. We conclude with a discussion on on-going and future work and especially on how we seek to further improve our system. We propose a road-map towards a holistic model for the fusion of information sources scattered across the Web. This research project lies at the crossroad of Web Information Extraction and Natural Language Processing and was initially presented at Hypertext 2008.

Motivation

Most - if not all - commercial Web Search Engines extensively rely on static URL summaries to provide default URL descriptions for their Search Engine Result Pages (SERPs). These summaries, which are either found directly in the URL content (meta-description) or in Web Directories such as DMOZ - are typically used when a URL is retrieved in response to a navigational query [Broder2002], or as a result of an anchor-text match if there is no match for the user query in the content of the page. Another typical case where static summaries are needed is when the content of a URL is predominantly multimedia, offering no text data towards the creation of a snippet. For instance at the time of this writing, Google uses the following DMOZ summary as a description for the official - mostly flash-based - web-site of *ABC Networks*:

> "Official site. Includes behind-the-scenes information, cast biographies, photos, and video clips."

Unfortunately this type of backup summaries are usually unavailable for less popular URLs, putting into light some fundamental limitations of Web Directories, which, and this is for obvious quality concerns, are currently manually curated. Those limitations are:

1. **Coverage**: while Search Engines are arguably able to successfully index the largest part of the visible Web, editors contributing to Web Directories are fighting a losing battle and can only cover a small fraction of the Web. As supporting evidence, at the time of this writing, DMOZ has about 5 million entries, while the number of pages on the visible Web already by far exceeds 5 billions. This means that most results in a typical SERP will not be listed in any Web Directory, leaving the Search Engine with no backup alternative in case the content-based data yields an empty - or inappropriate - title and/or description.
2. **Bias**: the URL summaries featured in Web Directories represent the views of a small set of individuals which might effectively differ from the way some of those URLs are perceived by the Web community as a whole.

To address these issues, we seek to develop a scalable, robust and non-biased alternative to DMOZ as a provider of static Web summaries.

Problem Space & Related Work

Before discussing specific solutions in the next sections, it is important to get a comprehensive view of what information that can be leveraged to summarize Web-sites. For an arbitrary Web-site, the range of information available spans the following categories: content-based, context-based and social. Intuitively an optimal summarization algorithm should be able to leverage all categories to generate summaries that are not only consistent with the target's content but also reflect the "aura" of the target throughout the Web.

Content Data

As with any text-based summarization task, the content of the target document is usually available and can be used as the main source of information for summary creation. In fact there has been a significant body of research on applying traditional summarization techniques to the specific area of Web summarization [Berger2000, Buyukkokten2001, Zhang2004].The Web-site summarization task however has certain specificities due to the multimedia nature and structural complexity of Web-data. Multimedia content hinders the location and extraction of important content. For instance it is important to be able to distinguish between navigational and content parts in a Web-page. Additionally in many instances the content of a Web-page can be entirely non-textual - as is the case with any flash-based site - making it very difficult for any Web-agent to extract substantial text information from the site itself.

Context Data

The main characteristic of Web content is in the ability for one document, or page, to link to another. While there are various ways of linking from one page to another - including through images, client side scripts, etc. - page creators predominantly choose to do so by using text that includes a short snippet about the target page. As a consequence, *Anchor-text*, that is the sum of all such snippets for a given URL, has attracted a lot of interest from the Information Retrieval (IR) community as an accurate representation of the content of the target page [McBryan1994,Glover2002,Sun2005]. However the question of how to use this context data for summarization purposes remains open.

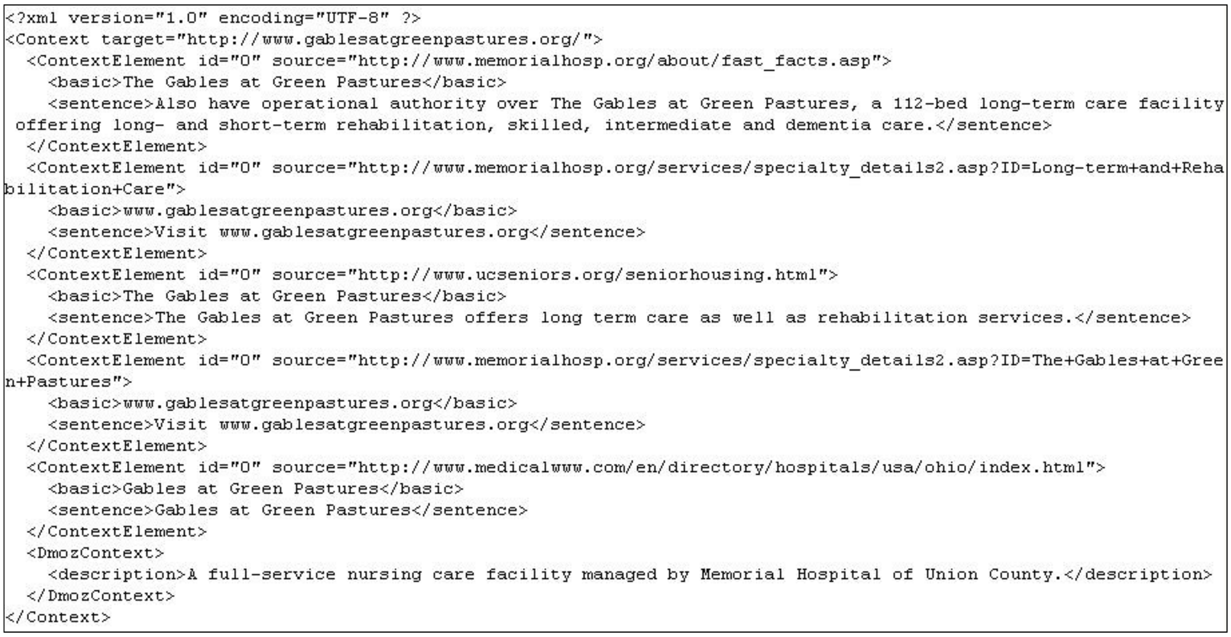
For summarization applications, elements of context, that is the text associated with a single link, can be extracted in various ways. For instance a Web page linking to the homepage of the Amtrak railway company might do so as follows:

<a href=”http://www.amtrak.com”>Amtrack Schedules</a>

We refer to the string *Amtrack Schedules* as an element of basic anchor-text for the URL *http://www.amtrak.com*. In turn, we refer to the sum of all such textual fragments as the basic context of a URL. Alternatively the concept of anchor-text can be extended to include the complete sentence surrounding such links, e.g.:

To schedule your trip, check <a href=”http://www.amtrak.com”>Amtrack Schedules</a>.

thereby contributing the complete sentence *To schedule your trip check Amtrack Schedules* to the extended anchor-text of *http://www.amtrak.com*. Although the extended context might only be referring to the target URL - as opposed to describing it - we can however see that it will typically help disambiguating the basic context.



```
<?xml version="1.0" encoding="UTF-8" ?>
<Context target="http://www.gablesatgreenpastures.org/">
  <ContextElement id="0" source="http://www.memorialhosp.org/about/fast_facts.asp">
    <basic>The Gables at Green Pastures</basic>
    <sentence>Also have operational authority over The Gables at Green Pastures, a 112-bed long-term care facility
 offering long- and short-term rehabilitation, skilled, intermediate and dementia care.</sentence>
  </ContextElement>
  <ContextElement id="0" source="http://www.memorialhosp.org/services/specialty_details2.asp?ID=Long-term+and+Reha
bilitation+Care">
    <basic>www.gablesatgreenpastures.org</basic>
    <sentence>Visit www.gablesatgreenpastures.org</sentence>
  </ContextElement>
  <ContextElement id="0" source="http://www.ucseniors.org/seniorhousing.html">
    <basic>The Gables at Green Pastures</basic>
    <sentence>The Gables at Green Pastures offers long term care as well as rehabilitation services.</sentence>
  </ContextElement>
  <ContextElement id="0" source="http://www.memorialhosp.org/services/specialty_details2.asp?ID=The+Gables+at+Gree
n+Pastures">
    <basic>www.gablesatgreenpastures.org</basic>
    <sentence>Visit www.gablesatgreenpastures.org</sentence>
  </ContextElement>
  <ContextElement id="0" source="http://www.medicalwww.com/en/directory/hospitals/usa/ohio/index.html">
    <basic>Gables at Green Pastures</basic>
    <sentence>Gables at Green Pastures</sentence>
  </ContextElement>
  <DmozContext>
    <description>A full-service nursing care facility managed by Memorial Hospital of Union County.</description>
  </DmozContext>
</Context>
```

***Figure 1**: Context data and DMOZ gold-standard for the URL http://www.gablesatgreenpastures.org*

While there have been previous attempts at leveraging URL contexts for summarization tasks [Amitay2000, Delort2003], all existing approaches have been based on the assumption that summaries, or their component sentences, can be extracted verbatim from the target’s context. [Amitay2000] proposed a solution to automatically collect entire summaries from web-pages that link to the site to be summarized. [Delort2003] built on this idea but, instead of looking for ”pre-built” summaries in the target’s context, proposed to construct summaries by extracting, clustering and concatenating sentences from any part of the target’s context. Irrespective of the approach taken, these works emphasize the importance of *subject sentences* where the target is effectively described - over *reference sentences* which only refer to the target. While extractive methods are able to identify subject sentences and to produce summaries of reasonable quality, they are, however, unable to guarantee the coherence and/or readability of the resulting summaries. More importantly, since they always work with unaltered content, they can potentially yield summaries that include content that is not supported by the target. We thus believe that extractive multi-document Web summarization is unsuitable as a reliable source of Web summaries, especially for consumer-facing applications.

Social Data

The context of a URL can be supplemented by additional social data, that is, data that does not directly link to the target URL, but can somehow be associated with it. This includes:

- **Web 2.0 tags**:(e.g. del.icio.us tags)

- **Search Engines Queries**: queries for which the target URL is one of the results in the Search Engine Result Page (SERP) and for which the URL was actually picked by some users. The fact that a result URL U is clicked on for a given query Q typically shows a strong relationship between U and Q. This type of relationship is obtained from query logs, and more specifically from the click-through data accumulated by the search engine. In practice we have noticed that related queries are a great source of noise and must therefore be integrated carefully with other sources: indeed if query can trigger the retrieval of the target URL it can contain keywords with limited connection with the target. As such extensive filtering would be desirable to make sure we do not introduce a significant amount of noise when including related queries in our context.

Temporal Data

If for now the content, contextual, and even social, data are considered to be completely static, each of these data sources can be seen has having both a static and dynamic component. The dynamic component of a source can help identify temporary or time-dependent characteristics of the target Web-site, while its static component will typically indicate long term characteristics of the target. This information could therefore be leveraged for summarization purposes [Jatowt2006]. This type of data analysis is currently beyond the scope of the work presented here, but is definitely one of the avenues we wish to explore in the near future.

Proposed Approach

The ability for context-based summarizers to go beyond extraction by transforming and/or abstracting content not fully supported by the target can lead to better summaries. This motivates our proposal for NLP-rich solutions that can not only support true fusion of distributed content [Barzilay2005] but also its transformation.

To test our intuition, we have implemented a context-based summarizer called CONFUSIUS (CONtext FUSIon Url Summarizer). CONFUSIUS uses context data only to guide the summarization process. This means that, unlike extractive approaches, it does not attempt to find a complete summary in the context data but simply a characterization of the target URL. This characterization is used to identify similar URLs for which we have descriptive content available. This descriptive content is then be combined, or fused, with the original context data to produce a summary for the target URL. Our current implementation uses a simple word-based vector model as a characterization of URLs and the DMOZ dataset as a large source of descriptive content.

The overall architecture of CONFUSIUS is presented in Figure 2. Below we review the summarization process as well as the functionality of each component.

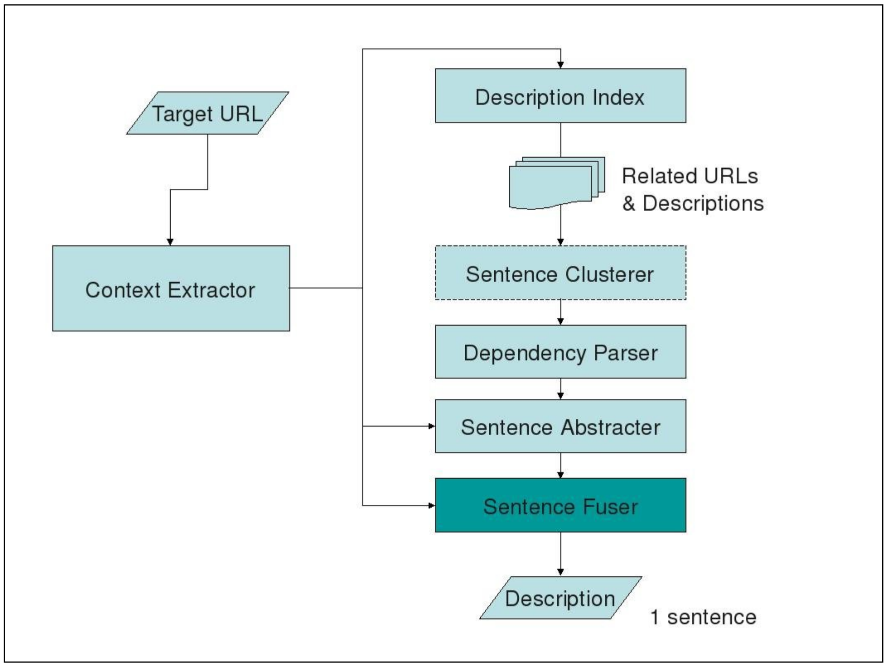


Figure 2*: Architecture of our context-based summarizer*

Context Extractor: Given a target URL TARGET_URL, the role of this component is to collect the complete context data of TARGET_URL. Collecting this data requires having access to a near complete Web-graph and is a fairly expensive operation. Similarly to [Delort2003] we use a commercial Search Engine to acquire the list of URLs forming the context of TARGET_URL. We download the content of each URL discovered through this operation, parse it and extract both the basic and extended context of the hyper-link(s) to TARGET_URL. The output of the context extractor is an XML file which gathers all flavors of context for each known link to TARGET_URL. A sample XML file is shown in Figure 1.

Description Index: In order to identify and retrieve a set of descriptions that are the most similar (i.e. most relevant) to the context of the target, we indexed the entire DMOZ dataset [DMOZ]. The DMOZ dataset consists of about 4 million entries, each consisting of a 4-tuple [URL, Title, Description, Topic] (note that a single URL can be listed under multiple Topics, and as a consequence can be associated with more than one Title/Description). The target context obtained from the context extractor is mapped to a single large query which is built according to the following procedure:

- Any token that appears in the target URL is filtered out, the intuition being that such tokens are very likely to bias the subsequent retrieval results towards unrelated sites.

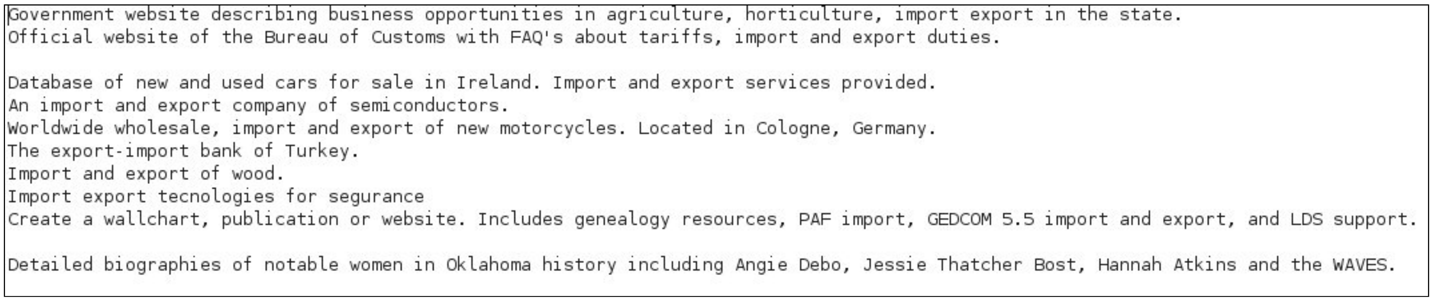
- Each query term is boosted according to its term-frequency (tf) in the context data

This query is submitted to the index which returns a ranked list of URLs (discarded), together with their associated descriptions (sample output), which are passed to the next component. At this stage the assumption is that the URL's returned by the index are matching TARGET_URL's context and are therefore similar to TARGET_URL. We then seek to generate a summary for TARGET_URL based on those URLs' descriptions.

Although we are currently limiting ourselves to DMOZ descriptions, this part of the system can be readily extended to support additional sources of descriptive content, such as meta-descriptions (i.e. descriptions that are directly embedded in Web-documents by their authors.)

Sentence Clusterer: The descriptions returned at the previous stage may not belong to a single theme (as defined in [Barzilay2005]), which can be caused either by limited or noisy context data or simply by a lack of related URLs in the description index. In order to be able to extract the common backbone from a group of descriptions it is important that those descriptions be comparable (strongly dissimilar descriptions may introduce a significant amount of noise in the fusion process presented below and therefore we choose to apply the fusion algorithm only to homogeneous description clusters). To this end we perform similarity-based string clustering. We don't know a-priori how many clusters should be created and therefore we use a parametric clustering algorithm where the maximum cluster diameter is the minimum expected similarity between any two descriptions belonging to the same cluster. The key issue in tuning our clusterer is to find the minimum similarity that causes fairly similar descriptions to be grouped together, while separating completely unrelated strings. We find that a fairly low similarity threshold actually works well in order to produce homogeneous description theme clusters. A potential future development will be the introduction of a parameter-free clustering algorithm. A sample output of the clusters output by this component is presented in Figure 3.

At this point note that we filter out clusters of size 1 (i.e. clusters that contain only one description): no abstraction is possible with a single description and we cannot assume that a single arbitrary description can be a valid description for the target. Finally we rank the remaining clusters by decreasing size and select the largest cluster for further processing: provided the amount of descriptions retrieved at the previous stage remains fairly small (we are currently keeping only the top 10 descriptions returned by the index), we expect the largest cluster to be representative of the theme of the target URL.



```
Government website describing business opportunities in agriculture, horticulture, import export in the state.
Official website of the Bureau of Customs with FAQ's about tariffs, import and export duties.

Database of new and used cars for sale in Ireland. Import and export services provided.
An import and export company of semiconductors.
Worldwide wholesale, import and export of new motorcycles. Located in Cologne, Germany.
The export-import bank of Turkey.
Import and export of wood.
Import export tecnologies for segurance
Create a wallchart, publication or website. Includes genealogy resources, PAF import, GEDCOM 5.5 import and export, and LDS support.

Detailed biographies of notable women in Oklahoma history including Angie Debo, Jessie Thatcher Bost, Hannah Atkins and the WAVES.
```

***Figure 3**: Related description clusters for the URL http://www.debo.it.*
The third cluster is discarded (size 1). The second cluster, which is the largest,
is selected as the best candidate and is used for text-to-text generation, leading to the description
"Import and export services provided." which is a valid summary for this URL.

Sentence Fusion: This component relies on a deeper analysis of the related description cluster generated at the previous stage. To this end we first generate a dependency parse of all sentences (currently using the Stanford Parser). Given this input (sample input), we then use the sentence-fusion algorithm described in [Barzilay2005] to generate a single sentence that combines common content from the – comparable - input sentences. More specifically, the input of the fusion algorithm is a collection of dependency parse trees for the input sentences. The fusion component attempts to generate a single sentence that reflects the information common to all the input sentences. To do so it first identifies sub-trees common to most input sentences. Those sub-trees are then combined into a fusion lattice that models all possible verbalizations of the information shared by the input sentences. Candidate sentences are finally obtained by linearizing this fusion lattice, the best sentence (sample output) being selected based on a language model (lowest entropy).

Experimental Results

We are interested in seeing how our approach would fare against the best known context-based summarizers. To this end we have developed an evaluation framework to assess the overall quality of Web summaries produced by various Web-summarizers. Below we present the results of a large scale evaluation of the approach described above and compare it with a state-of-the-art baseline.

Baseline Summarizer

In this set of experiments our baseline is the context-only summarizer described in [Delort2003]. At the time of this writing, we have not been able to get access to the original system, nor to the original results. We therefore resolved to create a best-effort implementation of this summarizer. As this system essentially behaves as a multi-document summarizer - where each sentence surrounding a link to the target is considered a document of its own - we chose to use the MEAD summarization framework [Radev2004] as a basis for our implementation. While we cannot guarantee that our corpus and the corresponding context data are identical to what was used in [Delort2003], we have been able to obtain a comparable performance on non-filtered context data and therefore believe that this baseline, which we refer to as MEAD in the following, is likely to represent the state-of-the-art of context-based summarizers.

Evaluation Corpus & Methodology

We follow the evaluation methodology described in [Delort2003] and randomly select 2000 URLs from the DMOZ dataset. Since our approach is currently targeting the English language (the Part-of-Speech tagger and sentence-fusion components expect English content) we restrict ourselves to the English part of the DMOZ dataset and discard any entry listed under the Top/International/ category.

The context data for each TARGET_URL is generated using the following process:

1. Collect the set CONTEXT_URLS of all known Web URLs linking to TARGET_URL. To obtain this kind of information we need to have access to a nearly complete Web graph. Our experimentation framework currently relies on Yahoo's Site Explorer API for this task.
2. For each URL in CONTEXT_URLS, download its content and extract various flavors of contextual data:
    - Basic context: the basic anchor-text for any hyper-link to TARGET_URL
    - Extended context: the complete sentence surrounding any hyper-link to TARGET_URL
3. Filter unwanted elements of context (see below).
4. All the context data is dumped to a single XML file that is used as a common input by all summarizers.

The filtering step (3) is necessary to remove unwanted noise or unwanted data from a TARGET_URL's context. In particular we observed that an interesting and significant side-effect of using DMOZ URLs for evaluation purposes is that unfiltered context data will frequently contain exact duplicates of the TARGET_URL's DMOZ summary. This is due to the significant amount of replication of DMOZ data throughout the Web, including DMOZ mirrors, as well as many directory-like pages that reuse DMOZ descriptions verbatim. If we were to allow those elements of context to be used by our context-based summarizers including our baseline we would be facing a chicken and egg problem where the summarization process would be intrinsically biased towards the DMOZ gold-standard. Our baseline summarizer, based after [Delort2003]'s algorithm and described below, is extremely effective at picking DMOZ-duplicates among all context elements and our initial experiments showed that not removing DMOZ duplicates would artificially boost the baseline performance by nearly 200% making it an invalid baseline.

The DMOZ summary of each URL is used as a gold standard against which to compare the quality of the summaries generated by the various summarizers we wish to evaluate. We use ROUGE-1 [Lin2004], a standard evaluation measure, as our similarity measure. Finally in order to perform as fair a comparison as possible between our summarizers, we truncate each summary to the length (in words) of the shortest summary. By doing so, we implicitly put more emphasis on precision than on recall.

Results

Experimental results for the summarizers and configurations we tested are presented in Table 1.

	MEAD w/ Extended Context	CONFUSIUS w/ Basic Context	CONFUSIUS w/ Exten
ROUGE-1 Average Recall (95% Conf. Int.)	0.08566 (0.07912 - 0.09200)	0.06927 (0.06460 - 0.07487)	0.08466 (0.07865 - 0.09
ROUGE-1 Average Precision (95% Conf. Int.)	0.09500 (0.08814 - 0.10158)	0.09556 (0.08866 - 0.10238)	0.10081 (0.09406 - 0.10
ROUGE-1 Average F-Measure (95% Conf. Int.)	**0.07829 (0.07304 - 0.08319)**	**0.07253 (0.06773 - 0.07762)**	**0.08273 (0.07708 - 0.08**

Table 1: *Experimental results for the MEAD baseline and two configurations of the CONFUSIUS summarizer.*

The first column in Table 1 shows the reference performance of the MEAD baseline while the second and third column show the performance of two configurations of our CONFUSIUS summarizer.

In the first configuration, the input to the retrieval stage of CONFUSIUS is the basic context data, that is the text strictly contained within the HTML links to the TARGET_URL. This is our original implementation of CONFUSIUS. Although the amount of basic context data is quite a lot less than the extended context data used by MEAD, we see that CONFUSIUS reaches an average performance (F-Measure) close to that of the baseline. While both systems have nearly comparable performance, we observe that the level of precision of CONFUSIUS is slightly better than that of MEAD which should translate into more reliable summaries. By analyzing the cases where MEAD did better than CONFUSIUS we observed that in most the ambiguity of the basic context data is causing CONFUSIUS to produce extremely irrelevant summaries.

In the second configuration, the input to the retrieval stage of CONFUSIUS is the extended context data, that is the entire sentence associated with the HTML links to the TARGET_URL. By using the extended context data we intended to reduce the level of ambiguity of the context information. The result of this experiment appears in the third column of Table 1 and shows that our intuition was correct. With more context information, CONFUSIUS is able to retrieve more relevant descriptions from the DMOZ index and to outperform MEAD with a 6% increase in Precision and F-measure (the set of similarities produced by both MEAD and CONFUSIUS are non-normally distributed, and the Wilcoxon signed rank test confirms that the difference is significant, with a p-value of 0.029).

Conclusions and Future Work

In this paper we presented a proof of concept for a novel approach to context-based URL summarization. Traditional summarization techniques, which have been proven to work well on text that is both grammatically-correct and abundant, are inappropriate when dealing with Web-content, be it from the target page or its context. We argued that novel summarization techniques are therefore needed to make the most of data that is inherently short and noisy. Specifically we proposed a novel approach to context-based summarization where the context data is used only to guide the summarization process and where the final summary is generated using related descriptive content that is fed to a text-to-text generation algorithm. While still at an early development stage, a large scale evaluation of our prototype showed that this approach is already able to outperform a state-of-the-art context-based summarizer.

Our current work is concerned with further developing this concept and to formulate a holistic summarization model that describes how to merge Web content from multiple sources. To this end, we believe it is fundamental to improve our ability to mine and model context data. We saw that while context data does not necessarily contain full-fledged summaries, it does contain substantial information about the entities (i.e. real-world objects, functionalities, topics, etc.) associated with the target. Being able to identify those entities, their types, and the relationships between them is key in reaching a near semantic representation of the knowledge associated with a URL. A human-readable summary can in turn be generated by verbalizing such a representation. As our verbalization process relies on reusing descriptive content, we are also working on an algorithm makes it possible to abstract and reuse existing descriptive content.

References

[Amitay2000]: Einat Amitay, Cecile Paris: Automatically Summarising Web Sites - Is There A Way Around It? CIKM 2000: 173-179

[Barzilay2005]: Regina Barzilay, Kathleen McKeown: Sentence Fusion for Multidocument News Summarization. Computational Linguistics 31(3): 297-328 (2005)

[Berger2000]: A. Berger and V. Mittal, OCELOT: a system for summarizing Web pages. In proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00), pages 144-151, 2000.

[Broder2002]: Andrei Broder, A taxonomy of web search, SIGIR Forum 36, page 3-10, 2002

[Buyukkokten2001]: O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for Web browsing on handheld devices. In Proceedings of the 10th International World Wide Web Conference, pages 652-662, 2001.

[Delort2003]: Jean-Yves Delort, Bernadette Bouchon-Meunier, Maria Rifqi: Enhanced web document summarization using hyperlinks. Hypertext 2003: 208-215

[Jatowt2006]:Adam Jatowt and Mitsuru Ishizuka: Temporal Multi-Page Summarization, Web Intelligence and Agent Systems: An International Journal (WIAS), IOS Press, 4(2), pp. 163-180, 2006.

[Lin2004]: Chin-Yew 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.

[Radev2004]: Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhang Zhu. MEAD - a platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*, Lisbon, Portugal, May 2004.

[Sun2005]: Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, Zheng Chen: Web-page summarization using clickthrough data. SIGIR 2005: 194-201

[Zhang2004]: Y. Zhang, N. Zincir-Heywood, E. Milios, World Wide Web Site Summarization, in Web Intelligence and Agent Systems, 2(1), pages 39-53, 2004.