Multidocument Summarization via Information Extraction

Michael White and Tanya Korelsky
CoGenTex, Inc.
Ithaca, NY
mike,tanya@cogentex.com

Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff
Department of Computer Science
Cornell University, Ithaca, NY
cardie,yung,pierce,wkiri@cs.cornell.edu

ABSTRACT

We present and evaluate the initial version of RIPTIDES, a system that combines information extraction, extraction-based summarization, and natural language generation to support user-directed multidocument summarization.

1. INTRODUCTION

Although recent years has seen increased and successful research efforts in the areas of single-document summarization, multidocument summarization, and information extraction, very few investigations have explored the potential of merging summarization and information extraction techniques. This paper presents and evaluates the initial version of RIPTIDES, a system that combines information extraction (IE), extraction-based summarization, and natural language generation to support user-directed multidocument summarization. (RIPTIDES stands for RapIdly Portable Translingual Information extraction and interactive multiDocumEnt Summarization.) Following [10], we hypothesize that IE-supported summarization will enable the generation of more accurate and targeted summaries in specific domains than is possible with current domain-independent techniques.

In the sections below, we describe the initial implementation and evaluation of the RIPTIDES IE-supported summarization system. We conclude with a brief discussion of related and ongoing work.

2. SYSTEM DESIGN

Figure 1 depicts the IE-supported summarization system. The system first requires that the user select (1) a set of documents in which to search for information, and (2) one or more scenario templates (extraction domains) to activate. The user optionally provides filters and preferences on the scenario template slots, specifying what information s/he wants to be reported in the summary. RIPTIDES next applies its Information Extraction subsystem to generate a database of extracted events for the selected domain and then invokes the Summarizer to generate a natural language summary of the extracted information subject to the user's constraints. In the subsections below, we describe the IE system and the Summarizer in turn.

2.1 IE System

The domain for the initial IE-supported summarization system and its evaluation is natural disasters. Very briefly, a top-level natural disasters scenario template contains: document-level information (e.g. *docno*, *date-time*); zero or more *agent* elements denoting each *person*, *group*, and *organization* in the text; and zero or more *disaster* elements. *Agent* elements encode standard information for named entities (e.g. *name*, *position*, *geo-political unit*). For the most part, *disaster* elements also contain standard event-related fields (e.g. *type*, *number*, *date*, *time*, *location*, *damage* sub-elements).

The final product of the RIPTIDES system, however, is not a set of scenario templates, but a user-directed multidocument summary. This difference in goals influences a number of template design issues. First, disaster elements must distinguish different reports or views of the same event from multiple sources. As a result, the system creates a separate *disaster* event for each such account. Disaster elements should also include the *reporting agent*, *date*, *time*, and *location* whenever possible. In addition, *damage* elements (i.e. *human* and *physical effects*) are best grouped according to the reporting event. Finally, a slight broadening of the IE task was necessary in that extracted text was not constrained to noun phrases. In particular, adjectival and adverbial phrases that encode *reporter confidence*, and sentences and clauses denoting *relief effort* progress appear beneficial for creating informed summaries. Figure 2 shows the scenario template for one of 25 texts tracking the 1998 earthquake in Afghanistan (TDT2 Topic 89). The texts were also manually annotated for noun phrase coreference; any phrase involved in a coreference relation appears underlined in the running text.

The RIPTIDES system for the most part employs a traditional IE architecture [4]. In addition, we use an in-house implementation of the TIPSTER architecture [8] to manage all linguistic annotations. A preprocessor first finds *sentences* and *tokens*. For syntactic analysis, we currently use the Charniak [5] parser, which creates Penn Treebank-style parses [9] rather than the partial parses used in most IE systems. Output from the parser is converted automatically into TIPSTER *parse* and *part-of-speech* annotations, which are added to the set of linguistic annotations for the document. The extraction phase of the system identifies domain-specific relations among relevant entities in the text. It relies on Autoslog-XML, an XSLT implementation of the Autoslog-TS system [12], to acquire extraction patterns. Autoslog-XML is a weakly supervised learning system that requires two sets of texts for training — one set comprises texts relevant to the domain of interest and the other, texts not relevant

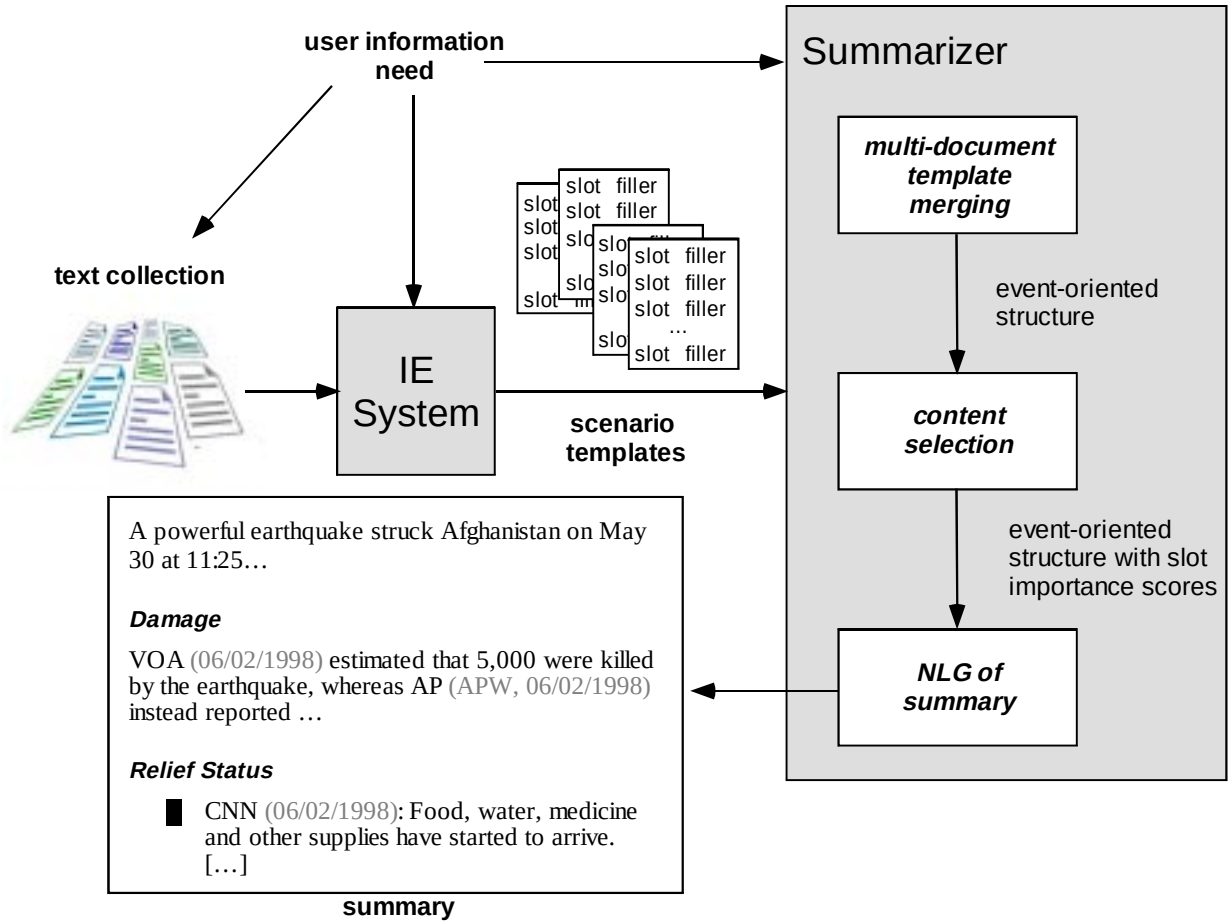

Figure 1. RIPTIDES System Design

to the domain. Based on these and a small set of extraction pattern templates, the system finds a ranked list of possible extraction patterns, which a user then annotates with the appropriate extraction label (e.g. victim). Once acquired, the patterns are applied to new documents to extract slot fillers for the domain. Selectional restrictions on allowable slot fillers are implemented using WordNet [6] and BBN's Identifinder [3] named entity component. In the current version of the system, no coreference resolution is attempted; instead, we rely on a very simple set of heuristics to guide the creation of output templates. The disaster scenario templates extracted for each text are provided as input to the summarization component along with all linguistic annotations accrued in the IE phase. No relief slots are included in the output at present, since there was insufficient annotated data to train a reliable sentence categorizer.

2.2 The Summarizer

In order to include relief and other potentially relevant information not currently found in the scenario templates, the Summarizer extracts selected sentences from the input articles and adds them to the summaries generated from the scenario templates. The extracted sentences are listed under the heading *Selected News Excerpts*, as shown in the two sample summaries appearing in Figures 3 and 4, and discussed further in Section 2.2.5 below.

2.2.1 Summarization Stages

The Summarizer produces each summary in three main stages. In the first stage, the output templates are merged into an event-oriented structure, while keeping track of source information. The merge operation currently relies on simple heuristics to group extracted facts that are comparable; for example, during this phase damage reports are grouped according to whether they pertain to the event as a whole, or instead to damage in the same particular location. Heuristics are also used in this stage to determine the most relevant damage reports, taking into account specificity, recency and news source. Towards the same objective but using a more surface-oriented means, simple word-overlap clustering is used to group sentences from different documents into clusters that are likely to report similar content. In the second stage, a base importance score is first assigned to each slot/sentence based on a combination of document position, document recency and group/cluster membership. The base importance scores are then adjusted according to user-specified preferences and matching

Document no.: ABC19980530.1830.0342
Date/time: 05/30/1998 18:35:42.49
Disaster Type: earthquake
•location: *Afghanistan*
•date: *today*
•magnitude: *6.9*
•magnitude-confidence: high
•epicenter: *a remote part of the country*
•damage:
human-effect:
victim: *Thousands of people*
number: *Thousands*
outcome: dead
confidence: medium
confidence-marker: *feared*
physical-effect:
object: *entire villages*
outcome: damaged
confidence: medium
confidence-marker: *Details now hard to come by / reports say*

PAKISTAN MAY BE PREPARING FOR ANOTHER TEST

Thousands of people are feared dead following... (voice-over) ...a powerful earthquake that hit Afghanistan today. The quake registered 6.9 on the Richter scale, centered in a remote part of the country. (on camera) Details now hard to come by, but reports say entire villages were buried by the quake.

Figure 2. Example scenario template for the natural disasters domain

criteria. The adjusted scores are used to select the most important slots/sentences to include in the summary, subject to the user-specified word limit. In the third and final stage, the summary is generated from the resulting content pool using a combination of top-down, schema-like text building rules and surface-oriented revisions. The extracted sentences are simply listed in document order, grouped into blocks of adjacent sentences.

2.2.2 Specificity of Numeric Estimates

In order to intelligently merge and summarize scenario templates, we found it necessary to explicitly handle numeric estimates of varying specificity. While we did find specific numbers (such as *3,000*) in some damage estimates, we also found cases with no number phrase at all (e.g. *entire villages*). In between these extremes, we found vague estimates (*thousands*) and ranges of numbers (*anywhere from 2,000 to 5,000*). We also found phrases that cannot be easily compared (*more than half the region's residents*).

To merge related damage information, we first calculate the numeric specificity of the estimate as one of the values NONE, VAGUE, RANGE, SPECIFIC, or INCOMPARABLE, based on the presence of a small set of trigger words and phrases (e.g. *several, as many as, from … to*). Next, we identify the most specific current estimates by news source, where a later estimate is considered to update an earlier estimate if it is at least as specific. Finally, we determine two types of derived information units, namely (1) the minimum and maximum estimates across the news sources, and (2) any intermediate estimates that are lower than the maximum estimate.[1]

In the content determination stage, scores are assigned to the derived information units based on the maximum score of the underlying units. In the summary generation stage, a handful of text planning rules are used to organize the text for these derived units, highlighting agreement and disagreement across sources.

2.2.3 Improving the Coherence of Extracted Sentences

In our initial attempt to include extracted sentences, we simply chose the top ranking sentences that would fit within the word limit, subject to the constraint that no more than one sentence per cluster could be chosen, in order to help avoid redundancy. We found that this approach often yielded summaries with very poor coherence, as many of the included sentences were difficult to make sense of in isolation.

To improve the coherence of the extracted sentences, we have experimented with trying to boost coherence by favoring sentences in the context of the highest-ranking sentences over those with lower ranking scores, following the hypothesis that it is better to cover fewer topics in more depth than to change topics excessively. In particular, we assign a score to a set of sentences by summing the base scores plus increasing coherence boosts for adjacent sentences, sentences that precede ones with an initial

[1] Less specific estimates such as "hundreds" are considered lower than more specific numbers such as "5000" when they are lower by more than a factor of 10.

Earthquake strikes Afghanistan

A powerful earthquake struck Afghanistan last Saturday at 11:25. The earthquake was centered in a remote part of the country and had a magnitude of 6.9 on the Richter scale.

Damage

Estimates of the death toll varied. VOA (06/02/1998) provided the highest estimate of 5,000 dead. CNN (05/31/1998) and CNN (06/02/1998) supplied lower estimates of 3,000 and up to 4,000 dead, whereas APW (06/02/1998) gave the lowest estimate of anywhere from 2,000 to 5,000 dead. People were injured, while thousands more were missing. Thousands were homeless.

Quake-devastated villages were damaged. Estimates of the number of villages destroyed varied. CNN (05/31/1998) provided the highest estimate of 50 destroyed, whereas VOA (06/04/1998) gave the lowest estimate of at least 25 destroyed.

In Afghanistan, thousands of people were killed.

Further Details

Heavy after shocks shook northern afghanistan. More homes were destroyed. More villages were damaged.

Landslides or mud slides hit the area.

Another massive quake struck the same region three months earlier. Some 2,300 victims were injured.

Selected News Excerpts

ABC (05/30/98):
PAKISTAN MAY BE PREPARING FOR ANOTHER TEST
Thousands of people are feared dead following...

ABC (06/01/98):
RESCUE WORKERS CHALLENGED IN AFGHANISTAN
There has been serious death and devastation overseas. In Afghanistan...

CNN (06/02/98):
Food, water, medicine and other supplies have started to arrive. But a U.N. relief coordinator says it's a "scenario from hell".

Figure 3. 200 word summary of simulated IE output, with emphasis on damage

pronoun, and sentences that preceded ones with strongly connecting discourse markers such as *however*, *nevertheless*, etc. We have also softened the constraint on multiple sampling from the same cluster, making use of a redundancy penalty in such

Earthquake strikes quake-devastated villages in northern Afghanistan

A earthquake struck quake-devastated villages in northern Afghanistan Saturday. The earthquake had a magnitude of 6.9 on the Richter scale on the Richter scale.

Damage

Estimates of the death toll varied. CNN (06/02/1998) provided the highest estimate of 4,000 dead, whereas ABC (06/01/1998) gave the lowest estimate of 140 dead.

In capital: Estimates of the number injured varied.

Selected News Excerpts

CNN (06/01/98):
Thousands are dead and thousands more are still missing. Red cross officials say the first priority is the injured. Getting medicine to them is difficult due to the remoteness of the villages affected by the quake.

PRI (06/01/98):
We spoke to the head of the international red cross there, Bob McCaro on a satellite phone link. He says it's difficult to know the full extent of the damage because the region is so remote. There's very little infrastructure.

PRI (06/01/98):
Bob McCaro is the head of the international red cross in the neighboring country of Pakistan. He's been speaking to us from there on the line.

APW (06/02/98):
The United Nations, the Red Cross and other agencies have three borrowed helicopters to deliver medical aid.

Figure 4. 200 word summary of actual IE output, with emphasis on Red Cross

cases. We then perform a randomized local search for a good set of sentences according to these scoring criteria.

2.2.4 Implementation

The Summarizer is implemented using the Apache implementation of XSLT [1] and CoGenTex's Exemplars Framework [13]. The Apache XSLT implementation has provided a convenient way to rapidly develop a prototype implementation of the first two processing stages using a series of XML transformations. In the first step of the third summary generation stage, the text building component of the Exemplars Framework constructs a "rough draft" of the summary text. In this rough draft version, XML markup is used to partially encode the rhetorical, referential, semantic and morpho-syntactic structure of the text. In the second generation step, the Exemplars text polishing component makes use of this markup to trigger surface-

oriented revision rules that smooth the text into a more polished form. A distinguishing feature of our text polishing approach is the use of a bootstrapping tool to partially automate the acquisition of application-specific revision rules from examples.

2.2.5 Sample Summaries

Figures 3 and 4 show two sample summaries that were included in our evaluation (see Section 3 for details). The summary in Figure 3 was generated from simulated output of the IE system, with preference given to damage information; the summary in Figure 4 was generated from the actual output of the current IE system, with preference given to information including the words *Red Cross*.

While the summary in Figure 3 does a reasonable job of reporting the various current estimates of the death toll, the estimates of the death toll shown in Figure 4 are less accurate, because the IE system failed to extract some reports, and the Summarizer failed to correctly merge others. In particular, note that the lowest estimate of 140 dead attributed to ABC is actually a report about the number of school children killed in a particular town. Since no location was given for this estimate by the IE system, the Summarizer's simple heuristic for localized damaged reports — namely, to consider a damage report to be localized if a location is given that is not in the same sentence as the initial disaster description — did not work here. The summary in Figure 3 also suffered from some problems with merging: the inclusion of a paragraph about thousands killed in Afghanistan is due to an incorrect classification of this report as a localized one (owing to an error in sentence boundary detection), and the discussion of the number of villages damaged should have included a report of *at least 80 towns or villages* damaged.

Besides the problems related to slot extraction and merging mentioned above, the summaries shown in Figures 3 and 4 suffer from relatively poor fluency. In particular, the summaries could benefit from better use of descriptive terms from the original articles, as well as better methods of sentence combination and rhetorical structuring. Nevertheless, as will be discussed further in Section 4, we suggest that the summaries show the potential for our techniques to intelligently combine information from many articles on the same natural disaster.

3. EVALUATION AND INITIAL RESULTS

To evaluate the initial version of the IE-supported summarization system, we used Topic 89 from the TDT2 collection — 25 texts on the 1998 Afghanistan earthquake. Each document was annotated manually with the natural disaster scenario templates that comprise the desired output of the IE system. In addition, treebank-style syntactic structure annotations were added automatically using the Charniak parser. Finally, MUC-style noun phrase coreference annotations were supplied manually. All annotations are in XML. The manual and automatic annotations were automatically merged, leading to inaccurate annotation extents in some cases.

Next, the Topic 89 texts were split into a development corpus and a test corpus. The development corpus was used to build the summarization system; the evaluation summaries were generated from the test corpus. We report on three different variants of the RIPTIDES system here: in the first variant (RIPTIDES-SIM1), an earlier version of the Summarizer uses the simulated output of the IE system as its input, including the relief annotations; in the second variant (RIPTIDES-SIM2), the current version of the Summarizer uses the simulated output of the IE system, without the relief annotations; and in the third variant (RIPTIDES-IE), the Summarizer uses the actual output of the IE system as its input.[2]

Summaries generated by the RIPTIDES variants were compared to a Baseline system consisting of a simple, sentence-extraction multidocument summarizer relying only on document position, recency, and word overlap clustering. (As explained in the previous section, we have found that word overlap clustering provides a bare bones way to help determine what information is repeated in multiple articles, thereby indicating importance to the document set as a whole, as well as to help reduce redundancy in the resulting summaries.) In addition, the RIPTIDES and Baseline system summaries were compared against the summaries of two human authors. All of the summaries were graded with respect to content, organization, and readability on an A-F scale by three graduate students, all of whom were unfamiliar with this project. Note that the grades for RIPTIDES-SIM1, the Baseline system, and the two human authors were assigned during a first evaluation in October, 2000, whereas the grades for RIPTIDES-SIM2 and RIPTIDES-IE were assigned by the same graders in an update to this evaluation in April, 2001.

Each system and author was asked to generate four summaries of different lengths and emphases: (1) a 100-word summary of the May 30 and May 31 articles; (2) a 400-word summary of all test articles, emphasizing specific, factual information; (3) a 200-word summary of all test articles, focusing on the damage caused by the quake, and excluding information about relief efforts, and (4) a 200-word summary of all test articles, focusing on the relief efforts, and highlighting the Red Cross's role in these efforts.

The results are shown in Tables 1 and 2. Table 1 provides the overall grade for each system or author averaged across all graders and summaries, where each assigned grade has first been converted to a number (with A=4.0 and F=0.0) and the average converted back to a letter grade. Table 2 shows the mean and standard deviations of the overall, content, organization, and readability scores for the RIPTIDES and the Baseline systems averaged across all graders and summaries. Where the differences vs. the Baseline system are significant according to the t-test, the p-values are shown.

Given the amount of development effort that has gone into the system to date, we were not surprised that the RIPTIDES variants fared poorly when compared against the manually written summaries, with RIPTIDES-SIM2 receiving an average grade of C, vs. A- and B+ for the human authors. Nevertheless, we were pleased to find that RIPTIDES-SIM2 scored a full grade ahead of the Baseline summarizer, which received a D, and that

[2] Note that since the summarizers for the second and third variants did not have access to the relief sentence categorizations, we decided to exclude from their input the two articles (one training, one test) classified by TDT2 Topic 89 as only containing brief mentions of the event of interest, as otherwise they would have no means of excluding the largely irrelevant material in these documents.

Table 1

Baseline	RIPTIDES-SIM1	RIPTIDES-SIM2	RIPTIDES-IE	Person 1	Person 2
D	C/C-	C	D+	A-	B+

Table 2

	Baseline	RIPTIDES-SIM1	RIPTIDES-SIM2	RIPTIDES-IE
Overall	0.96 +/- 0.37	1.86 +/- 0.56 (p=.005)	2.1 +/- 0.59 (p=.005)	1.21 +/- 0.46 (p=.05)
Content	1.44 +/- 1.0	1.78 +/- 0.68	2.2 +/- 0.65 (p=.005)	1.18 +/- 0.6
Organization	0.64 +/- 0.46	2.48 +/- 0.56 (p=.005)	2.08 +/- 0.77 (p=.005)	1.08 +/- 0.65 (p=.05)
Readability	0.75 +/- 0.6	1.58 +/- 0.61 (p=.005)	2.05 +/- 0.65 (p=.005)	1.18 +/- 0.62 (p=.05)

RIPTIDES-IE managed a slightly higher grade of D+, despite the immature state of the IE system. As Table 2 shows, the differences in the overall scores were significant for all three RIPTIDES variants, as were the scores for organization and readability, though not for content in the cases of RIPTIDES-SIM1 and RIPTIDES-IE.

4. RELATED AND ONGOING WORK

The RIPTIDES system is most similar to the SUMMONS system of Radev and McKeown [10], which summarized the results of MUC-4 IE systems in the terrorism domain. As a pioneering effort, the SUMMONS system was the first to suggest the potential of combining IE with NLG in a summarization system, though no evaluation was performed. In comparison to SUMMONS, RIPTIDES appears to be designed to more completely summarize larger input document sets, since it focuses more on finding the most relevant current information, and since it includes extracted sentences to round out the summaries. Another important difference is that SUMMONS sidestepped the problem of comparing reported numbers of varying specificity (e.g. *several thousand* vs. *anywhere from 2000 to 5000* vs. *up to 4000* vs. *5000*), whereas we have implemented rules for doing so. Finally, we have begun to address some of the difficult issues that arise in merging information from multiple documents into a coherent event-oriented view, though considerable challenges remain to be addressed in this area.

The sentence extraction part of the RIPTIDES system is similar to the domain-independent multidocument summarizers of Goldstein et al. [7] and Radev et al. [11] in the way it clusters sentences across documents to help determine which sentences are central to the collection, as well as to reduce redundancy amongst sentences included in the summary. It is simpler than these systems insofar as it does not make use of comparisons to the centroid of the document set. As pointed out in [2], it is difficult in general for multidocument summarizers to produce coherent summaries, since it is less straightforward to rely on the order of sentences in the underlying documents than in the case of single-document summarization. Having also noted this problem, we have focused our efforts in this area on attempting to balance coherence and informativeness in selecting sets of sentences to include in the summary.

In ongoing work, we are investigating techniques for improving merging accuracy and summary fluency in the context of summarizing the more than 150 news articles we have collected from the web about each of the recent earthquakes in Central America and India (January, 2001). We also plan to investigate using tables and hypertext drill-down as a means to help the user verify the accuracy of the summarized information.

By perusing the web collections mentioned above, we can see that trying to manually extricate the latest damage estimates from 150+ news articles from multiple sources on the same natural disaster would be very tedious. Although estimates do usually converge, they often change rapidly at first, and then are gradually dropped from later articles, and thus simply looking at the latest article is not satisfactory. While significant challenges remain, we suggest that our initial system development and evaluation shows that our approach has the potential to accurately summarize damage estimates, as well as identify other key story items using shallower techniques, and thereby help alleviate information overload in specific domains.

5. ACKNOWLEDGMENTS

We thank Daryl McCullough for implementing the coherence boosting randomized local search, and we thank Ted Caldwell, Daryl McCullough, Corien Bakermans, Elizabeth Conrey, Purnima Menon and Betsy Vick for their participation as authors and graders. This work has been partially supported by DARPA TIDES contract no. N66001-00-C-8009.

6. REFERENCES

[1] The Apache XML Project. 2001. “Xalan Java.” http://xml.apache.org/.

[2] Barzilay, R., Elhadad, N. and McKeown, K. 2001. "Sentence Ordering in Multidocument Summarization." In *Proceedings of HLT 2001*.

[3] Bikel, D., Schwartz, R. and Weischedel, R. 1999. "An Algorithm that Learns What's in a Name." *Machine Learning* 34:1-3, 211-231.

[4] Cardie, C. 1997. "Empirical Methods in Information Extraction." *AI Magazine* 18(4): 65-79.

[5] Charniak, E. 1999. "A maximum-entropy-inspired parser." Brown University Technical Report CS99-12.

[6] Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

[7] Goldstein, J., Mittal, V., Carbonell, J. and Kantrowitz, M. 2000. "Multi-document summarization by sentence extraction." In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA.

[8] Grishman, R. 1996. "TIPSTER Architecture Design Document Version 2.2." DARPA, available at http://www.tipster.org/.

[9] Marcus, M., Marcinkiewicz, M. and Santorini, B. 1993. "Building a Large, Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19:2, 313-330.

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

[11] Radev, D. R., Jing, H. and Budzikowska, M. 2000. "Summarization of multiple documents: clustering, sentence extraction, and evaluation." In *Proceedings of the ANLP/NAACL Workshop on Summarization*, Seattle, WA.

[12] Riloff, E. 1996. "Automatically Generating Extraction Patterns from Untagged Text." In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, Portland, OR, 1044-1049. AAAI Press / MIT Press.

[13] White, M. and Caldwell, T. 1998. "EXEMPLARS: A Practical, Extensible Framework for Dynamic Text Generation." In *Proceedings of the Ninth International Workshop on Natural Language Generation*, Niagara-on-the-Lake, Canada, 266-275.