LAMP-TR-113 June 2004
CAR-TR-997
CS-TR-4596
UMIACS-TR-2004-39

USE OF MINIMAL LEXICAL CONCEPTUAL STRUCTURES FOR SINGLE-DOCUMENT SUMMARIZATION

Bonnie J. Dorr, Nizar Y. Habash, and Christof Monz

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
{*bonnie,habash,christof*}*@umiacs.umd.edu*

Abstract

This reports provides an overview of the findings and software that have evolved from the ”Use of Minimal Lexical Conceptual Structures for Single-Document Summarization” project over the last six months. We present the major goals that have been achieved and discuss some of the open issues that we intend to address in the near future. This report also contains some details on the usage of some software that has been implemented during the project.

Keywords: Machine Translation, Document Summarization

The support of this research by the Department of Defense under the contract ITIC extension to NSF ITR Grant IIS-0326553 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2004		**00-06-2004 to 00-06-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Use of Minimal Lexical Conceptual Structures for Single-Document Summarization	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**12**	

1 PARTICIPANTS

PI:

Bonnie Dorr, University of Maryland (UMCP), bonnie@umiacs.umd.edu

OTHER SENIOR PERSONNEL (Ph.D.):

Co-PIs Richard Schwartz, BBN Technologies, schwartz@bbn.com

POSTDOCS:

Nizar Habash, University of Maryland, habash@umiacs.umd.edu

Christof Monz, University of Maryland, christof@umiacs.umd.edu

STUDENTS:

Stacy President, University of Maryland, stacypre@umiacs.umd.edu

Nathaniel Waisbrot, University of Maryland, waisbrot@umiacs.umd.edu

David Zajic, University of Maryland, dmzajic@umiacs.umd.edu

PARTNER ORGANIZATIONS THAT HAVE PROVIDED RESOURCES OR COLLABORATED ON RESEARCH

Richard Schwartz, BBN Technologies

2 COLLABORATIONS (BROADLY CONCEIVED)

Presentations by Bonnie Dorr to Georgetown on the use of linguistic information in hybrid statistical/symbolic tasks (summarization, machine translation, divergence unraveling).

3 PROJECT FINDINGS

1. We have shown the effectiveness of combining sentence compression and topic lists to construct informative summaries.
2. We carried out experiments were three approaches to automatic headline generation (Topiary, Trimmer and Unsupervised Topic Discovery) were compared using two automatic summarization evaluation tools (BLEU and ROUGE).

3. We have stressed the importance of correlating automatic evaluations with human performance of an extrinsic task, and have proposed event tracking as an appropriate task for this purpose.

4. Bonnie Dorr and her student David Zajic (in collaboration with Rich Schwartz at BBN) competed in the Document Understanding Conference (DUC)—a summarization evaluation conducted by NIST. Their headline generator, Topiary, was evaluated automatically using a new metric called Rouge. The Topiary system placed first (out of 40 systems) in the headline (up to 75 characters) summarization task.

5. In the area of single-document (monolingual) summarization, Topiary placed first (out of 40 systems) on 3 Rouge measures and was the only system on this task to score better than a human summary on one measure. On the single-document (cross-lingual) track, Dorr's team placed 2nd on 4 Rouge measures.

6. A preliminary user study where users have to judge the relevance of a document given the full document versus the headline shows using headlines lead to similar precision and recall, but reduce the time it takes to assess the documents by a factor of 4.

4 OPPORTUNITIES FOR TRAINING AND DEVELOPMENT (AT ALL GRADE LEVELS)

The automatically generated headlines allow users to assess the relevance of a document in a time efficient way.

In addition, for cross-lingual headline summarization, it allows user who does not understand the language in which the original document was authored, to assess quickly, whether it relevant enough for being translated by a human translator.

5 PUBLICATIONS AND PRODUCTS

5.1 JOURNAL/CONFERENCE PUBLICATIONS

D. Zajic, B. J. Dorr, and R. Schwartz. BBN/UMD at DUC-2004: Topiary. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Document Understanding*, Boston, MA, 2004.

Dorr, Bonnie J., David Zajic, and Richard Schwartz, "Hedge: A Parse-and-Trim Approach to Headline Generation", *Proceedings of the HLT-NAACL Text Summarization Workshop and Document Understanding Conference (DUC 2003)*, Edmonton, Canada, pp. 1–8, 2003.

Dorr, Bonnie J., Daqing He, Jun Luo, Douglas W. Oard, Richard Schwartz, Jianqiang Wang, and David Zajic, ”iCLEF 2003 at Maryland: Translation Selection and Document Selection”, *Proceedings of the Interactive track for the Cross-Language Evaluation Forum Workshop*, Trondheim, Norway, 2003.

Nizar Habash and Bonnie Dorr, ”CatVar: A Database of Categorial Variations for English”, in *Proceedings of the MT Summit*, New Orleans, LA, pp. 471–474, 2003.

Nizar Habash and Bonnie Dorr, ”A Categorial Variation Database for English”, *Proceedings of North American Association for Computational Linguistics*, Edmonton, Canada, pp. 96–102, 2003.

Nizar Habash. ”Matador: A large scale Spanish-English GHMT system”. *In Proceedings of the MT-Summit*, pages 149 156, 2003.

Habash, Nizar, Bonnie J. Dorr, and David Traum, ”Hybrid Natural Language Generation from Lexical Conceptual Structures”, *Machine Translation*, 18:2, 2003.

5.2 ONE-TIME PUBLICATIONS (INCLUDES BOOK CHAPTERS AND DISSERTATIONS)

David Zajic. *Automatic Generation of Informative Cross-Lingual Headlines for Text and Speech.* Thesis Proposal, University of Maryland, 2003.

5.3 OTHER PRODUCTS

1. Trimmer: Trimmer generates a headline for a news story by compressing the main topic sentence according to a linguistically motivated algorithm. The compression consists of parsing the sentence using the BBN SIFT parser and removing low-content syntactic constituents. Some constituents, such as certain determiners (the, a) and time expressions are always removed, because they rarely occur in human-generated headlines and are low-content in comparison to other constituents.

2. Topiary: Topiary is a modification of the Trimmer algorithm to take a list of topics with relevance scores as additional input. The compression threshold is lowered so that there will be room for the highest scoring topic term that isn’t already in the headline.

3. Generation Heavy Machine Translation system (GHMT). Currently, GHMT supports Spanish to English and Chinese to English translation. In this project, GHMT is adapted in a way that allows cross-lingual summarization.

Online demo of the Spanish-English GHMT system: `http://clipdemos.umiacs.umd.edu/matador/`

Download: `http://clipdemos.umiacs.umd.edu/ghmt/GHMT-PAK.tar.gz`.

Installation:

```
gunzip GHMT-PAK.tar.gz
tar -xf GHMT-PAK.tar
```

documentation:

GHMT: `GHMT-PAK/GHMT/install.readme`

4. depTrimmer. A cross-lingual headline generation extension for GHMT. depTrimmer is fully integrated into GHMT, where translation and sentence compression are applied in tandem. The benefit is that the summarization algorithm is applied to a language independent data structure, which makes it easy to adapt it to a new foreign language. This approach (depTrimmmer) is currently implemented as a prototype for Spanish-English GHMT, and no experimental results are available yet. depTrimmer works on the same data structures that are used within GHMT, viz. normalized dependency trees. The dependency trees are 'trimmed' based on linguistic information including part-of speech, syntactic function, and semantic type. depTrimmer requires the GHMT package, see above. Download: `http://clipdemos.umiacs.umd.edu/deptrimmer/DEPTRIM-PAK.tar.gz`.

   Installation:

   ```
   gunzip DEPTRIM-PAK.tar.gz
   tar -xf DEPTRIM-PAK.tar
   ```

   documentation:

   depTrimmer: `DEPTRIM-PAK/install.readme`

5. CatVar: A Categorial Variation Database for English. CatVar is an extensive is an extensive database of morphological variation for English. CatVar is integrated into GHMT in order to increase the flexibility of the generation of the English translation. Online demo: `http://clipdemos.umiacs.umd.edu/catvar/`

6 CONTRIBUTIONS

6.1 CONTRIBUTIONS WITHIN THE DISCIPLINE

1. An extensive database of morphological variation for English.

2. Robust Machine Translation system from Spanish to English and Chinese to English.

3. A suite of automatic summarization tools (mono-lingual and cross-lingual).

6.2 CONTRIBUTIONS TO OTHER DISCIPLINES (THIS IS NOT EXPECTED FROM ALL PROJECTS)

The project is relevant to the augmentation of capabilities useful for intelligence analysts, such as cross-lingual summarization and data mining.

6.3 CONTRIBUTIONS TO RESOURCES FOR RESEARCH

This work provides an integral part for many NLP applications that require cross-lingual information processing.

6.4 CONTRIBUTIONS BEYOND SCIENCE AND ENGINEERING (THESE CAN BE SPECULATIVE)

The research carried out in this project contributes to the development of cross-lingual information management and processing systems, which facilitates laymen and professionals in accessing information that is authored in a language they do not understand.

7 PLANS FOR THE NEXT YEAR, IF CHANGED

We intend to continue the integration of depTrimmer into Chinese-English and Arabic-English GHMT. Additionally, we plan to evaluate it on the DUC 2004 data sets.

Our funding for this project ends in early 2005. We will need additional funds for the 2 years after the project has expired to continue the high level of activity toward this effort that we have contributed over the last year. For this, we have one proposal currently under review:

> ”Divergence Resolution for Interlingual Variation Encoding (DRIVE)”, REFLEX submission, Broad Agency Announcement (BAA-04-01-FH), May 2004.

8 SPECIAL REPORTING REQUIREMENTS, IF ANY

None.

9 UNOBLIGATED FUNDS (ONLY IF OVER 20%)

N/A

10 SIGNIFICANT CHANGE IN USE OF HUMAN SUBJECTS

None.

A GHMT

REQUIRED RESOURCES

1. LISP: International Allegro CL Enterprise Edition 6.0 (Franz Inc.)
2. Perl: v5.8.0
3. Connexor parser (English and Spanish) (from www.connexor.com) See instructions below on hooking up the connexor client to the rest of the system.
4. Nitrogen Morphology Support
   Nitrogen is available at: http://www.isi.edu/natural-language/projects/nitrogen/
   Specifically, the morphology files, nitro.english.morph.lisp nitro.morph.8.98.lisp nitromorph-8-98.lisp must be placed under $PACKAGE/EXERGE/SOURCE/oxyexerge/
5. Halogen Forest Ranker
   Halogen is available at: http://www.isi.edu/licensed-sw/halogen/ All code from the forest ranker should be installed under $PACKAGE/HALOGEN/ForestRanker

Make sure the variables in sysVars.cshrc are added to your .cshrc

The source files for the Exerge system are included in this package in addition to created images on Solaris. to remake these images, run $PACKAGE/ake-Exerge.sh

See a Sample run of Matador below.

CONNEXOR SPECIFIC INSTRUCTIONS

1. Contact www.connexor.com to obtain a license for English and Spanish parsers.

2. Update the host/port in the files fdges-client.pl (for Spanish) and fdgen-client.pl (for English). The current values should look like this for fdges-client.pl:
   $remote_host="cheesecake.umiacs.umd.edu"
   $remote_port="11720"
   and as follows for fdgen-client.pl
   $remote_host="cheesecake.umiacs.umd.edu"
   $remote_port="11721"

SAMPLE RUN

```
> matador.pl test out2 x params=params.matador.2
parameter params = params.matador.2  ... loading
Processing Batch #0
PARSING...
TRANSLATING...
reading /fs/clip-plus/habash/PACKAGE/TRANSLEX/Spanish-English/span-eng.tralex ... done
translating torotemp.dep ...
done
CONVERTING,EXPANDING...
/fs/clip-plus/habash/PACKAGE/EXERGE/corexerge.sh torotemp.trans.amr torotemp.out.amr
T NIL T T T 10 10 10 10 NIL 1 T NIL T
<Running CorExerge>
; Exiting Lisp
LINEARIZAING...

<Running OxyGen 2.0>
; Exiting Lisp
RANKING...
/fs/clip-plus/habash/PACKAGE/EXERGE/halogenize torotemp.out.gls torotemp.out.txt 6
/fs/clip-plus/habash/PACKAGE/HALOGEN/ForestRanker/news.binlm
```

```
<GLS-to-Forest Conversion> && <Running  HALOGEN>

/fs/clip-plus/habash/HALOGEN/ForestRanker/polishsen.pl
/fs/clip-plus/habash/PACKAGE/MATADOR/halolin-temp.sen0
> /fs/clip-plus/habash/PACKAGE/MATADOR/halolin-temp.sen
; cpu time (non-gc) 420 msec user, 10 msec system
; cpu time (gc)     70 msec user, 0 msec system
; cpu time (total)  490 msec user, 10 msec system
; real time  23,139 msec
; space allocation:
;  332,622 cons cells, 7,882,064 other bytes, 0 static bytes; Exiting Lisp
REPORTING...
done!
```

B DepTrimmer

REQUIRED RESOURCES

1. GHMT System (specifically Matador installation)
2. The source files for DepTrimmer are included in this package in addition to created images on Solaris. to remake these images, goto $DEPTRIM-PAK/DEPTRIMMER/SOURCE run make

See a Sample run of DEPTrimmer below.

depTrimmer takes an AMR tree, removes parts of the sentence until the sentence length is below some threshold, then outputs the trimmed AMR tree.

The trimming algorithm:

1. Delete all determiners
2. Delete all punctuation
3. Delete all time expressions

4. Delete some conjunctions
5. Delete some relative clauses
6. If the sentence is too long, delete all conjunctions
7. If the sentence is too long, delete all relative clauses
8. While the sentence is too long, delete prepositional phrases which do not contain a proper noun
9. While the sentence is too long, delete all prepositional phrases
10. Clean up any dangling connectives

Note that steps 1-5 take place *even if the sentence is already below the threshold.*

More detailed explanation of steps:

1. Delete all determiners Determiners aren't generally needed for comprehension. Most real headlines don't have them. We delete anything tagged as 'D' (determiner).
2. Delete all punctuation Punctuation isn't very important for comprehension. The sophisticated use of punctuation that real headlines use is quite difficult. We delete anything tagged as 'PX' (punctuation)
3. Delete all time expressions Time expressions are generally superfluous. Relative expressions, like "today" are meaningless after that day has passed. When specific dates appear, they generally include the event that takes place on that date, which is more useful to keep. E.g. in "the November elections", 'November' is not as important as 'elections'. There may be specific time expressions which should be excluded from this, e.g. "the September 11th investigation committee". We delete anything tagged as 'TIME' (time expressions)
4. Delete some conjunctions In phrases like "he ran away and hid his face" or "the President and the Vice President", the subordinate phrase is generally less important. If any phrase (noun, verb, or prepositional) is connected to a phrase of the same type by a conjunction, the subordinate phrase is deleted. Optionally, we delete only phrases which do not contain a proper noun.
5. Delete some relative clauses In phrases like "actions which would bring about changes", the head of the sub-phrase "which would bring about changes" is the verb "bring". Verb phrases which are direct children of noun phrases are generally less important, and we delete them. Optionally, we delete only phrases which do not contain a proper noun.

6. If the sentence is too long, delete all conjunctions In step 4, we had the option of leaving phrases containing a proper noun intact. If we did so, and the sentence is too long, we now delete all these phrases.

7. If the sentence is too long, delete all relative clauses In step 5, we had the option of leaving phrases containing a proper noun intact. If we did so, and the sentence is too long, we now delete all these phrases.

8. While the sentence is too long, delete prepositional phrases which do not contain a proper noun We assume that the deepest prepositional phrase is the least important. E.g. in ”The prince of the smallest country in the world”, ”in the world” is probably the least important part of the phrase. Therefore, while the sentence is too long, we find the deepest prepositional phrase which does not contain a proper noun, and delete it.

9. While the sentence is too long, delete all prepositional phrases If the sentence is still too long, we repeat step 8, except that a phrase is deleted regardless of whether or not it contains a proper noun.

10. Clean up any dangling connectives The deletion process leaves some connectives dangling. Here, we delete any connective which doesn’t have siblings.

SAMPLE RUN

```
> more test
<doc doc_id="UN-20.100-23" sys_id="SRC-SP">
<segment>
Este ltimo misilpuede equiparse con las ojivas nucleares que se estn produciendo en Israel.
</segment>
</doc>

> depTrimmer+matador.pl test out TEST params=params.deptrimmer.1
parameter params = params.deptrimmer.1  ... loading
torotemp.*: No such file or directory
Processing Batch #0
PARSING...
TRANSLATING...
reading
/fs/clip-plus/habash/PACKAGE/GHMT-PAK/GHMT/TRANSLEX/Spanish-English/span-eng.tralex ... done
translating torotemp.dep ...
```

```
done
CONVERTING,EXPANDING...
<Running CorExerge>
; Exiting Lisp
<Running depTrimmer>
; Exiting Lisp
LINEARIZAING...
<Running OxyGen 2.0>
; Exiting Lisp
RANKING...
/fs/clip-plus/habash/PACKAGE/GHMT-PAK/GHMT/EXERGE/halogenize torotemp.out.gls
torotemp.out.txt 5 /fs/clip-plus/habash/PACKAGE/GHMT-PAK/GHMT/MATADOR/un500k.binlm

<GLS-to-Forest Conversion> && <Running  HALOGEN>

/fs/cliplab/muri/habash/HALOGEN/ForestRanker/polishsen.pl /tmp/halolin-temp.sen0
> /tmp/halolin-temp.sen
; cpu time (non-gc) 180 msec user, 0 msec system
; cpu time (gc)     50 msec user, 0 msec system
; cpu time (total)  230 msec user, 0 msec system
; real time  824 msec
; space allocation:
;  132,271 cons cells, 3,129,408 other bytes, 0 static bytes; Exiting Lisp
REPORTING...
done!

> more out
<doc doc_id="UN-20.100-23" sys_id="TEST">
<segment>
Last missile can be equipped with nuclear warheads
</segment>
</doc>
```