Summaries with SumUM and its Expansion for Document Understanding Conference (DUC 2002)

Atefeh Farzindar and **Guy Lapalme**
Département d´Informatique
et Recherche Opérationnelle
RALI / Université de Montréal
{Farzinda,Lapalme}@iro.umontreal.ca

Horacio Saggion
Department of Computer Science
Natural Language Processing Group
University of Sheffield
H.Saggion@dcs.shef.ac.uk

Abstract

We present the results of the DUC 2002 evaluation on an adaptation of SumUM which had previously been used for producing abstracts of scientific and technical papers. Although we did not modify the original manually developed algorithms or dictionary of concepts for this evaluation, the results are among the best of all systems that participated in DUC 2002 which was on a completely different domain.

1 Introduction

SumUM is a text summarization system developed by Horacio Saggion (2002), which produces short automatic abstracts of long scientific and technical documents. Abstracts are produced in two steps: the reader is presented with an indicative abstract, which identifies the topics of the document. After that, an informative abstract is presented that elaborates some topics selected by the reader. The implementation is based on shallow syntactic and semantic analysis, conceptual identification and text re-generation. For a complete description of the system the reader is referred to (Saggion and Lapalme, 2002).

For participation in DUC 2002, we modified SumUM without changing the algorithm or templates (developed manually from the study of a corpus of abstracts written by professional abstractors). We did not change the algorithm, because we wanted to evaluate the performance of the actual system before modifying it in our future investigations.

2 Description of the system

SumUM is a dynamic summarization system for producing the indicative-informative summaries of the long technical documents. The architecture of the system is shown in Figure 1.

This system is implemented in SICStus Prolog (Release 3.7.1) and Perl running on sun workstation (5.6) and Linux machines (RH 6.0).

The implementation of our method relies on the following: the section of particular types of information from the source text; the instantiation of different types of templates; the selection templates in order to produce an indicative abstract; the regeneration of a short but novel text which indicates the topics of the document and; the expansion of the indicative text with topic elaboration.

The input to the system is a technical article in English, a single-document without any markup, but with the following structural elements: Title of the article, Author information and affiliation, Introduction section, Main section without prototypical structure, References, and optionally, Author abstract, list of Keywords, and Acknowledgements.

2.1 Pre-processing and Interpretation

The raw text is tagged and transformed in a structured representation, allowing the following processes to access the structure of the text (words, groups of words, titles, sentences, paragraphs, sections, and so on). Domain specific transducers are applied in order to identify possible concepts in the discourse domain (such as the authors, the paper, references to other authors, institutions and so on) and linguistic transducers are applied in order to

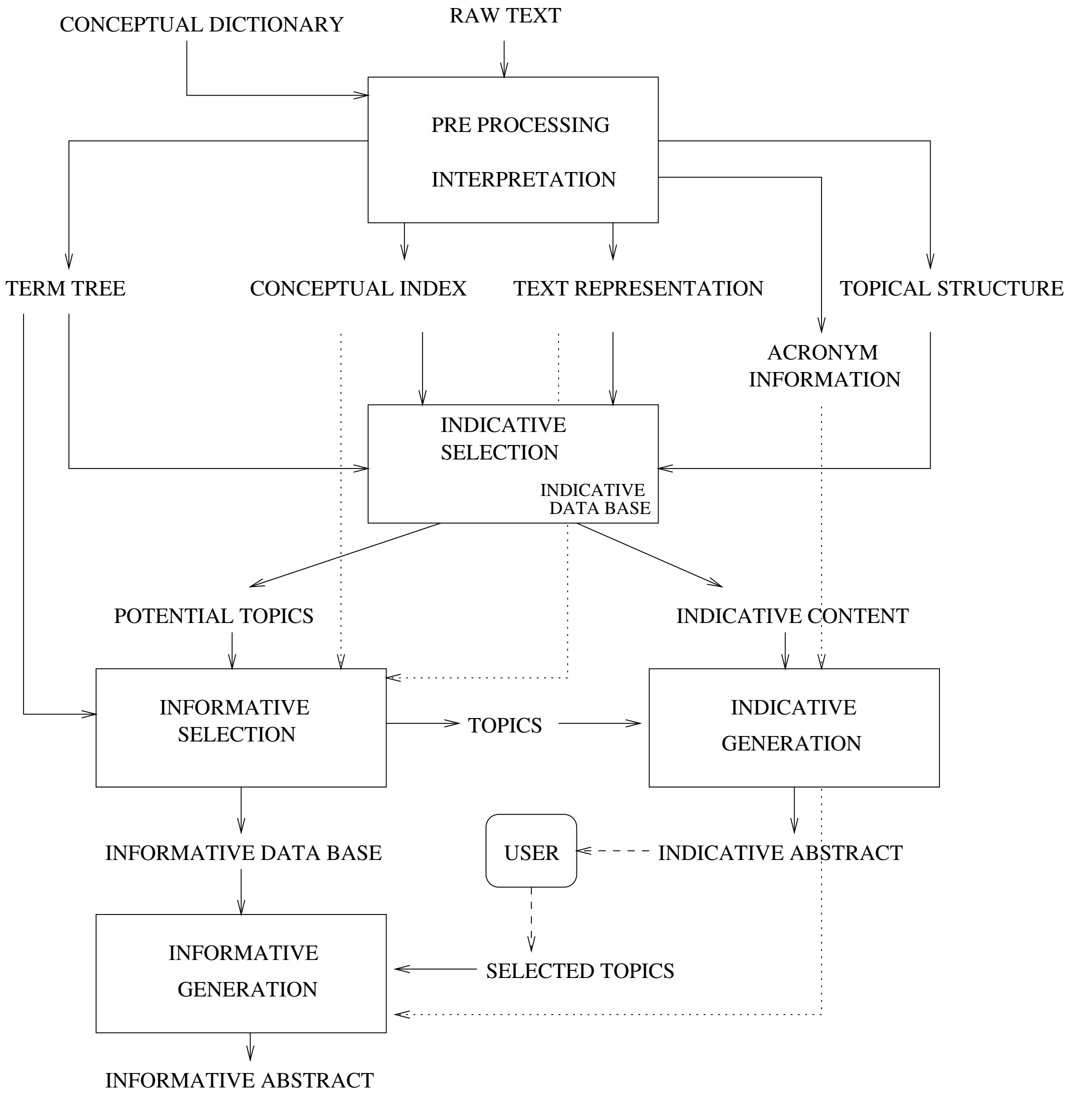

Figure 1: SumUM Architecture

identify noun groups and verb groups. Afterwards, semantic tags, marking discourse domain relations and concepts are added to the different elements of the structure.

2.2 Indicative Selection

Its function is to identify sentences of indicative type for constructing the content for the indicative abstract, which extracts the relevant information of the text and identifies potential topics of the document.

2.3 Informative Selection

This process determines the sentences of informative type. It elaborates the potential topics computed by the indicative selection according to the interest of the reader.

2.4 Generation

The output is an indicative abstract composed of complete re-generated sentences. The text is presented to the reader, with a list of topics available for expansion, which the reader can decide to elaborate.

3 Evaluation of SumUM in DUC 2002

We participated with SumUM in the DUC 2002 ”fully automatic summarization of a single document” track. SumUM produced abstracts 100 or less words length.

The National Institute of Standards and Technology of the U.S. (NIST) produced 60 reference sets of approximately 10 documents. The document sets are produced using data from the TREC disks used in the question-answering track in TREC-9. Each set contained the short documents defined by four different types of criteria, such as single natural disaster event, single event in any domain, multiple distinct events of a single type, and biographical information mainly about a single individual, but we did not have any long scientific documents.

Nevertheless, SumUM is designed for the technical articles and the sources of information that we used for implementing our system are based of the linguistic and conceptual pattern of technical documents. Obviously, the linguistic information, in the DUC corpus is different from information usually found in the technical domain. But before we consider any changes, we needed to test the general quality of the actual system in DUC 2002.

4 Expansion of SumUM for DUC 2002

In order to produce the abstract with SumUM for DUC 2002, we redesigned the input to our system. We had to format the source texts in order to be able to recognize the structure of the document. Then we changed the interactive mode of the system and its interface. In the first version of SumUM, in each step of pre-processing, interpretation, indicative selection, informative selection and generation, the user must guide the system and choose an option proposed by the system. When SumUM produces an indicative abstract and presents the potential topics, the user has to decide to continue elaborating the topics, or stop to exit the system. But in expanded version of SumUM all the processes are automatic: it takes a document set of DUC 2002 as input and it produces the abstracts of all included files in the set, as output.

For each set of 10 documents, we execute a single program allowing the system to do the following processes automatically and save the result with the submission requested in DUC 2002:

Input processing:

We have a Perl program for taking the single-document of DUC, removing the mark-up and adapting it with input structure recognizable by SumUM.

Pre-processing and Interpretation:

This process identifies the structure of the document and interprets the sentences of the text.

The main process of SumUM:

SumUM generates a short abstract of the text. For the DUC tests, the system is not interactive and will automatically abstract and elaborate the topics, without user intervention.

Output abstract:

A Perl program takes the abstract, generated by SumUM and adapts it, with the submission requested for DUC 2002.

5 Evaluation Results

For single document summaries there were two categories of evaluation, one done by humans (for the abstracts), and one done automatically (for the extracts) by NIST.

To compute the quantitative measure between the generated summaries by the summarization systems (Peer abstract) and the reference (Model abstract), the human-created summary was segmented by hand into model units (MUs), and the generated summaries by the summarization systems, were segmented into peer units (PUs), which are always sentences.

To evaluate the quality of the system-generated summaries, DUC 2002 used the twelve qualitative questions. The quality questions concern aspects like grammaticality, coherence and organization of abstracts. For each document set, the evaluator reads the peer summary and then makes overall judgments as to the peer summary's quality, independent of the model. Table 1 lists the results obtained for the evaluation of 13 systems, which participated in the Single-Document Summarization task at DUC 2002. In the table, Human denotes the result of human subject and Baseline is the result of the baseline which merely takes the first 100 words in the document. The lower score for the mean score and count of quality questions means the better performance.

SumUM was ranked first according to the mean score for questions about the quality of the produced output. This shows the interest of having Su-

Rank	Group	Mean Error Quality Questions	Group	Count of Quality Questions	Group	Mean Length-Adjusted Coverage
1	**Our system**	**0.408**	Microsoft	0.582	BBN	0.339
2	Microsoft	0.425	LCC	0.698	**Our system**	**0.299**
3	LCC	0.448	**Our system**	**0.758**	LCC	0.293
4	CCS-NSA	0.537	U.Nijmegen	0.885	Microsoft	0.272
5	U.Ottawa	0.551	U.Ottawa	0.986	NTT	0.272
6	NTT	0.552	Imperial College	0.997	CCS-NSA	0.261
7	U.Nijmegen	0.561	CCS-NSA	1.013	U.Leuven	0.251
8	Imperial College	0.565	NTT	1.014	U.Nijmegen	0.247
9	U.Michigan	0.644	U.Lethbridge	1.153	U.Lethbridge	0.240
10	U.Leuven	0.660	U.Leuven	1.210	U.Ottawa	0.232
11	U.Lethbridge	0.676	U.Michigan	1.441	Imperial College	0.228
12	BBN	1.040	BBN	2.637	ISI/GLEANS	0.220
13	ISI/GLEANS	1.281	ISI/GLEANS	3.200	U.Michigan	0.214
Human		0.354		0.505		0.336
Baseline		0.490		0.718		0.255

Table 1: Evaluation Results on Single-Document

mUM regenerating well-formed sentences instead of putting together arbitrary parts of text extracts.

Coverage metric measures how much of a model summary's content was expressed by a system-generated peer summary. For DUC 2002 there was a desire also to:

1) Look at the ability of a system to produce a summary shorter than the predefined target length

2) Devise a combined measure of coverage and compression.

The higher score for the mean coverage means the better performance. For the mean length-adjusted coverage, our system was ranked on the second place. We think the reason for the good performance of SumUM was that the templates were well designed, which regenerates the short complete sentences.

Table 2 lists the results of the evaluations of SumUM in DUC 2002 which the summaries responded the twelve questions and the rank of our summarization in single-document between 13 other systems. Table 3 lists also the experimental results we obtained in this competition.

6 Conclusions

We presented SumUM a text summarization system, which participated in Single-Document Summarization task, and the benefit of our experience in DUC 2002 as its expansion. Our system was specified for summarization of the long scientific and technical document and very different from the corpus of DUC, because the test corpus was very short for SumUM with the different linguistic information from the scientific paper. Unfortunately, in DUC 2002 we ended up with many empty outputs, where the patterns and templates of SumUM could not match with the new types of information of non-scientific documents. We just evaluated the automatic regenerated abstracts, which produced by the templates of system, without replacing the empty outputs with the base line abstract. This is the reason why we obtained the low scores in precision and recall. However, the automatic generated abstracts by SumUM were considered comparable to other systems summarization that participated in DUC 2002, which has been developed especially for this kind of data. The results so far indicate good general quality of system when compared with other summarization technologies.

We will continue to enhance our system and widen the semantic analysis and conceptual identification in order to develop the new patterns and templates for different types of document, in the text re-generation.

Acknowledgement

The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R D program.

Twelve questions used in DUC 2002	Evaluation Result	Rank of SumUM
Q1: Capitalization errors	0.06	1
Q2: Incorrect Word Order	0.03	7
Q3: Subject-verb agreement	0.00	1
Q4: Missing Component	0.12	5
Q5: Unrelated fragments joint	0.05	5
Q6: Missing articles	0.02	8
Q7: Pronouns lacking antecedents	0.06	5
Q8: Nouns with unclear referents	0.22	10
Q9: Noun/Noun phrase- pronoun replacement	0.01	9
Q10: Out of place conjunctions	0.02	7
Q11: Unnecessary information repetition	0.04	2
Q12: Sentences wrong place	0.01	3

Table 2: Result of the evaluation of SumUM, which the summaries responded the qualitative questions and the rank of our summarization in single-document between 13 other systems.

Evaluation	Result	Rank of SumUM
Score for quality questions with non-0 answe	0.31	2
Count of quality questions with non-0 answers	0.76	3
Mean Score for quality questions with non-0 answers	0.41	1
Unmodified Mean coverage	0.08	12
Mean length-adjusted coverage	0.30	2

Table 3: Evaluation Results of SumUM.

References

R. Barzilay and M. Elhadad. 1997. Using lexical chains for text summarization. *In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL.* Madrid, Spain.

E. Hovy and C. Lin. 1997. Automated text summarization in SUMMARIST. In Mani and Maybury (1997), 18–24. Cambridge:MIT Press.

Kevin, K. and D. Marcu, 2000. Statistics-Based Summarization - Step One: Sentence Compression, *The 17th National Conference of the American Association for Artificial Intelligence, AAAI'2000, Outstanding Paper Award*, pages = "703-710", Austin, Texas, 2000, July 30-Augest 3.

C. Lin. 2000. The automated acquisition of topic signatures for text summarization. *In Proceedings of the COLING Conference*, 2000, Saarbrcken, Germany.

C. Lin. 2001. SEE - Summary Evaluation Environment, "http://www.isi.edu/ cyl/SEE/",

Dragomir R. Radev and Kathleen McKeown. 1998. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 24(3):469–500. September 1998.

H. Saggion and G. Lapalme. 2002. Generating informative and indicative summaries with SumUM, to appear in *Computational Linguistics* , Special Issue on Automatic Summarization.