A Flexible Multitask Summarizer for Documents from Different Media, Domain, and Language

Author

Maria Fuentes Fort

Advisor

Horacio Rodríguez Hontoria

Tesi Doctoral presentada al
Departament de Llenguatges i Sistemes Informàtics
de la Universitat Politècnica de Catalunya
per optar al grau de doctor en Intel·ligència Artificial

Març de 2008

A en Tristan i a la seva gran família

Abstract

Automatic Summarization is probably crucial with the increase of document generation. Particularly when retrieving, managing and processing information have become decisive tasks. However, one should not expect perfect systems able to substitute human summaries. The automatic summarization process strongly depends not only on the characteristics of the documents, but also on user different needs. Thus, several aspects have to be taken into account when designing an information system for summarizing, because, depending on the characteristics of the input documents and the desired results, several techniques can be applied. In order to support this process, the final goal of the thesis is to provide a flexible multitask summarizer architecture. This goal is decomposed in three main research purposes. First, to study the process of porting systems to different summarization tasks, processing documents in different languages, domains or media with the aim of designing a generic architecture to permit the easy addition of new tasks by reusing existent tools. Second, to develop prototypes for some tasks involving aspects related with the language, the media and the domain of the document or documents to be summarized as well as aspects related with the summary content: generic, novelty summaries, or summaries that give answer to a specific user need. Third, to create an evaluation framework to analyze the performance of several approaches in written news and scientific oral presentation domains, focusing mainly in its intrinsic evaluation.

Resumen

El resumen automático probablemente sea crucial en un momento en que la gran cantidad de documentos generados diariamente hace que recuperar, tratar y asimilar la información que contienen se haya convertido en una ardua y a su vez decisiva tarea. A pesar de ello, no podemos esperar que los resúmenes producidos de forma automática vayan a ser capaces de sustituir a los humanos. El proceso de resumen automático no sólo depende de las características propias de los documentos a ser resumidos, sino que es fuertemente dependiente de las necesidades específicas de los usuarios. Por ello, el diseño de un sistema de información para resumen conlleva tener en cuenta varios aspectos. En función de las características de los documentos de entrada y de los resultados deseados es posible aplicar distintas técnicas. Por esta razón surge la necesidad de diseñar una arquitectura flexible que permita la implementación de múltiples tareas de resumen. Este es el objetivo final de la tesis que presento dividido en tres subtemas de investigación. En primer lugar, estudiar el proceso de adaptabilidad de sistemas a diferentes tareas de resumen, como son procesar documentos producidos en diferentes lenguas, dominios y medios (sonido y texto), con la voluntad de diseñar una arquitectura genérica que permita la fácil incorporación de nuevas tareas a través de reutilizar herramientas existentes. En segundo lugar, desarrollar prototipos para distintas tareas, teniendo en cuenta aspectos relacionados con la lengua, el dominio y el medio del documento o conjunto de documentos que requieren ser resumidos, así como aspectos relacionados con el contenido final del resumen: genérico, novedad o resumen que de respuesta a una necesidad especifica. En tercer lugar, crear un marco de evaluación que permita analizar la competencia intrínseca de distintos prototipos al resumir noticias escritas y presentaciones científicas orales.

Acknowledgements

I am most indebted and thankful to my supervisor, Horacio Rodríguez, who very generously shared his knowledge and time with me from the very first to the last day.

I want to thank Xavier Pueyo, head of the group Geometria i Gràfics at the Universitat de Girona, for supporting my research during my stay in the Department of Informàtica i Matemàtica Aplicada where everybody encouraged me constantly. I will always be very thankful to Marc Massot for making me believe in my aptitudes. I also want to express my gratitude to all those colleagues who, in a completely generous way, participated in the creation of the Spanish news summarization corpus for the HERMES project (funded by the Spanish Ministry of Science and Technology, TIC2000-0335-C03-02). It was in that project that I started collaborating with Laura Alonso, to whom I thank her encouragement and help I received from the very beginning, and who often compensated my lack of knowledge on more linguistic-based issues.

Xavi Carreras, Lluís Padró, Lluís Màrquez and German Rigau made it easy for me to be a telecommuting member of the group Processament del Llenguatge Natural, at the Universitat Politècnica de Catalunya, which is a consolidated research group according to the Comissió Interdepartamental de Recerca i Innovació Tecnològica, 2001SGR 00254.

I am very thankful to Jordi Turmo for inviting me to participate in the CHIL project (funded by the European Union Commission, IP 506909), thus allowing me to work on the new challenges targeted at the TALP research center, together with researchers such as Mihai Surdeanu, Dusan Macho, Joachim Neumann, Josep Ramon Casas and Climent Nadeu. I am deeply grateful to all of them for their hospitality and assistance. Moreover, I feel very lucky for the remarkable technical and scientific support I have received from Edgar Gonzàlez. Regarding the corpora used within this project, I am thankful to the ISL at the Karlsruhe University, which provided us with transcriptions through Matthias Wölfel, to the ELDA, which generated the reference summaries, and to all the volunteers in the Teoria del Senyal i Comunicacions and Llenguatges i Sistemes Informàtics departments at the Universitat Politècnica de Catalunya, who assisted in the automatic summaries evaluation.

My gratitude towards the people in the group Processament del Llenguatge i Sistemes d'Informació at Universitat d'Alacant. Mediating Yenory Rojas, they made it possible for me

to evaluate different Information Retrieval systems. I am as well indebted to Daniel Ferrès for the effort he put in adapting the modules of the TALP-QA Question Answering system. These experiments were partially funded by the ALIADO project (TIC2002-04447).

One of the articles included as part of this dissertation stemmed from my talks with Enrique Alfonseca. I will be always grateful to him for his availability and the interest he showed in my work lending me the UAM-Titech06 system and adapting it to use Machine Learning techniques.

My work has benefited from source components available online. Some of the used tools are: the Automatic Text Summarizer SweSum, the set of evaluation tools for the MEAD project, the chunk annotator YamCha, and the Information Retrieval system JIRS.

My research group and S107 office colleagues, Edgar Gonzàlez, Meritxell Gonzàlez, Daniel Ferrès, Pere Comas, Eli Comelles, Jordi Atseries, Lluís Villarejo, Francis Real, Manu Bertran, Jesus Giménez, Montse Cuadros, Roberto Asín, Muntsa Padró, Jordi Poveda, Emili Sapena, and Sergi Fernández, deserve a special mention. Their friendliness and drive make everyday life easier to deal with.

My work has also profited from appropriate comments and suggestions from Ani Nenkova, René Witte, Gemma Grau, Pere Pau Vázquez, Roser Saurí, Roser Morante, Astrid Magrans, as well as all those people that I have meet in my research path at work or when attending courses, conferences or project meetings.

I am very thankful to Isidoro Gil Leiva, professor of the master in Documentation of the Universitat Politècnica de València, and to the members of GLiCom, the Computational Linguistics group of the Universitat Pompeu Fabra for their interest in my work.

The comments provided by anonymous reviewers who read a previous version of my dissertation have been extremely helpful. They have been very useful and, in some cases, very encouraging.

I express my gratitude to each of the Committee members, David Farwell, Mihai Surdeanu, Irene Castellon, Manuel de Buenaga and Dragomir Radev, for their availability to participate in the examination panel of the dissertation I present here.

Finally, I would like to say that I feel fortunate because of the unconditional support that I received from my family and friends, despite the fact that "the thesis" had maximum priority during these past years. My mother has always been there for me, enjoying every finding and comforting me at every setback. I found in Christophe's love and understanding the strength that enabled me to finish this dissertation.

Contents

List of Figures

List of Tables

List of Acronyms and other Abbreviations

AI Artificial Intelligence

AS Automatic Summarization

ASR Automatic Speech Recognizer

BE Basic Element

CE Content Extractor

CHIL Computers in the Human Interaction Loop

CRF Conditional Random Fields

CLEF Cross-Language Evaluation Forum

CS Candidates Selector

DM Discourse Marker

DT Decision Tree

DUC Document Understanding Conference

ELDA Evaluation and Language resources Distribution Agency

EWN EuroWordNet

FE Feature Extractor

FEMsum Flexible Eclectic Multitask summarizer

GALE Global Autonomous Language Environment

HERMES Hemerotecas Electrónicas, Recuperación Multilingüe y Extracción Semántica

HG Headline Generation

HMM Hidden Markov Models

HTTP HyperText Transfer Protocol

IDF Inverse Document Frequency

IE Information Extraction

ILP Inductive Logic Programing

IR Information Retrieval

ISL Interactive Systems Laboratories

JIRS Java Information Retrieval System

LAC Length-Adjusted Coverage

LC Lexical Chain

LP Linguistic Processor

LSA Latent Semantic Analysis

MC Mapping-Convergence

MDS Multi-Document Summarization

ML Machine Learning

MMR Maximal Marginal Relevance

MRL Mutual Reinforcement Learning

MSE Multilingual Summarization Evaluation

MT Machine Translation

MUC Message Understanding Conferences

MW Multi-Word

NCM Noisy Channel Model

NE Named Entity

NEC Named Entity Classification

NER Named Entity Recognition

NERC Named Entity Recognition and Classification

NIST National Institute of Standards and Technology

NL Natural Language

NLG Natural Language Generation

NLP Natural Language Processing

NP Noun Phrase

NTCIR NII Test Collection for IR Systems

OCR Optical Character Recognizer

OCSVM One-Class Support Vector Machine

OPP Optimum Position Policy

POS Part Of Speech

PR Passage Retrieval

QA Question & Answering

QP Query Processor

RBF Radial Basis Function

RID Relevant Information Detector

ROUGE Recall-Oriented Understudy for Gisting Evaluation

SC Summary Composer

SCU Summary Content Unit

SDS Single Document Summarization

SEE Summary Evaluation Environment

SMG Similarity Matrix Generator

SS Speech Summarization

SVD Singular Value Decomposition

SVM Support Vector Machine

TAC Text Analysis Conference

TE Textual Entailment

TED Translingual English Database

TF Term Frequency

TFIDF Term Frequency Inverse Document Frequency

TREC Text REtrieval Conference

TS Text Summarization

TU Textual Unit

UPC Universitat Politècnica de Catalunya

UTD Unsupervised Topic Discovery

WER Word Error Rate

WN WordNet

WSD Word Sense Disambiguation