

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 específica. 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 González. 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

<b>Abstract</b>	<b>iii</b>
<b>Resumen</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>List of Acronyms and other Abbreviations</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem to be solved . . . . .	1
1.1.1 Automatic Summarization Aspects . . . . .	2
1.1.2 Automatic Summarization Evaluation . . . . .	6
1.2 Aim of the Thesis . . . . .	11
1.3 Research Method . . . . .	12
1.4 Document Outline . . . . .	12
<b>2 State of the Art</b>	<b>15</b>
2.1 Automatic Summarization from a classical perspective . . . . .	17
2.1.1 Surface level . . . . .	18
2.1.2 Entity level . . . . .	19
2.1.3 Discourse level . . . . .	21
2.1.4 Combined Systems . . . . .	22
2.2 Automatic Summarization from a Multitask perspective . . . . .	22

---

2.2.1	DUC Summarization tasks . . . . .	23
2.2.2	Automatic Summarization Multitask classification . . . . .	24
2.3	Automatic Summarization Evaluation . . . . .	35
2.3.1	Multilingual Summarization Evaluation . . . . .	44
2.3.2	Speech Summarization Evaluation . . . . .	45
2.3.3	Query-Focused Summarization Evaluation . . . . .	45
2.3.4	International Evaluation Future . . . . .	45
<b>3</b>	<b>Overall FEMsum Architecture</b>	<b>47</b>
3.1	Global Architecture . . . . .	48
3.2	Component Architecture Design Considerations . . . . .	50
3.3	Components used by the instantiated FEMsum approaches . . . . .	54
3.4	Linguistic Processor Component . . . . .	59
3.4.1	Catalan and Spanish LP . . . . .	62
3.4.2	English LP . . . . .	63
3.5	Query Processor Component . . . . .	64
3.5.1	Splitting QP . . . . .	65
3.5.2	Generating QP . . . . .	65
3.6	Relevant Information Detector Component . . . . .	66
3.6.1	Lexical Chain based RID . . . . .	68
3.6.2	Feature Extractor RID . . . . .	72
3.6.3	Passage Retrieval RID . . . . .	76
3.7	Content Extractor Component . . . . .	78
3.7.1	Lexical Chain based CE . . . . .	80
3.7.2	Machine Learning based CE . . . . .	80
3.7.3	Question & Answering System based CE . . . . .	82
3.7.4	Semantic information based CE . . . . .	83
3.8	Summary Composer Component . . . . .	87
3.8.1	Complex SC . . . . .	87

---

3.9	Implementation FEMsum framework . . . . .	93
3.9.1	Integration from Source Code . . . . .	94
3.9.2	Integration from Binaries . . . . .	97
<b>4</b>	<b>Multilingual FEMsum</b>	<b>99</b>
4.1	Agency news multilingual Single Document Summarization . . . . .	100
4.1.1	Multilingual evaluation framework . . . . .	100
4.1.2	Exploiting Textual Properties . . . . .	102
4.2	News Stories Headline Extraction . . . . .	114
4.2.1	Re-using DUC data for evaluation . . . . .	115
4.2.2	A mixed approach to Headline Extraction for DUC task . . . . .	116
4.3	Conclusions . . . . .	121
<b>5</b>	<b>Spontaneous Speech FEMsum</b>	<b>123</b>
5.1	Spontaneous speech evaluation framework . . . . .	124
5.1.1	TED manual transcript gold corpus . . . . .	124
5.1.2	JANUS-ViaVoice automatic transcript gold corpus . . . . .	125
5.1.3	JANUS automatic transcript gold corpus . . . . .	127
5.1.4	Evaluation method . . . . .	129
5.2	Original textual FEMsum instantiations . . . . .	129
5.2.1	Summarizing manual transcripts . . . . .	130
5.2.2	Summarizing automatic transcripts . . . . .	134
5.2.3	Analysis of the results . . . . .	137
5.3	Exploiting general textual properties . . . . .	138
5.3.1	Analysis of the results . . . . .	142
5.4	Conclusions . . . . .	144
<b>6</b>	<b>Query-focused FEMsum</b>	<b>145</b>
6.1	Query-focused Multi-Document Summarization DUC task . . . . .	146
6.1.1	AutoPan, a pyramid-based automatic evaluation method . . . . .	147

---

6.1.2	Using the TALP-QA system at DUC 2005 . . . . .	150
6.1.3	Comparing several FEMsum approaches at DUC 2006 . . . . .	155
6.1.4	SEMsum and SVM based approaches evaluated with autoPan . . . . .	160
6.1.5	SEMsum at DUC 2007 . . . . .	165
6.2	Query focused Multi-Document Summarization update task . . . . .	166
6.2.1	Modifications with respect to DUC 2007 SEMsum . . . . .	167
6.2.2	Analysis of the results . . . . .	169
6.3	Conclusions . . . . .	170
<b>7</b>	<b>Multi-source FEMsum</b>	<b>173</b>
7.1	Evaluation framework . . . . .	174
7.2	Configurations of the evaluated approaches . . . . .	177
7.3	Analysis of the results . . . . .	178
7.4	Contrasting the results obtained in different tasks . . . . .	180
7.5	Conclusions . . . . .	181
<b>8</b>	<b>Conclusions</b>	<b>183</b>
8.1	Contributions . . . . .	185
8.2	Future Work . . . . .	187
	<b>References</b>	<b>189</b>
	<b>Index</b>	<b>207</b>

# List of Figures

1.1	Diagram of the research method applied to obtain this thesis. . . . .	13
2.1	A summarization multitask classification. . . . .	25
2.2	Example of an annotated peer summary, using a pyramid created from 7 human 250-word summaries. . . . .	43
2.3	Discussion on the annotated peer summary, using a pyramid created from manual summaries. . . . .	44
3.1	Flexible Eclectic Multitask summarizer (FEMsum) Global Architecture . . . . .	49
3.2	FEMsum data flow . . . . .	53
3.3	Notation used in the diagrams depicted in Figures 3.4, 3.6, 3.9, 3.13, 3.20 . . . . .	58
3.4	Linguistic Processor instantiated FEMsum component. . . . .	60
3.5	Linguistic Processor constituents and text represented by <i>sent</i> , <i>sint</i> and <i>env</i> . . . . .	61
3.6	Query Processor instantiated FEMsum component. . . . .	64
3.7	An example of the information provided in one DUC 2007 topic. . . . .	65
3.8	DUC 2007 complex natural language query processor output. . . . .	65
3.9	Relevant Information Detector instantiated FEMsum component. . . . .	67
3.10	Terms related in an agency news document. . . . .	70
3.11	Terms related with the concept “candidato.” . . . . .	71
3.12	Terms related with the concept “debate.” . . . . .	71
3.13	Content Extractor instantiated FEMsum component. . . . .	79
3.14	Mapping-Convergence algorithm. . . . .	82
3.15	TALP-QA Question Types. . . . .	83

3.16	Semantic Content Extractor modules . . . . .	84
3.17	Example of the graph of an <i>env</i> representation . . . . .	84
3.18	Environment ( <i>env</i> ) representation of a sentence. . . . .	85
3.19	Candidates Selector procedure. . . . .	86
3.20	Summary Composer instantiated FEMsum component. . . . .	88
3.21	The MetaServer tasks. . . . .	93
3.22	Client MetaServer Dialog. . . . .	94
3.23	Servers to summarize a scientific oral presentation in English. . . . .	95
3.24	Condition Notation. . . . .	95
3.25	Configuration of a <i>LP</i> server used to process documents in Spanish or Catalan. . . . .	96
4.1	Web interface to create the Spanish summarization corpus. . . . .	101
4.2	<i>LCsum</i> FEMsum instantiation process. . . . .	103
4.3	Example of an original agency news document in Catalan from EFE with the corresponding lexical chains. . . . .	106
4.4	Example of the rhetorical structure of a sentence. . . . .	108
4.5	A text represented by lexical chains. Boxes are used for the identity chain members and circles for similarity chain members. . . . .	109
4.6	A text represented by rhetorical structure. Grey text contains rhetorically subordinate text segments, while underlined text contains central segments and discourse markers are in circles. . . . .	110
4.7	<i>LCsum (DiscStr)</i> , integration of discursive information: lexical chains (left) and discourse structural (right). . . . .	111
4.8	Experiments to assess the impact of discourse structure on lexical chain members. . . . .	112
4.9	<i>MLsum</i> FEMsum instantiation process. . . . .	116
4.10	Manual evaluations for <i>Coverage</i> and <i>Length-Adjusted Coverage</i> for the Document Understanding Conference (DUC) 2003 participant systems [7 - 26] and the DUC Baseline [1]. . . . .	117
4.11	Steps of the Full Parser Compressor process. . . . .	119



---

4.12	For each DUC 2003 participant system 4 values are depicted: the DUC <i>Coverage</i> , the approximated <i>Coverage</i> , the DUC <i>Length-Adjusted Coverage (LAC)</i> and the approximated <i>LAC</i> . For each new evaluated system only approximated values are depicted. . . . .	121
5.1	Summary Generation Tool. . . . .	126
5.2	Automatic transcript chunks aligned with chunks from the manual transcript. . .	128
5.3	Example of a TU of 169 words. . . . .	133
5.4	NE lexical chain examples. . . . .	133
5.5	<i>PreSeg</i> and <i>PostSeg</i> modules. . . . .	139
6.1	Scores obtained by manual and automatic constituent annotation of the human and the participant DUC system summaries, averaged by system. . . . .	149
6.2	Comparison of automatic metrics applied to DUC 2005 participants. . . . .	150
6.3	Modules of <i>QASUM-TALP</i> , a <i>QAsum</i> approach. . . . .	151
6.4	Example of a sentence linguistically processed and the generated questions. . . .	152
6.5	AutoPan scores of the DUC systems and the new approaches, averaged by system.	154
6.6	FEMsum approaches at DUC 2006. . . . .	156
6.7	Procedure to train a Support Vector Machine to rank sentences. . . . .	161
6.8	DUC 2007 FEMsum architecture for the update task. . . . .	168
7.1	CHIL manual evaluation web interface. . . . .	177



# List of Tables

1.1	Example of 10-word summaries produced from an automatic oral presentation transcript. The human summaries and the corresponding title and the keywords of the paper are used as reference and summaries. . . . .	8
1.2	Given a user query, example of 100-word manual and automatic summaries produced from different sorts of documents from an oral presentation. . . . .	9
1.3	Given a user need, example of 250-word manual and automatic summaries produced from a set of 25-50 documents. . . . .	10
2.1	Systems or Groups working in several tasks . . . . .	36
2.2	Some on-line demos of summarization systems, both commercial and academic. . . . .	37
2.3	Example of a pyramid SCU (D633.CDEH.pyr) with the <b>contributors</b> shown in their original sentential contexts. . . . .	40
2.4	Peer annotation examples. . . . .	42
3.1	Association of summary production process model stages to the correspondent architecture component. . . . .	52
3.2	Components of the instantiated FEMsum approaches . . . . .	55
3.3	FEMsum approaches related with a summarization model. . . . .	57
3.4	Question Generation patterns by NE type. . . . .	66
4.1	<i>LCsum</i> performance when summarizing agency news. . . . .	104
4.2	<i>LCsum</i> performance when summarizing the same documents in Catalan or Spanish. . . . .	105
4.3	Example of Spanish Proper Names, NEs or acronyms translated as a common nouns in Catalan. . . . .	105
4.4	Results of the integration of lexical chains and discursive structural information . . . . .	113

5.1	Example of content from an automatic transcript. . . . .	127
5.2	ROUGE-1 scores taking as reference: 2 extract-based human summaries, 2 abstract-based author paper summaries (T: title, K:list of keywords) or all of them as 10-word summary models. Types of evaluated summaries: human ideal automatic systems ( <b>upper bound</b> ), paper abstract based, and FEMsum ( <i>best approach</i> ). . .	131
5.3	ROUGE-1 scores taking as reference: 2 extract-based human summaries, 1 abstract-based author paper summaries (T-K: title and list of keywords) or all of them as 30-word summary models. Types of evaluated summaries: human ideal automatic systems ( <b>upper bound</b> ), paper abstract based, and FEMsum ( <i>best approach</i> ). . .	131
5.4	ROUGE-1 scores taking as reference: 2 extract-based human summaries, 1 abstract-based author paper summaries (A: abstract) or all of them as 70-word summary models. Types of evaluated summaries: human ideal automatic systems ( <b>upper bound</b> ), paper abstract based, and FEMsum( <i>best approach</i> ). . . . .	132
5.5	Performance of the FEMsum approaches producing 10,30,70-word summaries for 29 transcripts. . . . .	136
5.6	Corpus comparative study. . . . .	136
5.7	Example of strong lexical chains members and their score. . . . .	137
5.8	Performance of the FEMsum approaches producing 10,30,70-word summaries for 29, 34 and 39 manual transcripts. . . . .	138
5.9	Example of collocations filtered by syntactic patterns. . . . .	139
5.10	Example of discourse markers, their relevance and syntactic class. . . . .	140
5.11	Upper bound (Human ideal automatic systems) and system performances when summarizing manual or automatic transcripts. ROUGE unigram overlap measure have been computed when taking, extract-based human summaries (H1:human1 and H2:human2), abstract-based author paper summaries (T:title and K:list of keywords) or all of them as model summaries. . . . .	142
5.12	Pearson correlation coefficients between the different models in each evaluation set. . . . .	143
6.1	DUC 2005 complex natural language questions. . . . .	146
6.2	Correlation values for several metrics, evaluating average scores of non-human systems. Values which exceed the p-value for $p = 0.01$ are shown in <b>bold</b> . Values which exceed the p-value for $p = 0.05$ are shown in <b><i>bold italics</i></b> . . . . .	149
6.3	Precision, Recall and F-measure obtained when selecting relevant sentences. . . . .	155

---

6.4	FEMsum linguistic quality scores by approach, as well as the mean of the 34 participant systems obtained in the associated subset of summaries. . . . .	158
6.5	DUC FEMsum manual content responsiveness score. . . . .	159
6.6	DUC content responsiveness scores by approach. . . . .	159
6.7	DUC content responsiveness scores distribution by approach . . . . .	159
6.8	DUC ROUGE measures when considering 4 manual summaries as references. . .	160
6.9	ROUGE and AutoPan results on the original UAM-Titech06 and SEMsum or when using several SVMs. . . . .	164
6.10	FEMsum linguistic quality scores by approach, as well as the mean of the 32 participant systems obtained in the associated subset of summaries. . . . .	166
6.11	Content responsiveness score and mean distance for human, the best system, our submission and the baseline. . . . .	166
6.12	<i>SEMsum (Update)</i> responsiveness evaluation, 22 participants. . . . .	169
7.1	Example of ISL seminar topics. . . . .	174
7.2	Example of queries for three of the ISL seminar topics. . . . .	175
7.3	Example of the three reference summary models of one of the 20 summaries to be evaluated. . . . .	176
7.4	ROUGE measures when considered 3 manual summaries as references. . . . .	179
7.5	Responsiveness considering 3 human models when evaluating automatic summaries and 2 when evaluating human summaries. . . . .	179
7.6	Responsiveness scores distribution by automatic system. . . . .	179
8.1	Details of several tested approaches: input, purpose and output aspects. . . . .	184



# 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 Programming
- 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