Event-based Multi-document Summarization

Luís Carlos dos Santos Marujo

CMU-LTI-15-010

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., PA 15213
United States of America

Dep. of Computer Science and Engineering
Instituto Superior Técnico
University of Lisbon
Av. Rovisco Pais, 1, 1049-001 Lisbon
Portugal

Thesis Advisers:

Prof. Jaime Carbonell, Carnegie Mellon University
Prof. Anatole Gershman, Carnegie Mellon University
Prof. David Martins de Matos, Instituto Superior Técnico, Universidade de Lisboa
Prof. João Paulo Neto, Instituto Superior Técnico, Universidade de Lisboa

Thesis Committee:

Prof. Ricardo Baeza-Yates, Yahoo!Research/Univ. Pompeu Fabra/Univ. de Chile
Prof. Jaime Carbonell, Carnegie Mellon University
Prof. Eduard Hovy, Carnegie Mellon University
Prof. David Martins de Matos, Instituto Superior Técnico, Universidade de Lisboa
Prof. Ana Paiva, Instituto Superior Técnico, Universidade de Lisboa
Prof. Isabel Trancoso, Instituto Superior Técnico, Universidade de Lisboa

Submitted in partial fulfillment of the requirements
for the Dual degree of Doctor of Philosophy
in Language and Information Technologies.

To my parents and grandmother,

Acknowledgements

> *No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude.*
>
> Alfred North Whitehead (1861-1947), *English Mathematician and Philosopher*

There is a long list of exceptional people to whom I wish to express my gratitude, who have supported and motivated me during this journey which I started six years ago. I would like to start by thanking my excellent advisors at IST and CMU: Anatole Gershman, David Martins de Matos, Jaime Carbonell, and João P. Neto. Thanks to their diverse areas of expertise and extremely insightful advice, I was able to understand and navigate between several very interesting research topics including Information Retrieval, Information Extraction, Natural Language Processing, Machine Learning, and Speech processing. The countless hours of great guidance and enthusiastic discussions either physically or via Skype meetings were priceless and truly shaped this thesis. I also need to thank all committee members - Ana Paiva, Eduard Hovy, Ricardo Baeza-Yates, and Isabel Trancoso - for their time and amazing comments that tremendously improve this thesis. Having such distinguished thesis committee members born in four different continents (Africa, America, Asia, and Europe) has definitely brought different perspectives to this thesis. During my stay at CMU, I was very fortunate to attend amazing courses, collaborate in Anatole Gershman and Robert Frederking research group, which included Kevin Dela Rosa, Bo Lin, Sushant Kumar, and Rushin Shah. I really learned a lot from the interactions, presentations, and classes with several CMU Faculty: Alan Black, Alon Lavie, Bhiksha Raj, Carolyn Rosé, Chris Dyer, Eugene Fink, Eric Nyberg, Florian Metze, Jack Mostow, Jaime Callan, Lori Levin, Manuela Veloso, Noah Smith, Norman Sadeh, Sharon Carver, Ralf Brown, Ravi Starzl, Rita Singh, Roni Rosenfeld, Scott Fahlman, Teruko Mitamura, and Yiming Yang. During two semesters I was teaching assistant of two inspiring courses: Inventing the Future of Services, and Mobile & Pervasive Computing services. Clearly, I would like to thank Anatole Gershman, and Norman Sadeh for accepting me as their Teaching Assistant. I also wish to thank Hazim Almuhimedi, Justin Cranshaw, Bin Liu, and Linda Francona for their help. As a student at CMU, I am also grateful to have

interacted with several amazing colleagues and in some cases part of their families: Almer Tigelaar, Alok Parlikar, Anagha Kulkarni, André Martins, Arnold Overwijk, Avner Maiberg, Christoph Schrey, Dani Yogatama, Guang Xiang, Hideki Shima, Jaime Arguello, Jose Gonzalez, Juan Pino, Jun Zhu, Konstantin Salomatin, Leid Zejnilovic, Le Zhao, Long Qin, Luís Brandão, Mahesh Joshi, Mathew Marge, Maxim Makatchev, Michael Denkowski, Ming Sun, Narjes Sharif, Ni Lao, Prasanna Kumar, Reyyan Yeniterzi, Rodrigo Belo, Vasco Pedro Calais, Shoou-I Yu, Thuy Linhn Nguyen, Ting-Hao Kenneth, Yi-Chia Wang, Weisi Duan, Zi Yang, Zhou Yu, among many others. I am thankful to Bob Frederking, João Barros, João Claro, José Moura, Sara Brandão, and Stacey Young for helping to clarify several questions regarding the CMU requirements and CMU—Portugal program. Thanks to Benjamin Cook, Carolina Carvalho, Dana Gates, Krista McGuigan, Kelly Widmaier, Eric Riebling, Kathleen Schaich, Lori Spears, Mary Jo Bensasi, Nicole Hillard-Hudson, and the remaining LTI and ICTI/CMU Portugal staff. During my stay in Portugal, I also had the opportunity to interact with many great colleagues. At IST/INESC, I need to give a very special thanks to Ricardo Ribeiro and Wang Ling (王零) for actively collaborating in my research work. I would also thank João Miranda, João Paulo Carvalho, José Portêlo for very fruitful discussion. As a student at IST, I have also benefited from the courses taught by Mário Figueiredo, and David Matos. These courses gave me insights into Machine Learning and Information Extraction.

Of course, I also wish to express my gratitude to several current and past L2F group members including Alberto Abad, Annamaria Pompini, António Serralheiro, Anabela Barreiro, Fernando Batista, Helena Moniz, Hugo Meinedo, Hugo Rosa, João Graça, Jorge Batista, José David, Luísa Coheur, Miguel Bugalho, Nuno Mamede, Ramón Astudillo, ... Special thanks to Hugo Meinedo for his tremendous system admin work in collaboration with David Matos, which was very important to running my experimental work.

I would like to thank several members of the Voice Interaction start up for helping with the AUDIMUS speech recognition system and testing some of my research models: Carlos Mendes, João Nuno Neto, Márcio Viveiros, Sérgio Paulo, Renato Cassaca, Tiago Luís, ...

I would like to thank Aurélia Constantino, Elisabete Ferreira, Teresa Mimoso, Vanda Fidalgo, and the remaining L2F/INESC-ID staff.

Along the six years of my Ph.D., I clearly benefited from attending Summer Schools and Conferences. Special thanks the organizers of S3MR including Touradj Ebrahimi, Naeem Ramzan, Murat Tekalp for giving me a travel grant and a best poster award. There is also very long list of people that gave me feedback about my work in conferences, but listing them all here is unfeasible. To them all, thank you!

To some friends in Lisbon: Ana Carapeto, Ana Cristina Gomes, Miguel Costa, Filipa Peleja, Pedro Costa, Gabriel Barata, Ricardo Dias, ... thank you!

Last but not least to my parents, Cidália and Francisco, to my grandmother, Deolinda, thank you for your unconditional love and support.

My Ph.D. studies were funded by Fundação para a Ciência e Tecnologia (FCT), Information and Communication Technology Institute (ICTI), Carnegie Mellon Portugal Program, FCT grant SFRH/BD/33769/2009 and CMUP-EPB/TIC/0026/2013

Lisbon, May 30^{th}, 2015
Luís Carlos dos Santos Marujo

Resumo

Diariamente o número de notícias, sobre eventos ocorrendo no mundo, está a crescer exponencialmente. Simultaneamente, organizações estão à procura de informação sobre eventos atuais e passados que as afectem, como por exemplo fusões e aquisições de empresas. As organizações necessitam, para tomarem decisões, de obter informação sobre eventos de uma forma rápida e sumarizada. Sistemas de recuperação de informação e sumarização baseados em eventos oferecem uma solução eficiente para este problema. A maioria dos trabalhos de investigação em sumarização utiliza notícias. Apesar de este tipo de documentos ser caracterizado por transmitir informação sobre eventos, a maioria do trabalho foca-se em metodologias que não têm em conta este aspeto.

Os métodos propostos para a sumarização de vários documentos são baseados na combinação hierárquica de vários sumários individuais. Utilizamos informação sobre eventos para melhorar sumarização de vários documentos. A nossa metodologia é baseada num método composto por duas etapas, desenhado para sumarização individual de documentos, que extrai uma coleção de expressões chave, que são depois usadas num modelo de centralidade como forma de capturar relevância e auxiliar na seleção de passagens. Também exploramos como adaptar o modelo de sumarização para documentos individuais baseado no modelo de centralidade para sumarização de vários documentos e utilizando informação sobre eventos, visto que necessitávamos de ter um bom sistema de base. Devido à extração de expressões chave desempenhar um papel importante na sumarização, nós melhoramos uma ferramenta, que é o estado da arte, de extração de expressões chave, com quatro novos conjuntos de descritores semânticos. O método de deteção de eventos é baseados em impressões digitais difusas, que é um método supervisionado treinado em documentos anotados com eventos. Nós exploramos três formas de integrar informação sobre eventos, obtendo resultados que são estado de arte para sumarização de um e vários documentos, utilizando filtros e descritores baseados em eventos. Para lidar com a possível utilização de termos diferentes descrevendo o mesmo evento, exploramos representações distribuídas de texto na forma de palavras embebidas, que contribuiu para melhorar os resultados da sumarização de vários documentos.

A avaliação automática e humana mostram que estes métodos melhoram o estado da arte de sistemas de sumarização de vários documentos em dois corpora de avaliação frequentemente utilizados, o DUC 2007 e o TAC 2009. Obtemos um melhoramento, em termos de ROUGE-1, de 16% no caso do TAC 2009 e de 6% no DUC 2007. Também obtivemos melhoramentos em

termos de ROUGE-1 sobre sistemas estado da arte variando entre 32 % para texto limpo e 19% para texto com ruído. Estes melhoramentos derivam da inclusão de expressões chave e informação sobre eventos. A extração de expressões chave foi também refinada com etapas de pre-processamento e características que levaram a um melhoramento relativo em termos de valores de NDCG de 9%. A introdução de impressões digitais difusas para deteção e classificação de eventos possibilitou a deteção de todos os tipos de eventos, enquanto o melhor sistema adversário, um SVM melhorado com mais descriptors, só conseguiu detetar 85% dos diferentes tipos de eventos. Isto levou a um grande melhoramento em termos de G-Mean e variantes quando utilizamos impressões digitais difusas.

Abstract

Daily amount of news reporting real-world events is growing exponentially. At the same time, Organizations are looking for information about current and past events that affects them, such as mergers and acquisitions of companies. The Organizations need to obtain event information in a fast and summarized form to make decisions. Event-based retrieval and summarization systems offer an efficient solution to this problem. Most summarization research work uses news stories. Although this type of documents is characterized by conveying information about events, almost all work concentrates on approaches that do not take into account this aspect.

The proposed multi-document summarization methods are based on the hierarchical combination of single-document summaries. We improved our multi-document summarization methods using event information. Our approach is based on a two-stage single-document method that extracts a collection of key phrases, which are then used in a centrality-as-relevance passage retrieval model. To adapt centrality-as-relevance single-document summarization for multi-document summarization that is able to use event information, we needed a good and adaptable baseline system. Because the key phrase extraction play a significant role in the summarization, we improved a state-of-the-art key phrase extraction toolkit using four additional sets of semantic features. The event detection method is based on Fuzzy Fingerprint, which is a supervised method trained on documents with annotated event tags. We explored three different ways to integrate event information, achieving state-of-the-art results in both single and multi-document summarization using filtering and event-based features. To cope with the possible usage of different terms to describe the same event, we explored distributed representations of text in the form of word embeddings, which contributed to improve the multi-document summarization results.

The automatic evaluation and user study performed show that these methods improve upon current state-of-the-art multi-document summarization systems on two mainstream evaluation datasets, DUC 2007 and TAC 2009. We show a relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007. We have also obtained improvements in ROUGE-1 upon current state-of-the-art single-document summarization systems of between 32% in clean data and 19% in noisy data. These improvements derived from the inclusion of key phrases and event information. The extraction of key phrases was also refined with additional pre-processing steps and features, which lead to a relative improvement in NDCG

scores of 9%. The introduction of Fuzzy Fingerprints for event detection enabled the detection of all event types, while the best competitor, an SVM with enhanced features, only detects roughly 85% of the different types of events. This lead to a large increase in the G-Mean and variants results when using the Fuzzy Fingerprints method.

Key phrases

Automatic Key Phrase Extraction (AKE)
Co-reference Normalization
Crowdsourcing
Event detection
Event-based single-document summarization
Event-based multi-document summarization
Imbalanced Multi-class learning
Important Passage Retrieval
KP-Centrality
FEME-KP-Centrality
Light Filtering
Multi-document summarization
Rhetorical Signals
ROUGE
User study

List of Abbreviations

ACE — Automatic Content Extraction is a research program to advance information extraction techniques.

ACM — Association for Computing Machinery is the world's largest computing society.

AKE — Automatic Key phrase Extraction is the process of extracting the most relevant phrases from a document.

AMT — Amazon's Mechanical Turk is a crowdsourcing Internet marketplace.

ASR — Automatic Speech Recognition, also known as computer speech recognition, is the process of converting the speech signal into written text.

AWE — Average Word Embeddings is a language model that average the language model of words.

BN — Broadcast News transmitted over radio and TV.

CE-KPC — Combination of Event Filtering-based and Event-Enhanced KP-Centrality.

CKP-Centrality — Key Phrase Confidence-based Centrality is a summarization method.

CMU — Carnegie Mellon University is a private research university in Pittsburgh, Pennsylvania, United States.

CRFs — Conditional Random Fields are a discriminative undirected probabilistic graphical model for labeling and segmenting structured data.

DUC — Document Understanding Conferences are summarization evaluation workshops.

EE-KPC — Variation of the KP-Centrality method with event-based features.

EF-KPC — Variation of the KP-Centrality method with filtering of events.

ENER — English Event Reports dataset.

FAO — Food and Agriculture Organization of the United Nations is an intergovernmental organization.

HIT — Human Intelligence Task is individual job designed for crowdsourcing workers.

IBM — International Business Machines Corporation is an American multinational corporation.

IDF — Inverse Document Frequency is a information retrieval metric, which measures how important a term is.

INESC-ID — Institute for Systems and Computer Engineering: Research and Development is a non-profit organization devoted to research in the field of information and communication technologies.

IR — Information Retrieval is a research branch of artificial intelligence, computer science, and Natural Language Processing that studies how to obtain information resources relevant to an information need from a collection of information resources

ITF — Inverse Topic Frequency is an adaptation of IDF to fuzzy fingerprints.

JSD — Jensen-Shannon Divergence is a method to measure similarity between two probability distributions.

KBP — (TAC) Knowledge Base Population is research trend organized by National Institute of Standards and Technology.

KEA — Key Extraction Algorithm is a toolkit designed to extracted key phrases.

KLD — Kullback–Leibler Divergence is a method to measure similarity between two probability distributions.

KP-Centrality — Key Phrase-based Centrality is a summarization method.

L^2F — Spoken Language Systems Laboratory is a research department at INESC-ID.

LDA — Latent Dirichlet Allocation is a generative probabilistic topic model.

LDC - Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories that creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes.

LSA - Latent Semantic Analysis is an unsupervised vector-space method created for automatic indexing and retrieval.

LTI — Language Technologies Institute is a research department in the School of Computer Science at Carnegie Mellon University.

MMR — Maximal Marginal Relevance is a query-oriented method for summarization.

nDCG — Normalized Discounted Cumulative Gain is evaluation metric for ranking quality and measure the importance (gain) of an item based on its relevance and position in an order list.

NER — Named Entity Recognizer is a system that labels sequences of words in a text which are names of things, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

NLP — Natural Language Processing is a branch of artificial intelligence, computer science, and linguistics that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages.

OKP-Centrality — Only Key phrase containing passages-based Centrality is a summarization method.

POS — Part-of-Speech tag, also known as word class, lexical class or lexical class are traditional categories of words intended to reflect their functions within a sentence.

ROUGE — Recall-Oriented Understudy for Gisting Evaluation is the most used summarization evaluation metric.

SMO — Sequential Minimal Optimization is an algorithm used to train SVMs.

SPER — Spanish Event Reports dataset.

SU — Sentence Unit.

SVM — Support Vector Machines are a set of supervised learning methods used for classification and regression based on the Structural Risk Minimization inductive principle.

TAC — The Text Analysis Conference (TAC) is a series of evaluation workshops, organized by National Institute of Standards and Technology. TAC is divided in a sets of tasks known as tracks, each of which focuses on a particular subproblem of NLP: Question Answering, Recognizing Textual Entailment, Summarization, and Knowledge Base Population.

TDT — Topic Detection and Tracking is a DARPA-sponsored initiative to investigate the state-of-the-art in finding and following new events in a stream of broadcast news stories.

TF-IDF — Term Frequency-Inverse Document Frequency is one of the simplest ranking functions as it is a algorithm to score the importance of words (or terms) in a document based on how often they appear in multiple documents

WEKA — Waikato Environment for Knowledge Analysis is a collection of machine learning algorithms for data mining tasks.

Contents

List of Figures

List of Tables

1 Introduction

> *What Information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention.*

Carnegie Mellon University Professor Herbert A. Simon (1978), *Nobel Laureate and Turing Award winner*

Humans have reported problems in understanding or making decisions when faced with excessive amounts of information, which is nowadays known as *Information Overload* problem [199]. The abundance of information makes the search for relevant information more complex, such as finding a needle in a haystack.

At the same time, the abundance of information does not always cover relevant information to understand or make decisions, also known as, *Information Scarcity* problem. The *Information Scarcity* problem is one of the reasons behind the invention of the printing press, in around 1440 A.D. It enabled the mass production of books. Using the haystack metaphor again, it means that the haystack grows with more books being available, but at the same time the likelihood of the haystack actually having the needle also increases.

The scale and complexity of the Information Overload and Scarcity problems increased with the rise of modern computers in the 1960s. The modern computers were connected to create the Internet in the late 1970s. This network became global and brought access to a very large amount of information. Daily news articles and broadcast news are a good example of huge amounts of information daily published in the Internet. Again, people have more difficulty finding information (needle) to understand events in news documents (haystack). There is a very famous citation presented in an article from The New York Times: "Can't Grasp Credit Crisis? Join the Club" [5], noting how complex is to follow the Economic crisis events. Since the number of news articles covering complex events is large, a reasonable solution is to generate a summary containing the most important sentences (needles) from a set of news documents (haystack).

1.1 Motivation

Many automatic summarization systems have been proposed in order to cope with the growing number of news stories published online. The main goal of these systems is to convey the important ideas in these stories by eliminating less crucial and redundant pieces of information. In particular, most of the work in summarization has been focused on the news domain, which is strongly tied to events, as each news article generally describes an event or a series of events. However, few attempts have focused on the use of automatic techniques for event classification for summarization systems for this domain [71]. In fact, most of the work on multi-document summarization are either Centrality-based [173, 61, 212, 177], Maximal Marginal Relevance-oriented (MMR) [45, 78, 187, 112], or Coverage-base methods [116, 195, 64, 108, 71, 122, 230]. Generally, centrality-based models are used to generate generic summaries, the MMR family generates query-oriented ones, and coverage-based models produce summaries driven by topics or events.

The use of event information in multi-document summarization can be arranged in the following categories: pioneer **hand-based experiments** [54]; **pattern-based approaches** based on enriched representations of sentences, such as the cases of the work presented by Zhang et al. [230] and by Wenjie Li et al. [108], which define events using an event key term and a set of related entities, or centrality-based approaches working over an event-driven representation of the input [71], where events are also pattern-based defined; and, **clustering-based** event definition [107].

The major problem of these approaches is that it is difficult to relate different descriptions of the same event due to different lexical realizations. In our work, we address this problem by using an event classification-based approach and including event information supported by a distributed representations of text — the skip-ngram model [151]. Our event detection and classification framework is based on vector-valued fuzzy sets [86, 138], which is able to detect all event types in the ACE 2005 Multilingual Corpus [208].

Concerning the summarization framework, we use the two-stage summarization approach (KP-Centrality) [178] that we proposed for generic single-document summarization. The two-stages framework starts by extracting the most meaningful words and phrases (key phrase extraction) that are then used to guide the centrality-as-relevance summarization method. This summarization provides an adequate framework for the integration of additional information. Within this framework, we explore different ways of incorporating event information, attaining state-of-the-art results in both single and multi-document summarization tasks.

The final evaluation of our work is performed using the standard summarization evaluation metric, ROUGE [114]. Moreover, to better understand the impact of using event information,

we also performed a human evaluation using Amazon Mechanical Turk (AMT)[1]. Through the human evaluation, we show that despite ignoring the coherence computation in the summarization process, we are still able to obtain good results.

1.2 Thesis statement

Event-based multi-document summarization is feasible and useful to deliver important information about news events. Human evaluation shows that our summaries are on average more useful for humans than reference and baseline summaries.

To show the challenges involved in determining an event-based multi-document summary from a set of news documents, let us look at the events that culminated in Muammar Gaddafi's death. It occurred on 20^{th} October 2011 during the Battle of Sirte [6].

Querying a search engine with "Muammar Gaddafi's death OR War in Libya OR battles" returns news documents about War in Libya connected with Gaddafi's death. Consider that a user is reading a news document about "Muammar Gaddafi's death" and he did not follow the war in Lybia. But wishes to be informed about the events leading to Gaddafi's death.

An example of event-based multi-document summary describes the key battles (events) in the war in Libya that culminated in Gaddafi's death. There were battles in nine cities: in some of these cities, there was more than one battle. The last column of Table 1.1 shows an example of an event-based multi-document summary. This table is a manual summary representing the aspirational goal of our work.

Table 1.1: Example of Event-based Multi-document Summary of Muammar Gaddafi's death (source of news: The Guardian).

Date	Event	Source News Title	Summary
17-20 Feb	1^{st} Battle of Benghazi	Libya protests: gunshots, screams and talk of revolution	Benghazi student says fear of Muammar Gaddafi's regime is ebbing away.
24 Feb - 11 Mar	1^{st} Battle of Zawiya	Zawiya town centre devastated and almost deserted	Gaddafi's men are cleaning up Zawiya, the town they have finally taken after bombarding it for a week. They have brought in road sweepers to brush away the evidence of the worst fighting between Libyans in a century. It is certainly the worst devastation I've seen in any town centre.
18 Feb - 15 May	Battle of Misrata	Libyan rebels pay a heavy price for resisting Gaddafi in Misrata	Libyan rebels pay a heavy price for resisting Gaddafi in Misrata with 1,000 dead and a further 3,000 injured, the two-month-old war has taken its toll on the people of the city.

[1]https://www.mturk.com/

2 Mar	1^{st} Battle of Brega	Battle for Brega could mark start of real war in Libya	Battle for Brega could mark start of real war in Libya at least six people die as eastern town fights off attack by pro-Gaddafi forces.
13-15 Mar	2nd Battle of Brega	Gaddafi forces rout rebels in eastern Libya	Rebels driven out of town of Brega under heavy bombardment as pro-regime forces advance towards Benghazi.The Gaddafi forces' advance came as Hillary Clinton, the US secretary of state, prepared to travel to the region to meet representatives of the rebels' revolutionary council.
15-17 Mar	1st Battle of Ajdabiya	Gaddafi's effort to defeat rebels before international support pays off	Muammar Gaddafi's effort to defeat the rebels before international support can come seems to be paying off, with the uprising close to collapse as the US ended weeks of stalling to join Britain and France in supporting a United Nations resolution to impose a no-fly zone over Libya.
17 Mar	Battle of Zueitina	Gaddafi threatens retaliation in Mediterranean as UN passes resolution	Muammar Gaddafi has pledged to retake the rebel stronghold of Benghazi and warned that any foreign attack on Libya would endanger air and maritime traffic in the Mediterranean area, as the UN security council voted for military intervention.
19-20 Mar	2nd Battle of Benghazi	Libya crisis: Gaddafi troops launch bloody assault on Benghazi	Coalition air strikes relieve pressure on rebel forces as Gaddafi defies ceasefire. Gaddafi's assault on the rebel stronghold was led by forces that broke away from the attack on the town of Ajdabiya, 90 miles along the coast, in what appeared an effort to seize Benghazi before Tripoli is forced to halt its bid to crush the month-long popular uprising.
21-26 Mar	2nd Battle of Ajdabiya	Libyan rebels rejoice in Ajdabiya as air strikes drive Gaddafi loyalists out	Fall of Ajdabiya is first significant victory for rebels since coalition strikes began a week ago.
31 Mar - 7 Apr	3rd Battle of Brega	Nato air strike 'kills Libyan rebels'	At least 13 rebel fighters were killed, according to one report. Fighters who fled to the town of Ajdabiya said the attack, on the outskirts of Brega, involved a number of Nato bombing runs, and several tanks were destroyed.
14-21 Jul / 9-22 Aug	4th Battle of Brega	Libyan rebels in Brega fall back	Rebels in Libya's east pulled back Friday after a failed advance on an oil town, as embattled ruler Muammar Gaddafi called on his followers to strike back at NATO.
13-18 Aug	Battle of Gharyan	Libyan rebels capture demoralised Gaddafi troops	But Brahim said the rebels would struggle to take Gharyan, a crucial gateway to Tripoli, as Gaddafi's forces had numerous troops and heavy weapons there.
13-20 Aug	2nd Battle of Zawiya	Rebel advances mask uncertainty over Libya's future	It is rebels in the west – from the Nafusa mountains and Misrata – that have captured Zawiya, 30 miles west of the capital, Garyan, 40 miles south and Zlitan, 80 miles to the west. Their commanders and politicians will, if they storm the Libyan capital, demand a greater say in what is currently a Benghazi-centred administration.

20-28 Aug	Battle of Tripoli	Battle for Tripoli: pivotal victory in the mountains helped big push	Yet within two weeks, the tide in the Libyan conflict has changed dramatically. Gaddafi's defences have crumbled, one government stronghold after another has collapsed and the rebels now control most of Tripoli.
15- 20 Oct	Battle of Sirte	Gaddafi loyalists hold out in last desperate resistance at Sirte, as families flee	The war in Libya is almost over, but for ordinary people in Sirte's District 2 the misery gets deeper.
20 Oct	**Gaddafi's death**	**Muammar Gaddafi, the 'king of kings' dies in his hometown**	**Colonel Muammar Gaddafi was born near Sirte, (...) On Thursday, after a brutal – and ultimately hopeless – last stand, it was the place where he died.**

1.3 Thesis Structure and Contributions

After this introductory chapter, we will review important background work for this thesis in Chapter 2. The following chapters will present our solution to the problem of generating event-based multi-document summaries:

- **Chapter 3 - Single-document summarization Extraction of News Documents Description**: this chapter focused on the first steps to create a single-document summarization. Our focus will be on the extraction of key phrases as a first step to improve summarization.

  Some of the contributions we will present in Chapter 3 have been published:

  - Luís Marujo, Márcio Viveiros, João P. Neto, Keyphrase Cloud Generation of Broadcast News, In proceeding of Interspeech 2011: 12^{th} Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy, August 2011 (won **Best Poster award** on S3MR 2011 - 2^{nd} Summer School on Social Media Retrieval)
  - Luís Marujo, Miguel Bugalho, João P. Neto, Anatole Gershman, Jaime Carbonell, Hourly Traffic Prediction of News Stories, In 3^{rd} International Workshop on Context-Aware Recommender Systems held as part of the 5^{th} ACM RecSys Conference, Chicago, USA, October 2011
  - Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João P. Neto, Anatole Gershman, Jaime Carbonell, Key Phrase Extraction of Lightly Filtered Broadcast News, In Proceedings of 15^{th} International Conference on Text, Speech and Dialogue (TSD 2012), Brno, Czech Republic, September 2012

- Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, João P. Neto, Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization, In Proceedings of the 8^{th} International Conference on Language Resources and Evaluation (LREC '12), Istanbul, Turkey, May 2012
- Ricardo Ribeiro, Luis Marujo, David Martins de Matos, João P. Neto, Anatole Gershman, Jaime Carbonell, Self Reinforcement for Important Passage Retrieval, In Proceedings of the 36^{th} Annual ACM Special Interest Group on Information Retrieval (SIGIR 2013), Dublin, Ireland, July 2013

- **Chapter 4 - Event-based Single-Document Summarization** - Since news stories cover events, limiting the retrieval of sentences for the summary to the sentences describing events is a logical solution. For this purpose, we used the new Fuzzy Fingerprint method [138], which performed better than several supervised machine learning classifiers including Support Vector Machine (SVM) and Random Forests (state-of-the art methods). We also analyzed the impact of several new features on the machine learning classifiers, which were insufficient to outperform our Fuzzy Fingerprint method.

  Part of this research work has been presented in the following publication:

  - Luís Marujo, Wang Ling, Anatole Gershman, Jaime Carbonell, João P. Neto, David Martins de Matos, Recognition of Named-Event Passages in News Articles, In Proceedings of 24^{th} International Conference on Computational Linguistics (Coling 2012), Mumbai, India, December 2012
  - Luís Marujo, João Paulo Carvalho, Anatole Gershman, Jaime Carbonell, João P. Neto, David Martins de Matos, Textual Event Detection using Fuzzy Fingerprint, In Proceedings of IEEE Intelligent Systems IS'14, Warsaw, Poland, September 2014

- **Chapter 5 - Event-based Multi-document Summarization** - This chapter will describe how we extend the generic single-document summarization method (KP-Centrality) to multi-document summarization. After that, the generic multi-document summarization is extended to include event information using the same strategies proposed for the single-document summarization method.

  - Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João Neto, Anatole Gershman and Jaime Carbonell, Extending a Single-Document Summarizer to Multi-Document: a Hierarchical Approach, In Proceedings of *SEM: the 4^{th} Joint Conference on Lexical and Computational Semantics, Denver, Colorado, USA, June 2015

- **Chapter 6 - Conclusions and Future Work** - closes the document with the conclusions and proposes the future work.

2 Background and Related Work

> *Everything that needs to be said has already been said. But since no one was listening, everything must be said again.*
>
> André Gide, Nobel Prize in Literature, *Le traité du Narcisse*, 1891

The generation of *Event-Based Multi-document Summaries* is a research problem that combines Event Detection and Extractive Summarization tasks. Both tasks start with the extraction of representations (descriptive metadata) of the input news documents. These representations included information such as title, source, subject, keywords, and key phrases.

The Term Frequency-Inverse Document Frequency (TF-IDF) weighting is the most frequently used method to extract descriptive metadata, such as keywords, because of its simplicity and fast calculation. TF-IDF [185] is formally defined in equation 2.1:

$$TF\text{-}IDF(i, D) = tf(t, d) \times idf(t, D) \tag{2.1}$$

$$idf(t, D) = log\left(\frac{|D|}{1 + |d \in D : t \in d|}\right) \tag{2.2}$$

where $tf(t, d)$ is the number of occurrences of term or phrase t in document d; $|D|$ is the number of documents in the corpus; $|d \in D : t \in|$ is the number of documents containing term or phrase t.

There are also four extra variants to the $tf(t, d)$ calculation defined for large documents [128]. The TF-IDF method is also used to obtain features used in Automatic Key phrase Extraction (AKE) task.

The next section reviews the key phrase extraction literature. Then, Section 2.2 reviews the literature on Event Detection and Tracking. Extractive Summarization methods are presented in Section 2.3. Section 2.4 describes the methods exploring both event detection and summarization closing the Chapter.

2.1 News Descriptions: Key Phrase Extraction

Key phrases are defined as the most relevant words or phrases from a news document. While key phrases provide a concise representation of news documents, their automatic identification is a non-trivial task [82]. This Natural Language Processing (NLP) task is defined as AKE. Key phrases are also used in Information Retrieval tasks to enhance Information Retrieval indexing, to help users in completing queries [215], and to improve the prediction of web traffic [130].

Information extraction tasks, such as event detection [166, 132], and text summarization [56, 228, 178] also benefit from key phrase extraction. As key phrases provide concise descriptions of news, they can also be used as tags in Tag clouds [137]. Tag clouds are weighted renditions of collections of words (tags) that represent the concepts, in a visually appealing way to summarize vast amounts of information [100].

Both supervised and unsupervised AKE methods have been explored in the literature, and both have their strengths and weaknesses. Both supervised and unsupervised AKE have two stages: generation of candidate key phrases, and classification/ranking/filtering of candidates. Unsupervised methods do not require tagged training data to classify/rank/filter the candidate key phrases. However, the supervised AKE methods outperform unsupervised AKE methods.

Despite the large amount of work on unsupervised key phrase extraction, TF-IDF remains the most used and robust unsupervised baseline on several datasets from a variety of domains [82]. Other unsupervised AKE methods investigated were simple statistics (e.g., word frequency [126], PAT-tree [49], language modeling [200]), graph-based ranking (Spectral Graph Clustering [228], TextRank [149], SingleRank [209], ExpandRank [209]), and clustering [142, 124].

Supervised AKE methods follow a fairly traditional approach of training a binary classifier to select an ordered list of the most likely key phrase candidates in news documents. To apply this approach to any input news documents, it is necessary to execute two-steps:

candidate generation step - retrieves all possible candidate phrases and filters malformed ones (e.g., phrases starting and/or ending in stop words).

classification step - judges if a candidate phrase is a key phrase. The decision has a score or confidence that is used to rank the key phrases.

The first two supervised AKE systems have in common the fact that both employed the two-step approach. One system was GenEx [203] and the other system was Key Extraction

Algorithm (KEA) [216]. GenEx implemented a genetic algorithm, using frequency and position as features, to classify and rank candidate key phrases. KEA [216] opted to use the Naïve Bayes classifier and to represent the candidate key phrases using TF-IDF. In 2006, Medelyan introduced KEA++ [147] extending KEA representation of candidate phrases (feature set) with three additional features: the position of the first occurrence of a candidate phrase, the length of a candidate phrase in words, and the node degree. The node degree represents the number of thesaurus (e.g., WordNet [62]) links that connect the term to other candidate phrases.

Three years later, Medelyan introduced Maui [145], the current state-of-the-art AKE system, which is built upon the KEA and KEA++ architectures. KEA's feature set grew from three to nine features in Maui. Five out of nine features are simple statistics (TF-IDF, first occurrence position, keyphraseness - measures how often a candidate phrase is a key phrase, candidate phrase length, and spread - difference between first and last position) and the remaining features are Wikipedia-based [19] features (the wikipedia is modeled as a graph to measure the number of links pointing to or leaving from candidate key phrases, number of Wikipedia page entries, distance between words composing the candidate key phrases). Some of the new features are not independent, for instance: first occurrence position and spread, or node degree and semantic relatedness are dependent pairs of features. Some classifiers deal with dependencies between features better than others. For example, KEA's Naïve Bayes classifier assumes that all features are independent. Thus, Naïve Bayes was replaced by a Bagged C4.5 decision tree in Maui [42, 170]. Additional supervised classifiers, such as SVM [231], and Conditional Random Fields (CRFs) [229], were tested for AKE.

More recent AKE methods explored the inclusion of pre-processing steps. These steps are executed on the input news documents before executing the AKE methods.

In 2012, we introduced the introduction of pre-processing steps in AKE with Light Filtering and Co-Reference Resolution. **Light Filtering** [135, 131] (explained in detail in Section 3.1.2) uses centrality-based summarization to eliminate about 10% of the document's sentences. **Co-Reference Normalization** [131] (described in detail in Section 3.1.3) transforms several forms of references to the same named entity into a single form, using a co-reference resolution system.

Later, also in 2012, another pre-processing step, Text Denoising, was presented by Shams et al. [194]. Text Denoising, is a heuristic-based text reduction method which assumes a correlation between the high readability level sentences (high complexity of vocabulary and syntax) and the most content-rich sentences. The method uses the Gunning's Fog Index Readability score [76, 77] (Eq.2.3) to rank sentences.

$$0.4\left(\frac{\text{nr. words}}{\text{nr. sentences}}\right) + 100\left(\frac{\text{nr. complex words}}{\text{nr. words}}\right) \tag{2.3}$$

Words with three or more syllables are defined as complex words. The exceptions to this definition are proper nouns, familiar jargon, or compound words.

Ranking sentences by the readability metric or by the number of characters is roughly the same. A more detailed description of readability metrics is included in [129].

There are several datasets online [13] to train and test AKE methods. These datasets cover several domains with inherently different text document characteristics:

- ◇ **DUC-2001** [162, 81]): is a collection of 308 news articles annotated by Wan and Xiao [209]. Only the annotations are freely available.
- ◇ **Inspec** [88]: is a compendium of 2000 abstracts. The abstracts come from journal papers *Computer and Control, and Information Technology.*
- ◇ **NUS Keyphrase Corpus** [161]: contains 211 scientific conference papers. Each paper has one or more sets of key phrases assigned by the authors.
- ◇ **ICSI Meeting Corpus** [89]: includes 75 news meetings collected at the International Computer Science Institute in Berkeley during the years 2000-2002. This corpus was annotated with key phrases by Liu et al. [121] at the University of Texas. However, these annotations are not available. This dataset was also used in an evaluation study of unsupervised AKE performed at University of Texas [82].
- ◇ **FAO780** [147]: is a collection of 780 documents downloaded from the Food and Agriculture Organization (FAO) repository [17].
- ◇ **Citeulike180** [145]: is a collection of 183 papers obtained from the CiteULike.org bookmarking service, indexed by 332 annotators. However, the number of documents annotated per annotator ranges from 1 to 25 documents.
- ◇ **Schutz-2008** [190]: is a dataset of 1,323 articles spanning across 254 different journals published by PubMed Central, ranging from Abdominal Imaging to World Journal of Urology. It is a subset of the PubMed Central 3 corpus consisting of 77,496 peer-reviewed papers.
- ◇ **Krapivon-2009** [97]: contains 2,304 papers from Computer Science domain, which were published by the Association for Computing Machinery (ACM) in the period from 2003 to 2005.

◇ **SemEval-2010 Task 5** [95]: is a corpus released for the Task five of the Workshop on Semantic Evaluation 2010 in ACL 2010. It comprises a set of 284 papers with key phrases picked by both authors and annotators.

Table 2.1 summarizes the AKE corpora available before our work. There is a clear focus on the extraction of key phrases from scientific papers. This fact is justified by the availability of papers annotated with key phrases.

Table 2.1: AKE Corpora statistics.

Corpora	Domain	#Docs	#Tokens/ #Docs	#Keys/ #Docs	#Tokens/ #Keys
DUC-2001	News Articles	308	876	8.1	2.1
Inspec	Paper abstracts	2000	134	9.8	2.3
NUS	Full Papers	211	8291	11.0	2.1
ICSI	Meeting Transcripts	161	1611	3.6	1.3
Schurz-2008	Full Papers	1323	3720	6.9	5.0
Krapivon-2009	Full Papers	2304	7855	5.3	2.1
Citeulike180	Full Papers	183	6980	25.3	1.2
FAO780	Reports	779	29432	5.4	1.6
SemEval-2010	Full papers	284	8075	15.2	2.2

2.1.1 AKE evaluation

The standard evaluation metrics used in text classification tasks, such as AKE, are Precision (P), Recall (R), and F-measure (F_1). Precision is the fraction of key phrases correctly classified (true positives, tp) over all phrases classified as key phrases, i.e., the sum of tp with false positives (fp):

$$P = \frac{\#tp}{\#tp + \#fp} \tag{2.4}$$

Recall is the fraction of key phrases over the total number of key phrases that were successfully identified:

$$R = \frac{\#tp}{\#tp + \#fn} \tag{2.5}$$

F-measure combines the precision and recall in the following way:

$$F_1 = \frac{2PR}{R + P} \tag{2.6}$$

F_1 is a specific case of a more general formula when $\beta = 1$:

$$F_\beta = (1 + \beta^2)\frac{PR}{(\beta^2 P) + R} \tag{2.7}$$

The disadvantage of these metrics is that they only consider one reference. However, for datasets with several references (created by human annotators) and low agreement between them or to evaluate a ranked list of results like the one produced by AKE methods produce, it is more suitable to use the Normalized Discounted Cumulative Gain (NDCG) [90] metric [131]:

$$NDCG = \frac{DCG}{iDCG} \tag{2.8}$$

$$DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2 i} \tag{2.9}$$

Where rel_i represents the relevance score of each key phrase at rank i, i.e., the number of human annotators that selected a phrase as relevant. For normalization, DCG is divided by the ideal ordered list of key phrases (iDCG).

2.2 Event Detection and Tracking

The earliest work on *Event Detection* proposed a rule-based system to classify news documents into event types [83]. In 1992, Massand et al. [139] replaced the ruled-based system by a supervised classifier (K-Nearest Neighbors - K-NN [53]).

In the late 1990s, the event detection problem was also investigated under the Topic Detection and Tracking (TDT) effort [27, 226, 46, 223]. The TDT project was organized into two primary tasks: First Story Detection, or New Event Detection (NED), and Event Tracking. The goal of the NED task was to discover documents covering breaking new events in a news stream. The other task, Event Tracking, was focused on the tracking of articles describing the same event or topic over time. More recent work using the TDT datasets focused on Event Threading consisting of tracking and linking several related events. Current work [154, 63, 87] tried to organize news articles about armed clashes into a sequence of events, but still assumed that each article described a single event. Another related type of task, Passage Threading [63], extends Event Threading by relaxing the one-event-per-news-article assumption and using a binary classifier to identify "violent" paragraphs.

Even though the TDT project ended in 2004, new event detection research followed. Automatic Content Extraction (ACE) is the most pertinent example for this work. The goal of ACE Event task is the detection of events at sentence level. In addition to the identification of events, the ACE 2005 [208] task identifies the participants, relations, and attributes of each event. This extraction is an important step towards the overarching goal of building a knowledge base of events [91].

Event datasets are usually composed by several news articles. Experts defined a list of event types and annotated each sentence of the news articles. In practice, only a few sentences contain these types of events. Frequently, these sentences describe only one event and are complemented with other sentences describing other unrelated events (a type of event not included in the list) or "no events", such as a dateline, leading to an imbalanced dataset. As a result, it is hard to obtain good classification results in these imbalanced datasets with few examples of events. There are several ways to address this problem: namely, to increase the number of examples of events through bootstrapping techniques, or augmenting the event-labeled dataset by including documents from other collections (cross-document techniques) such as MUC-6 (Message Understanding Conference, edition 6) [109, 110, 87]. Other works explore a wide range of features using supervised classifiers [155]. Generally, a drawback of these approaches is that performance rapidly decreases as the total number of event types or labels increases. In fact, for multi-label document classification in large datasets, probabilistic generative methods can outperform discriminative methods such as Support Vector Machines [182]. For small datasets, such as the ACE 2005, there is not enough data to successfully learn a traditional supervised classifier, such as SVM. The alternative for these cases is to prefer the fuzzy fingerprints [86, 181, 138] method detailed in Section 4.1.3, over the traditional supervised classifier. This method not only has low computational requirements for both training and classification, especially when compared to the other methods. In addition, it is able to detect examples of all event types. This is particularly important for this work as it will have an impact in the filtering of sentences without events before generating summaries.

An alternative method to track events is to use Probabilistic Topic models. These models were designed to work in collections of documents, facing limitations when used for individual documents or sentences. Another limitation of Probabilistic Topic models is that they do not model the structure of sequences of events or topics as research in News Threading, Timeline Generation, and Temporal Summarization do. The following sub-sections review each approach separately and we will conclude the section with the evaluation metrics.

2.2.1 Probabilistic Topic Tracking

Probabilistic topic models are a class of statistical models in which the semantic properties of words and the documents are expressed in terms of probabilistic topics. They are an evolution of statistical methods such as Latent Semantic Analysis (LSA) (that also claimed that semantic information can be derived from a word-document co-occurrence matrix and that dimensionality reduction is an important part of this derivation).

More precisely, in Topic Models, such as Probabilistic Latent Semantic Analysis (PLSA) [85] and Latent Dirichlet Allocation (LDA) [39], documents are mixtures of topics, where a topic is a probability distribution over words, regardless of the event-like characteristics of the topic (purity) or a clear definition of the number of topics. Several papers have proposed extensions and modifications to LDA, to model the temporal dimension of the information [38, 214, 23]. Blei and Laferty [38] analyzed how topics evolve over time in the Science Journal with dynamic topic models (DTM). Topics over Time (TOT) [214] is an extended version LDA model with a time dimension extension. Ahmed and Xing [23] generalized DTM to infinite DTM (iDTM), also in the topic tracking of scientific literature. However, Ahmed et al. [22, 21] found that previous models, such as DTM and TOT, have not been shown to successfully model rapidly-changing corpora such as news or blogs. Structured Determinantal Point Processes (SDPP) [70] overcame the time complexity of DTM methods (seventy-five times faster).

Kim and Oh [94] identified one of the most important limitations of Probabilistic Topic models. The methods only model change of word distributions of the topics. They assume that the set of topics stays constant thought time, so it does not model the appearance and disappearance of topics over time.

This means that Probabilistic Topic models do not model the structure of the sequence of events or topics. One example of sequence of events is the Restaurant Script in the Book of Schank and Abelson [189], where a customer goes to a restaurant, orders something, eats it, pays, and leaves. The idea of filling in scripts is to have a set of defined actions that only some information varies, such as food and bill. This information is stored in a database or knowledge base. There were two research trends to fill templates (simple scripts with one action): MUC [3, 75], ACE [1], and TAC [14] research trends. The idea of temporal sequence is also explored in recent work on news threading, and timeline generation.

2.2.2 News Threading, Timeline Generation, and Temporal Summarization

Threading is the automatic discovering of connections between documents. This task has been initially proposed in the email domain [106] where the task is simpler because there is a

strong structure of referenced messages, which means they share the same header or include a copy of the previous messages.

Contrary to emails, in the news domain a well-defined linking structure does not exist, making news threading a complex task. Also, the presentation of threading results is made in different ways, for example; forms, timelines, graphs, tag clouds, summaries, or combinations of these representations. Swan and Allan [197] projected the automatic construction of news timelines by extracting clusters of noun phrases and named entities. Later, Allan et al. [28] used the timelines to introduce the task of *temporal summarization*, which takes a stream of news articles (timeline) on a particular topic and extracts one sentence (per date).

Chieu and Lee [50] considered the generation of event timelines based on queries and a single thread of events. In addition, their method also explored the existence of bursts of news events, which is not visible in the TDT datasets.

The TDT datasets were also used to introduce *Event Threading* [154, 63], which organizes news articles into sequences of events, but still assumed that each article described a single event. *Passage Threading* [63] extended the event threading concept by relaxing the one event per news article assumption and used a binary classifier to identify paragraphs containing "violent" events.

With Café, Yang et al. [224] extracted information over temporally sequenced documents from the TDT4 corpus to answer queries. The answers are "information nuggets" as designated in NIST's TREC-QA evaluations [206], which correspond, in general, to one or several key phrases that answer a complex information need or question. The method relies on user feedback to select the nuggets, adaptive filtering (logistic regression classifier) to evaluate relevance against the query, and Maximal Marginal Relevance (MMR) to remove redundancy.

With Connecting-the-Dots, Shahaf and Guestrin [192] proposed a method to generate individual threads of news articles from The New York Times. These threads have well defined start and end points, where a "point" is a news document). Their method combines a linear programing framework over a bipartite graph representation of documents and words. They claim that their methods allow to optimize the relevance, coherence, and redundancy of the individual threads.

Recently, Shahaf et al. [193] proposed metro maps, a metaphoric railway map of intersecting concepts from threads where time order is relaxed. The Connecting-the-Dots method motivated Zhu and Oates [236] to study the problem and consider pruning the least relevant and redundant "dots" (documents) during random walks in bipartite graphs.

Leskovec et al. [105] explored event tracking in the social media domain. In this domain, the events tracked were short, distinct phrases ("memes"), e.g., "lipstick on a pig" and "our

entire economy is in danger". This research observed two interesting temporal patterns. A lag of two hours and a half between peaks of memes is one pattern. The other pattern is a burst or heartbeat behavior in the publication of memes. The emergence and popularity of microblogs, such as Twitter [16, 113] and Weibo [18], motivated the exploration of timeline generation for microblogs. The unique characteristics of microblogs, that is, the small size of their messages (maximum number of characters is 140), hashtags (words or phrases prefixed with a "#" sign), and mentions ("@" sign followed by a username is used for replying or referring to other users).

Both news articles and microblogs frequently include illustrative images. These images provide additional context. There are some initial efforts to combine text and images in the generation of timelines [219, 221]. The timelines have also been generated to forecast some future events [175, 176].

The intersection between timelines and temporal summarization motivated the creation of ETTS (Evolutionary Trans-Temporal Summarization) and ETS (Evolutionary Timeline Summarization) [222, 220] methods. Their goal is to generate individual, but correlated, summaries for several dates along a timeline created about a user query, such as "Obama", or "BP Oil".

Other lines of research distilled temporal expressions from documents [30, 31, 93] to improve clustering, timeline generation, and visualization of search results. The recent formalization of the ISO-TimeML specification language [168] to represent temporal information, enabled the creation of standard evaluation corpora (e.g., TempEval-2 [205]) and temporal taggers [47].

2.2.3 Event Detection Evaluation

Just like other text classification tasks, event detection is typically evaluated using: Precision (P_i), Recall (R_i), and F-measure ($F1$). The disadvantage of these metrics is in the sensitivity to imbalanced distribution of events. For instance, it is possible to obtain high F-measure values while still failing to detect a relevant number of event types. To overcome this limitation, Kubat et al. [98] proposed the G-mean metric (Equation 2.10) to evaluate generic imbalanced binary classification problems. The extension of G-mean to imbalanced multiclass classification problems was proposed by Sun et al. [196]. G-mean is defined as the geometric mean of the recall values R_i, and therefore has the disadvantage of assuming the value zero when at least one recall value Ri is zero. To overcome this limitation, we introduce a smoothing G-Mean version, the SG-Mean (Equation 2.11) [138]. A smoothing constant (e.g., $\delta = 0.001$) added to each R_i solves the problem of multiplication by zero if a class is not detected. With this metric it is possible to evaluate the performance of a method while still considering the loss of classes. In some situations, such as when the number of elements of

some events is greater or when the detection of some events are more important than other, it might be useful to weight the recall values. For this purpose, we proposed WSG-Mean (Equation 2.12), where w_i is the percentage of sentences belonging to an event i over the total number of sentences.

$$G\text{-}Mean = (\prod_{i=1}^{n} R_i)^{\frac{1}{n}} \tag{2.10}$$

$$SG\text{-}Mean = \left(\prod_{i=1}^{n} (R_i + \delta)\right)^{\frac{1}{n}}, \ \delta > 0 \tag{2.11}$$

$$WSG\text{-}Mean = \left(\prod_{i=1}^{n} (R_i + \delta) \times w_i\right)^{\frac{1}{n}}, \ \delta > 0 \tag{2.12}$$

To complement these metrics, we include the number of classes that the methods fail to detect ($\#R_i = 0$).

2.3 Extractive Summarization

Automatic Text Summarization is the process of reducing one or more texts to the essential information presented in a shorter text - a summary. Several methods have been created to generate summaries. There are two approaches to automatic summarization: *extractive* and *abstractive* summarization. Extractive summarization methods are primarily concerned with what the summaries content should be, relying exclusively on the retrieval of sentences. Abstract summarization methods aim to create summaries closer to what humans generate. Creating such summaries involves paraphrasing the original documents. Paraphrasing is a complex Natural Language generation task. Most research has been focused on extractive summarization methods, so the state-of-the-art abstractive summarization methods are still very weak. By weak we mean that they create summaries with poor paraphrasing, grammatical errors, and missing relevant phrases or topics. In general, the abstractive summarization methods are not applied to speech corpora, such as Broadcast News, because the accumulation of speech recognition errors with abstractive summarization could render the final summary useless.

In addition, some abstractive summarization methods use extractive summarization methods as pre-processing steps.

Besides the division of automatic summarization into extractive and abstractive methods, it

is also common to characterize automatic summarization according to the following categories [123]:

- ◇ *Query-oriented* versus *Generic* - A query-oriented summary favors specific topics or passages of the original text, in response to user's information needs encoded in queries. A generic summary gives the same importance to all major topics in the original documents.

- ◇ *Single-Document* versus *Multi-Document* - Single-document summarization methods generate summaries from one input document, though the summarization process itself may use information obtained from other documents. Multi-document summaries produce summaries using two or more input topic-related documents.

- ◇ *Input/Output source(s)* - The media type of the input documents and output summaries has a strong influence in the summarization methods. Whether the source of information is a clean text or a noisy speech source, influences the complexity of the methods. While text summarizers use syntactic [204] and semantic information [202], depending on the amount of speech recognition errors, the syntactic and semantic information loses importance in speech summarization. Speech-specific information (for example prosodic features [140], recognition confidence [227]) become more important. The number of input sources and languages also have impact in the summarization methods.

- ◇ *Interactivity* - Most summarization systems do not rely on users to improve the quality of the summary. Few systems use user feedback. Different types of user feedback include asking users to identify the most important sentences during an iterative summarization process [119], or the identification key words/phrases to increase their importance in the summarization [148, 179].

- ◇ *Other categories* - During the last 20 years, the National Institute of Standards and Technology (NIST) ran several summarization evaluations, much as SUMMAC [2], TREC [15], DUC [4], and TAC [14]. These evaluations created new training and evaluation datasets, and new summarization tasks, including Update Summarization and Opinion Summarization.

  The goal of the *Update Summarization* task, introduced in the last year of DUC (2007) and continued in TAC, is to create "short (100-word) multi-document summaries under the assumption that the reader has already read some previous documents". TAC (in 2008) included a new task called *Opinion Summarization*.- It had as objective to generate summaries of opinion posts from blogs.

A more detailed and extended discussion about these categories and corresponding summarization methods can be found in several automatic summarization surveys [55, 157, 158, 103].

These surveys also indicate that the very first automatic summarization works [127, 35] were published in 1958. These works, done at the International Business Machines Corporation (IBM), proposed both word frequency and sentence position to represent technical documents information.

Extractive summarizers execute three tasks. The first task is to obtain a representation of the input document, such word vectors with TF-IDF values. The remaining two tasks are, respectively, scoring and retrieval of sentences/passages. In the second task, each sentence/-passage receives a score that represents its importance. In the final task, the summarizer has to select the best combination of sentences/passages to form a summary.

These three tasks are recurrent in the following family of summarization models: Maximal Marginal Relevance, Centrality, and Two-stage extractive summarization methods.

2.3.1 Maximal Marginal Relevance Family

Maximum Marginal Relevance (MMR) [45] is a query-oriented summarization method. It works by iteratively selecting the sentence S_i from input document(s) D that maximize the following equation:

$$MMR = \underset{S_i \in D \setminus U}{\operatorname{argmax}}[\lambda Sim_1(S_i, Q) - (1 - \lambda) \max_{S_j \in D}(Sim_2(S_i, S_j))] \tag{2.13}$$

where λ is a linear combination parameter ranging between $[0, 1]$ that simultaneously rewards relevant sentences and penalizes redundant ones; Q is a query or user profile; U is the set of sentences already selected; $D \setminus U$ is the set difference; Sim_1 and Sim_2 are two similarity measures, which are commonly set to the standard vector space cosine similarity:

$$Cos(\theta) = Sim(V_1, V_2) = \frac{V_1 \cdot V_2}{\parallel V_1 \parallel \parallel V_2 \parallel} = \frac{\sum_i V_{1i} V_{2i}}{\sqrt{\sum_i V_{1i}^2} \times \sqrt{\sum_i V_{2i}^2}} \tag{2.14}$$

The popularity and adaptability of the MMR method originated a family of methods. The Portfolio Theory [211] is one recent example based on the idea of ranking under uncertainty. Another example is the extension of the MMR as a probabilistic model (Probabilistic Latent MMR) in the the Expected n-call@k [78, 187] method.

2.3.2 Centrality Family

Centrality is a family of summarization methods that identify the most important content based on the detection of the most central content of the input source(s). Ribeiro and

Matos [177] divided this family into two sub-families: centroid-based summarization and pair-wise passage similarity-based summarization.

Centroid-based summarization methods define a central point or centroid. The centroid is defined by the passages of the input source(s) in a geometrical representation space. The centroid is also called topic signature in the literature [116]. The summary contains the passages that are closer to the centroid.

Raved et al. [171, 172, 173] built MEAD [12], the first centroid-based summarizer for multi-document summarization. This method creates clusters of documents using $TF \times IDF$ vectors (vector space model). The weighted averages of the vectors in each cluster represent the centroid. The centroid value (C_i), position value (P_i), and first sentence overlap value (F_i) are linear combination features used by MEAD sentence's (s_i) score function:

$$score(s_i) = w_c C_i + w_p P_i + w_f F_i, \quad 1 \leq i \leq N \tag{2.15}$$

where centroid value C_i is the sum of the centroid values $C_{w,i}$ of all words in the sentence S_i:

$$C_i = \sum_w C_{w,i} \tag{2.16}$$

position value P_i is given by (with C_{max} as a constant value):

$$P_i = \frac{(n - i + 1)}{n} \times C_{max} \tag{2.17}$$

Finally, the first sentence overlap value (F_i) is the inner product of sentence s_i and the first sentence of the document:

$$F_i = \vec{S_1}\vec{S_i} \tag{2.18}$$

The highest scored passages, according to Equation 2.15, define the summary.

In the pair-wise passage-based similarity, each passage is scored against every other passage or a defined set of passages. Graph-based methods, such as LexRank [61], and TextRank [149], that are based on PageRank [44], are some of the most best known examples. These methods build a graph representation of the source document(s) having the passages as vertices with edge connecting vertices when the similarity between sentences meets a certain threshold value.

The Centrality-as-Relevance method, or simply Centrality, as described by Ribeiro and de Matos [177], is based on the notion of support set: after dealing with the representational aspects, the first step of the method is to compute a set consisting of the most semantically related passages, designated support set. Then, the most important passages are the ones that occur in the largest number of support sets.

Given a segmented information source $I \triangleq p_1, p_2, ..., p_N$, a support set is computed for each passage p_i (Eq. 2.19, $sim()$ is a similarity function, and ε_i is a threshold).

$$S_i \triangleq \{s \in I : sim(s, p_i) > \varepsilon_i \wedge s \neq p_i\} \tag{2.19}$$

Passages are ranked in accordance with Eq. 2.20.

$$\underset{s \in \cup_{i=1}^{n} S_i}{\arg\max} \left|\{S_i : s \in S_i\}\right| \tag{2.20}$$

Another set of interesting centrality models is proposed by Kurland and Lee [101, 102] for re-ranking a previously retrieved set of documents. Their model is based on the notion of top generators of a document that is similar to our concept of support set. The definition of the top generators of a document is based on a k-nearest-neighbor (kNN) approach using generation probabilities, while in our model the definition of cardinality of the supports sets can be seen as resulting from a generalization of both kNN and εNN approaches (threshold-based), since each support set can use differentiated thresholds (ε_i, Equation 2.19). Additionally, we base semantic similarity on geometric proximity.

2.3.3 Two-stage extractive summarization methods

Two-stage extractive summarization methods divide the problem of retrieval of important passages into two steps. In the first step, the methods identify important words or phrases, which are used to improve the selection of passages performed in the second step. The second step ranks the passages.

Our two-stage KP-Centrality method, presented in Chapter 3, combines supervised AKE and the Centrality-as-relevance method. Closely related to our work two-stage KP-Centrality method are the unsupervised key phrase extraction approaches that have been explored to reinforce summarization methods [228, 210, 120, 180, 195]. Litval and Last [120] and Riedhammer et al. [180] propose the use of key phrases to summarize news articles [120] and meetings [180] Litval and Last explore both supervised and unsupervised methods to extract key phrases as a first step towards extractive summarization. Moreover, our adaptation of the centrality-based summarization model plays an important role in the whole process, an inexistent step in their work. Riedhammer et al. [180] propose the method closest to ours: the first stage consists of a simple key phrase extraction step, based on part-of-speech patterns (unsupervised); then, these key phrases are used to define the relevance and redundancy components of an MMR summarization model. Beyond the differences in the key phrase extraction (supervised vs. unsupervised), our method differs from Riedhammer et al. [180] in the way the key phrases are used. Our two-stage KP-Centrality method does not restrict

key phrases as the unique document representation, instead it complements the bag-of-words representation by using key phrases as additional sentences or passages.

Recently, there is also our work [135, 131] exploring the use of extractive summarization to guide automatic key phrase extraction (detailed in Section 3.1).

2.3.4 Summarization Evaluation

The most widespread automatically summary evaluation metric is ROUGE [114, 117]. ROUGE compares human reference summaries with the automatic generated summaries measuring n-gram co-occurrences. The reduction of complexity of text summarization evaluation and high-level of correlation with manual evaluations are the reasons behind the wide adoption of ROUGE. The metric measures the percentage of n-gram matches between a set of reference summaries (human produced) and a automatically generated summary. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) can be used in the following five ways:

- ◇ *ROUGE-N* is an n-gram recall co-occurrence statistic. N is usually 1 (*ROUGE-1* - uses unigrams) or 2 (*ROUGE-2* - considers bigrams). *ROUGE-N* is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{2.21}$$

  where n stands for the n-gram length; $Count_{match}(gram_n)$ is the maximum number of times $gram_n$ co-occurs on both generated summary and reference summaries. As noted by Lin [114], ROUGE-N is a metric related to BLEU [164], frequently used in automatic machine translation, which is precision-based. The rationale of using recall for summarization is to analyze information coverage, while for translation what is relevant is to measure the number of words translated correctly.

- ◇ *ROUGE-S* uses skip-bigrams, that is, it considers any pair of words in their sentence order.

- ◇ *ROUGE-SU* uses both skip-bigram and unigram co-occurrence statistics.

- ◇ *ROUGE-L* uses the longest common subsequence (LCS) to count the number of matches.

- ◇ *ROUGE-W* is a weighted version of LCS.

ROUGE is an example of automatic summary evaluation metric with models, where the models are human summaries. These summaries are also used in the Pyramids evaluation

method [160, 159], BLEU [164], ParaEval [234], cosine similarity [58], Kullback–Leibler Divergence (KLD) [115], and Jensen-Shannon Divergence (JSD) [115].

The last two metrics introduced an automatic summarization evaluation with models based on distances (divergences) between two probability distributions of words. They are defined in the following way:

- ◇ KLD [99] between two probability distributions (P and Q, where P represents the automatically-produced summary and Q represents the reference, human-produced summary) is defined as follows:

$$KLD(P||Q) = \sum_{w} P(w) log \frac{P(w)}{Q(w)} \tag{2.22}$$

  Since the KLD is not symmetric, both divergences between the pairs of probability distribution source-document/summary and summary/source-document can be used as metrics. Moreover, KLD is undefined when $Q(W) = 0$ and $P(w) > 0$. But the simplest type of smoothing [48], called Additive Smoothing [111], can be used to overcome the undefined problem.

- ◇ JSD [118] is a symmetrized and always defined version of the KLD (Equation 2.22). It is given by the following equation:

$$JSD(P||Q) = \frac{1}{2}[KL(P||A) + KL(Q||A)] \tag{2.23}$$

  where $A = \frac{P+Q}{2}$ is the mean distribution of P and Q.

The JSD has always outperformed the KLD [115] in the summarization evaluations.

Recently, Louie and Nenkova [125] proposed the use of the cosine similarity, KLD, and JSD evaluation metrics for automatic summary evaluation without human model summaries. The evaluation compares the content-based probability distribution of words between the automatic generated summaries and source documents. Among the three, the Jensen-Shannon divergence was the most reliable metric, because it has a stronger correlation with Pyramids and Responsiveness (human-based evaluation that indicates how well the summary satisfied a given information need in a scale of one to five [163]). One year later, Saggion et al. [184] extended the evaluation of Jensen-Shanon divergence metric without models for other summarization evaluation tasks. JSD without models showed moderate to high correlation to ROUGE for French and Spanish documents. But it was found to be unreliable for more complex summarization tasks, such as biographical [235] and opinion summarization [233].

2.4 Event-Based Summarization

Although most summarization approaches are for the news domain, in which events are the most important concepts, very few works have attempted to combine event information and summarization towards event-based summarization.

Daniel et al. [54] proposed the first event-based summarization approach. However, it is not automatic, since it needs a human to annotate the relevance score of sentences for all topic sub-events. The sub-events are topics or document-level events ignoring sentence-level events mentions. The sentences with the highest sum of relevance score over all sub-events generate the best summaries.

Filatova and Hatzivassiloglou [64] proposed sentence-level events to improve summarization. They defined an atomic event as a triplet composed by a named entity, a verb or action noun, and another named entity, where the verb/action noun defines a relation between the two named entities. However, this definition excludes important events that to not have more than one named entity, such as "In the capital, Tripoli, cars clogged the city centre." [6] or events that are described using more complex grammar structure: "The number of people with Ebola in west Africa has risen above 16,000, with the death toll from the outbreak reaching almost 7,000, the World Health Organisation (WHO) says." [20] "Croatia has become the 28th member of the European Union, with crowds joining celebrations in the capital Zagreb" [9]. The event information is used to minimize redundancy since sentences without events are excluded from the summary.

Li et al. [108] extended the Filatova and Hatzivassiloglou [64] method by using PageRank [44] to perform sentence selection. To apply PageRank, they build a co-occurrence graph of named entities and other event terms (e.g., verbs). As event terms express the semantic relations between named entities, Liu et al. [122] and Zhang et al. [232] proposed to cluster event terms to identify similar events. More recent work [71] explores temporal relations between individual events. They propagate the importance of very important events to other past or future events. Others have tried to use time information and word overload to summarize the same events [37, 36]. In Chapter 4, we will present our event representation for sentences. We use a vector where each entry models the likelihood of a sentence describe an event type, such as meet.

3 Single-document Summarization

> *Most people don't have time to master the very mathematical details of theoretical physics*
>
> Stephen Hawking, *English Physicist (1942-)*

The information overload caused by the massive amount of content published nowadays calls for methods to retrieve the desired information automatically [186, 92]. Extractive single-document summarization is a task where the most important parts of a document are selected to produce a human comprehensible summary of that document. Several approaches have been proposed to detect the most salient content: approaches based on significance measures, such as LSA [72, 153], graph-based relevance [69], or MMR [45, 218]; approaches based on classification, which formulate the sentence selection process as a binary classification problem, where the method has to decide if the passage should be included in the summary or not [68, 217]; and, approaches based on passage position [51, 84]. The challenges in this field can be placed into three major areas: the efficiency of the methods to keep up with the continuously increasing rates of information generation; the adaptability to different types of information sources, such as video and audio; and, the effectiveness of the method in terms of the quality of the retrieved information. In this work, we address the third challenge, as our main goal is to improve the quality of the retrieval using the standard ROUGE [114] evaluation metric. However, we are conscious of the importance of efficiency and show that our algorithm can be tuned for speed. While we do not create a method for audio data specifically, we present results on both textual and speech transcriptions sources using the ROUGE evaluation metric.

PageRank [44] is among the most popular retrieval models for the extractive summarization task [61, 101, 33, 102, 66]. This model belongs to the family of centrality methods where the most salient items are the most central ones, under a representation where such notion makes sense (graph, spatial). Centrality-based methods [41, 40, 24, 25, 26, 172, 61, 149, 101, 177] detect the most salient passages by computing the central passages of the input source(s). One of the main members of this family is centroid-based summarization [173]. Centroid-

based methods take advantage of the idea of creating one pseudo-passage that represents the central topic of the input source. The passages included in the summary are the ones closer to the pseudo-passage or centroid. The distance between the passages and the centroid is obtained using the cosine similarity. Another approach to centrality estimation is to compare each candidate passage to every other passage and select the ones with higher scores (the ones that are closer to every other passage). One simple way to do this task is to represent passages as vectors using a weighting scheme like TF-IDF. Then, passage similarity can be assessed using, for instance, the cosine similarity, assigning to each passage a centrality score. These scores are then used to create a sentence ranking: sentences with highest scores are selected to create the summary.

Since this kind of models treats all passages equally [177], estimating either centroids or average distances between input source passages, we may be selecting extracts that, being central to the input source, are, however, not the most important ones. The degree of centrality, that is, the number of links or similar passages is an approximation of the concept of importance in all centrality-based methods. In cognitive terms [60], the summarization process relies on the removal of least relevant, information. This means that it is common to find, in the input sources to be summarized, inadequate content, secondary topics, or irrelevant information. These aspects affect centrality-based summarization methods by inducing inadequate centroids or decreasing the scores of more suitable sentences. For instance, an article about a start-up acquisition may contain a small paragraph about the start-up's history. In this case, the main topic would relate the acquisition itself, and a secondary topic would include the history of the start-up. Since the centrality-based methods cannot distinguished between main and secondary topic, they use the degree of centrality of the topic for the estimation of importance. Obviously, the number of topics within a document is not pre-determined and may vary. For instance, a document may contain one primary topic and two secondary ones, or even two main topics. Furthermore, the notion of topic is not strictly defined as the same passage may contain artifacts from different topics. In this situation, current methods are suboptimal for one of the two following reasons: firstly, centroid-based methods [172] build the summary using only the most relevant topic due to the larger term frequency of the terms of that topic and ignore the secondary topic; secondly, centrality-based methods [44, 61, 177] place equal importance on extracting passages from both topics for the summary, which is undesirable, as the final summary may contain an excessive amount of passages for the secondary topic.

We partially addressed this problem in previous work [178], which we now present here in an extended format. We present the biased centrality model, which is an extension of the model presented by Ribeiro and de Matos [177] to perform generic single document extractive summarization. This biased method improves the importance of the different passages by

initially extracting a set of key phrases that is used to guide the underlying summarization model. Key phrases are the most important words or contiguous sequences of words in a document. We explored several ways to integrate key phrases in centrality methods. The most successful integration uses the key phrases as pseudo-passages. This approach effectively reduces the excessive number of passages from secondary or irrelevant topics.

To evaluate the success of our methods, we use the standard summarization metric ROUGE [114]. Results show that the use of key phrases yields significant relative improvements ranging between 17% and 31%, across different languages and datasets. Additional statistically significant relative improvements, ranging between 1% and 4%, over the biased method were obtained by the new iterative method. Interestingly, the summaries generated by the new method are smaller, in terms of number of words.

The rest of the chapter is organized as follows: Section 3.1 introduces our work on automatic key phrase extraction that is included in our two-stage single-document summarization method presented in Section 3.2.

3.1 Automatic Key Phrase Extraction

Summarization systems need high level descriptions of news for selecting the most important content. Fast and effective automated indexing is a critical problem for such services. Key phrases that consist of one or more words and represent the main concepts of the document are often used for the purpose of indexing. The precision and F-measure of current state-of-the-art AKE is in the 30-50% range [137, 146, 216]. This makes improvements in AKE an urgent problem.

In this work, we followed a fairly traditional approach of training a classifier to select an ordered list of the most likely candidates for key phrases in a given document. The main novelty of our AKE method is the use of additional semantic features and pre-processing steps. We tested several features, which, to the best of our knowledge, have not been used for this purpose. These features include the use of signal words and freebase categories, among others. Some of these features lead to significant improvements in accuracy. We also experimented with two forms of document pre-processing that we call light filtering and co-reference normalization. Light filtering removes sentences from the document, which are judged peripheral to its main content. Co-reference normalization unifies several written forms of the same named entity into a unique form. In our experiments, both light filtering and co-reference normalization lead to noticeable improvements in the resulting accuracy of key phrase extraction.

We also needed a set of labeled documents for training and evaluation (gold standard). We

used the AMT service to obtain these documents for English [131].

3.1.1 Features

In general, classifier-based key phrase extraction methods use vector space models with TF-IDF [185] over words. Other commonly used features are the phrase position on the page (first, current, and last) [216], distance between the last and first occurrence of the phrase, number of words in the phrase [146], Part-Of-Speech (POS) tags [137], among others. Our baseline set of features included the commonly used ones and an initial level of shallow semantic features:

1. the number of characters in the phrase - empirically, nouns that are long tend to be relevant.
2. the number of named entities - very often, named entities are important key phrases; typically, this number is zero, one, or two.
3. the number of capital letters - the identification of acronyms is the main reason to include this feature.
4. the POS pattern of the phrase (e.g., <noun>, <adj, noun>, <adj, adj, noun>, etc.) - <noun> and <noun phrase> are the most common patterns observed in key phrases, <verb> and <verb phrase> are less frequent, and the remaining POS tags are rare.
5. the frequency of the phrase in a 4-ngram language model using the HUB4 dataset [74]. The model was compressed using the Minimal Perfect Hash method [79], implemented in the smooth-nlp toolkit [8], to optimize the computational performance.

We used two additional kinds of features: semantic and rhetorical. We used two levels of semantic features: top-categories and sub-categories. The top-categories we used are the following: *Technology, Crime, Sports, Health, Art and Culture, Fashion, Science, Business, World Politics,* and *U.S. Politics*. We also used 85 sub-categories taken from the Freebase domain names [11]. These included *American Football, Baseball, Book, Exhibitions, Education Engineering, Music,* among others. Both the top-categories and the sub-categories are used as binary features of a phrase. The top-category of each phrase is obtained from the document source category and the sub-categories are extracted by matching the phrase against the Freebase dump.

Authors of news articles use various rhetorical devices to direct the reader's attention. The following eleven types of **rhetorical signals** have been identified in the literature [67]:

1. **Continuation** - there are more ideas to come, e.g., *moreover, furthermore, in addition, another.*

2. **Change of direction** – there is a change of topic, e.g., *in spite of, nevertheless, the opposite, on the contrary.*

3. **Sequence** – there is an order in the presenting ideas, e.g., *in first place, next, into (far into the night).*

4. **Illustration** – gives an example, e.g., *to illustrate, in the same way as, for instance, for example.*

5. **Emphasis** – increases the relevance of an idea, these are the most important signals, e.g., *it all boils down to, the most substantial issue, should be noted, the crux of the matter, more than anything else.*

6. **Cause, Condition, or result** – there is a conditional or modification coming to following idea, e.g., *if, because, resulting from.*

7. **Spatial** – denote locations, e.g., *in front of, between, adjacent, west, east, north, south, beyond.*

8. **Comparison/contrast** – comparison of two ideas, e.g., *analogous to, better, less than, less, like, either.*

9. **Conclusion** – ending the introduction of the idea and may have special importance, e.g., *in summary, from this we see, last of all, hence, finally.*

10. **Fuzz** – there is an idea that is not clear, e.g., *looks like, seems like, allegedly, maybe, probably, sort of.*

11. **Non-word emphasis** - all visual signals that indicate emphasis and are not words, e.g., exclamation mark (!),"quotation marks".

We hypothesized and confirmed that sentences containing such signals are more likely to contain key phrases. We used each of these eleven types of signals as a feature of a phrase. The feature values are the number of signals occurring in the sentence containing the candidate key phrase.

3.1.2 Light Filtering

Light Filtering is the process of elimination of about 10% of low-relevance passages from the body of news documents. It is based on assigning a measure of relevance to each sentence of the article using centrality-as-relevance methods [177]. Centrality-as-relevance calculates pair-wise distances between sentences and finds a centroid for the article. The K sentences closest to the centroid are called the support set (SS). The distance between a sentence and the

support set is used as a measure of this sentence relevance. Based on our previous experiments, we used five support sentences per document and removed the 10% of the most distant sentences from all documents using the Euclidean distance (x and y are vectorial sentence representations and n designates the length in number of words of the longest sentence):

$$D_{euclidean} = |x - y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3.1}$$

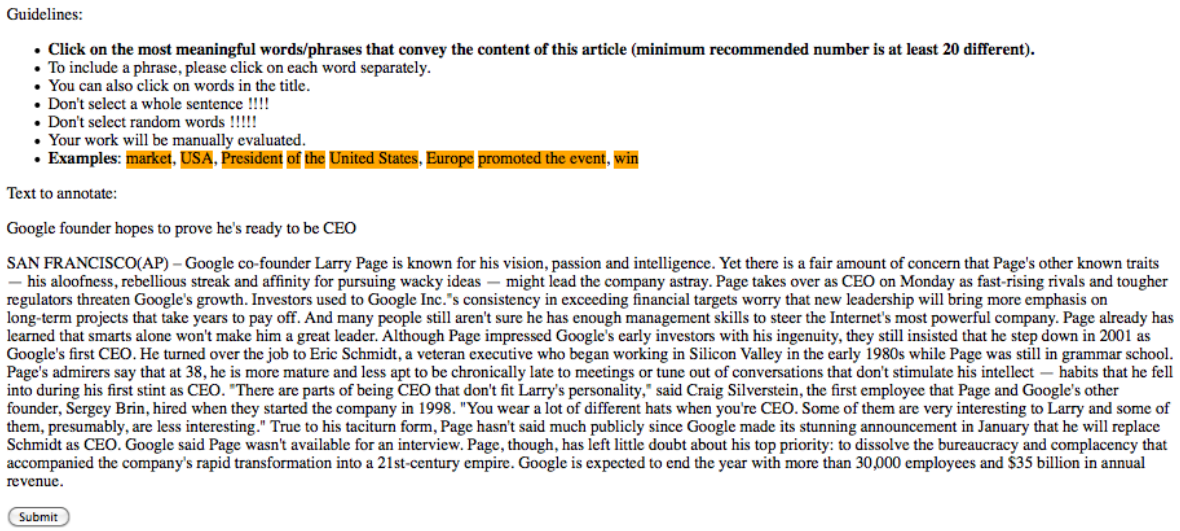
3.1.3 Co-reference Normalization

For stylistic reasons, journalists often use different forms of reference to the same named entities. For example, they might refer to Michael Jackson as Jackson or Michael. We hypothesized that normalizing such references would improve the AKE performance. We used ENCORE [191], a semi-supervised, ensemble co-reference resolution system to identify multiple forms of the same named entity and to normalize them into a single form (e.g., Michael Jackson).

3.1.4 English Gold Standard Corpus

To evaluate our hypotheses, we needed a set of news documents with the corresponding key phrases. The only dataset in the news domain available at the time these experiments were performed was DUC-2001 [10]. The disadvantage of this corpus is that the annotated list of key phrases does not completely describe all topics. It was, therefore, important to create a new dataset to address the issue just described. However, creating such a dataset presented both conceptual and practical difficulties. Designations of key phrases are subjective decisions of each reader with relatively little agreement among them. Our solution was to use multiple annotators for the same news document and assign to each phrase a score equal to the number of annotators who selected this phrase as a key phrase. Then, we ordered the phrases based on these scores and kept only the phrases selected by at least 90% of the annotators. We used AMT service to recruit and manage our annotators. To the best of our knowledge, this has not been done before for this purpose. Each assignment, called Human Intelligence Task (HIT), consisted of clicking on the most meaningful sequences of words in a news document. We provided several examples shown on Figure 3.1. Annotating one news document was a HIT and it paid \$0.02 if accepted. We selected 50 stories for each of the 10 categories and created twenty HITs for each of the 500 stories in our set. An individual annotator could only do one HIT per news document. Unfortunately, this creates a practical problem of uneven quality of annotators: some of the annotator used bad shortcuts to do a HIT, producing meaningless results. We used several heuristics to weed out bad HIT. For example, the inclusion of stop

words, very long sequences (limiting the length of key phrases to ten words), and very fast work completion (less than 30 seconds) usually indicated a bad HIT. As a result, we were able to keep 90% of HITs for each news document. We created a gold standard set of 500 annotated news documents. The average number of key phrases per story was about 39.72. This number includes all of the key phrases occurring in all good HITs. However, the average agreement between workers was only 55% (10 workers).


Guidelines:

- **Click on the most meaningful words/phrases that convey the content of this article (minimum recommended number is at least 20 different).**
- To include a phrase, please click on each word separately.
- You can also click on words in the title.
- Don't select a whole sentence !!!!
- Don't select random words !!!!!
- Your work will be manually evaluated.
- **Examples**: market, USA, President of the United States, Europe promoted the event, win

Text to annotate:

Google founder hopes to prove he's ready to be CEO

SAN FRANCISCO(AP) – Google co-founder Larry Page is known for his vision, passion and intelligence. Yet there is a fair amount of concern that Page's other known traits — his aloofness, rebellious streak and affinity for pursuing wacky ideas — might lead the company astray. Page takes over as CEO on Monday as fast-rising rivals and tougher regulators threaten Google's growth. Investors used to Google Inc."s consistency in exceeding financial targets worry that new leadership will bring more emphasis on long-term projects that take years to pay off. And many people still aren't sure he has enough management skills to steer the Internet's most powerful company. Page already has learned that smarts alone won't make him a great leader. Although Page impressed Google's early investors with his ingenuity, they still insisted that he step down in 2001 as Google's first CEO. He turned over the job to Eric Schmidt, a veteran executive who began working in Silicon Valley in the early 1980s while Page was still in grammar school. Page's admirers say that at 38, he is more mature and less apt to be chronically late to meetings or tune out of conversations that don't stimulate his intellect — habits that he fell into during his first stint as CEO. "There are parts of being CEO that don't fit Larry's personality," said Craig Silverstein, the first employee that Page and Google's other founder, Sergey Brin, hired when they started the company in 1998. "You wear a lot of different hats when you're CEO. Some of them are very interesting to Larry and some of them, presumably, are less interesting." True to his taciturn form, Page hasn't said much publicly since Google made its stunning announcement in January that he will replace Schmidt as CEO. Google said Page wasn't available for an interview. Page, though, has left little doubt about his top priority: to dissolve the bureaucracy and complacency that accompanied the company's rapid transformation into a 21st-century empire. Google is expected to end the year with more than 30,000 employees and $35 billion in annual revenue.

Submit

Figure 3.1: Example of AMT HIT used for creating the English AKE reference corpus.

3.1.5 Evaluation and Results

For our experiments, as a baseline, we used Maui [146] – a state-of-the-art supervised key phrase extractor based on a bagging over a C4.5 decision tree classifier [169]. The extraction of some shallow semantics features needs a Named Entity Recognizer (NER) and a POS tagger. We used the MorphoAdorner name recognizer [7] for Named Entity Recognition and the Stanford Log-linear POS tagger [201].

The probability of the phrase occurring in a text document is obtained from a 4-gram domain model - about 62K unigrams, 11,000M bigrams, 5,700M trigrams, and 4,000M 4-grams generated from the LDC HUB4 dataset [74].

In our experiments, we limited the number of extracted key phrases to 10. This made the calculation of recall irrelevant. Consequently, we used precision (P) and NDCG (Eq. 2.8) to evaluate the results.

Table 3.1 presents the AKE results with the new features and pre-processing steps using

Condition	NDCG	Precision
Baseline	0.7045	0.4790
Baseline + SS	0.7329	0.5301
Baseline + SS + TC	0.7504	0.5180
Baseline + SS + TC + RS	0.7657	**0.5430**
Baseline + SS + TC + RS + SC	0.7356	0.5140
Baseline + SS + TC + RS + SC + CN	0.7577	0.5278
Baseline + SS + TC + RS + CN + LF	0.7560	0.5170
Baseline + SS + TC + RS + SC + CN + LF	**0.7702**	0.5401

Table 3.1: Results of our AKE system when extracting 10 key phrases ($p - value < 0.05$) (SS - Shallow Semantics, TC - Top Categories, RS - Rhetorical Signals, SC - Sub-Categories from Freebase, CN - Co-reference Normalization pre-processing, LF - Light Filtering pre-processing).

10-fold cross-validation. The baseline corresponds to the Maui standard system, without the Wikipedia-based features because they did not improve the results of our preliminary experiments. A more detailed table containing the results of each shallow semantic and rhetorical signals features in provided in Appendix A.

3.1.6 Discussion

Our data indicates that the largest improvements in performance were due to shallow semantics features, top news categories, and rhetorical signals (NDCG 77.02% vs. 70.45%). The inclusion of Freebase sub-categories did not provide any beneficial improvements when used alone, but in combination with pre-processing it did cause slight improvements in the NDCG scores. It is interesting to compare our results with human performance. Since human annotators did not order their key phrases, we run hundred trials where we randomly ordered the key phrases for each annotator. Then, we computed the average NDCG score against the gold standard. The result was 64.63%, which is considerably lower than the system's performance. This may be due to the relative lack of agreement among human annotators and to sorting. Since the system is trained on the intersection of phrases (90% agreement among annotators), it seems to produce better results when measured against the weighted ordering. While the accuracy of automatic key phrase extraction may never be very high, we show in the following chapter that it is sufficient to improve the summarization by boosting the weights of more significant words and phrases when compared to the traditional TF-IDF scores.

3.2 Two-Stage Single-Document Summarization

To determine the most important sentences of an information source, we use the centrality model described by Ribeiro and de Matos [177]. The main reasons behind this choice are its adaptability to different types of information sources (e.g., text, audio, and video), its extensibility to other languages, and the state-of-the-art performance on both clean and noisy input documents.

The centrality (single-stage method) was extended with an initial step. This step corresponds to the extraction of key phrases. The idea motivating this additional step is that key phrases complement the bag-of-words model. We used the key phrases in three different ways to bias the centrality model.

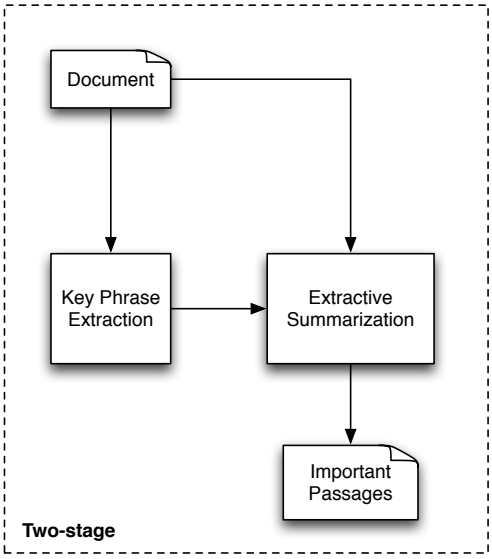


Figure 3.2: The flowchart of two-stage methods extractive summarization methods where the first stage is the extraction of key phrases and the second stage is the retrieval of important passages. KP-Centrality, CKP-Centrality and OKP-Centrality methods are examples of two-stage methods.

The centrality model is based on the notion of support set: after addressing the representational aspects, the first step of the method (Block "Extractive Summarization" in Figure 3.2) is to compute a set consisting of the most semantically related passages, designated support set (similar to the concept of cluster). Then, the most important passages are the ones that occur in the largest number of support sets.

A key factor that distinguishes this model from its predecessors is the fact that by allowing different thresholds (ε_i in Eq. 2.19) to each support set defined either manually or estimated using heuristics, centrality is influenced by the latent topics that emerge from the groups of

related passages. In the degenerate case where all ε_i are equal, the model behaves as the degree centrality model proposed by Erkan and Radev [61]. However, using a naïve approach of having dynamic thresholds (ε_i) set by limiting the cardinality of the support sets (a kNN approach), centrality is changed because each support set has only the most semantically related passages of each passage. From a graph theory perspective, this means that the underlying representation is directed, and the support set can be interpreted as the passages recommended by the passage associated to the support set. This contrasts with LexRank [61], which is based on undirected graphs. On the other hand, the models proposed by Mihalcea and Tarau [150] are closer to this centrality model in the sense that they explore directed graphs, although only in a simple way (graphs can only be directed forward or backward). Nonetheless, semantic relatedness (content overlap) and centrality assessment (performed by the graph ranking algorithms HITS [96] and PageRank [44]) is quite different from this model. Generally, methods based on directed graphs are directly or indirectly modeling the order or position of the passages or sentences in the original text. This order is particular important in some domains, such as the news, where the most important content appear in the beginning of the document.

The most important passages (summary) are the ones that appear in the largest number of support sets. Ribeiro and de Matos [177] explore several metrics to compute semantic relatedness and propose different ways to estimate the cardinality of the support sets. The cardinality or number of sentences in the support sets influences the number of support sets and has a direct impact in the centrality results. In this work, we use the heuristics based on passage order. This type of heuristics explores the structure of the input source to partition the candidate passages (to be included in the support set) into two subsets: the ones closer to the passage associated with the support set under construction, and the ones further apart (see Algorithm 1).

For our two-stage single-document summarization method (Figure 3.2), we adapted the model in three different ways. They are explained in Sections 3.2.1, 3.2.2, and 3.2.3.

3.2.1 OKP-Centrality

At stage 1 of Only Key Phrase-based Centrality (OKP-Centrality), we obtained key phrases ($K \triangleq k_1, k_2, ..., k_M$). At stage 2, we remove the passages that do not contain key phrases from the support sets, because they are less likely to be relevant. The idea behind considering only passages containing key phrases is to remove redundancy and at the same time to filter secondary topics. A formal definition of OKP-Centrality is as follows: given the subset of passages of I defined as $O \triangleq o_1, o_2, ..., o_K$, with $o_i \in I \wedge o_i \supset k_j, \quad j = 1, ..., M$, the resulting model is defined by Equation 3.2 and 2.20.

ALGORITHM 1: Generic passage order-based heuristic.

Input: Two values r_1 and r_2, each represents a subset, and the set of the passages p_k and corresponding distances d_k^i to the passage p_i in the support set under construction

Output: The support sets R_1 and R_2

$R_1 \leftarrow \emptyset$, $R_2 \leftarrow \emptyset$;
for $k \leftarrow 1$ *to* $N-1$ **do**
  if $|r_1 - d_k^i| < |r_2 - d_k^i|$ **then**
    $r_1 \leftarrow (r_1 + d_k^i)/2$;
    // e_k is a passage $R_1 \leftarrow R_1 \cup \{e_k\}$;
  else
    $r_2 \leftarrow (r_2 + d_k^i)/2$;
    // e_k is a passage $R_2 \leftarrow R_2 \cup \{e_k\}$;
  end
end
$l \leftarrow \arg\min_{1 \leq k \leq N-1}(d_k^i)$;
if $p_l \in R_1$ **then**
  return R_1;
else
  return R_2;
end

$$S_i \triangleq \{s \in O : sim(s, p_i) > \varepsilon_i \wedge s \neq p_i\}, \quad i = 1, ..., N \tag{3.2}$$

The ranking of passages of the OKP-Centrality is defined in Equation 2.20.

3.2.2 CKP-Centrality

Key Phrase Confidence-based Centrality (CKP-Centrality) models the influence of the key phrases by weighting the passages. The weights improve OKP-Centrality by reducing the importance of passages with less important key phrases, as well as those containing false positive key phrases. Those false positives key phrases typically obtain low confidence scores in the AKE nethod. As a result, the weights, used by the CKP-Centrality method, are the confidence scores obtained from the AKE method: $weight(p_i) = conf(k_j), p_i \supset k_j$ (Equation 2.19 and 3.3). High confidence values mean that the phrases are very likely to be key phrases, while low confidence scores reveals the opposite. Frequently, the most relevant key phrases also have high confidence scores. Then, we use the key phrase confidence scores as weights that approximate the relevance of passages.

As a result, passages are ranked in accordance with Equation 3.3.

$$\underset{s\in\cup_{i=1}^{n}S_i}{\arg\max}\Big(\sum_{s\in\cup_{i=1}^{n}S_i} weight(s) * \big|\{S_i : s \in S_i\}\big|\Big) \tag{3.3}$$

Both OKP-Centrality and CKP-Centrality approaches have in common the fact the input is a terms by passages matrix (Equation 3.4), but containing only the passages from the input source, as the key phrases influence the computation of the supports set as shown in Equations 3.2 and 2.19. For these approaches, passages not containing key phrases could be removed from matrix A if function w does not use them. Matrix A is a representation of all passages $p_i, ...; p_n$ in the document based on features (terms $t_1, ..., t_N$). This is representation is used to build the support sets and to rank the passages.

$$A = \begin{bmatrix} w(t_1, p_1) & \dots & w(t_1, p_N) \\ \dots & & \\ w(t_T, p_1) & \dots & w(t_T, p_N) \end{bmatrix} \tag{3.4}$$

3.2.3 KP-Centrality

In the KP-Centrality approach, key phrases $K \triangleq k_1, k_2, ..., k_M$ are considered regular passages, augmenting the number of support sets and, therefore, changing centrality: $I \cup K \triangleq q_1, q_2, ..., q_{N+M}$ (Equation 3.5 and 3.6). By augmenting the support sets with key phrases, we hope to bias the centrality method towards the most important topic(s). This means that we will have, not only always some relevant passages (either or both relevant regular and pseudo-passages, i.e., key phrases), but also reinforce the importance of the relevant regular passages in the support set with the passage being ranked by the centrality method (see Algorithm 2).

$$S_i \triangleq \{s \in I \cup K : sim(s, q_i) > \varepsilon_i \wedge s \neq q_i\}, \quad i = 1, ..., N + M \tag{3.5}$$

Passages are ranked excluding the key phrases K ("artificial passages") according to Equation 3.6. The input is represented as a terms by passages matrix (Equation 3.7), where w is a function of the number of occurrences of term t_i in passage p_j or key phrase k_l. The only difference between the A matrix in Eq. 3.4 and Eq. 3.7 is in the inclusion of key phrases as additional passages.

$$\underset{s\in(\cup_{i=1}^{n}S_i)-K}{\arg\max} \big|\{S_i : s \in S_i\}\big| \tag{3.6}$$

ALGORITHM 2: Key Phrases-Centrality algorithm (**KP-Centrality**).

Input: $D = Text\ Document,\ K = KeyPhrases(D),\ L = Number\ of\ Passages$;

Output: $S = Passages\ Retrieved\ (Summary)$;

// Create Compact Matrix Representation using TF-IDF
$M \longleftarrow$ Matrix(D);
// Add key phrases as new passages to the matrix M
$M \longleftarrow M \cup K$;

// Build support sets (clusters)
$C \longleftarrow$BuildSupportSets(M);

// Rank all passages using cosine similarity (default distance) excluding Key Phrases
$D_r \longleftarrow$ RankPassages(M, C, K);

// Select top L passages
$S \longleftarrow$SelectTopN(D_r, L);

$$A = \begin{bmatrix} w(t_1,p_1) & \dots & w(t_1,p_N) & w(t_1,k_1) & \dots & w(t_1,k_M) \\ \dots & & & & & \dots \\ w(t_T,p_1) & \dots & w(t_T,p_N) & w(t_T,k_1) & \dots & w(t_T,k_M) \end{bmatrix} \tag{3.7}$$

KP-Centrality intuitively improves over the OKP-Centrality and CKP-Centrality because it does not eliminate passages without key phrases. This is an advantage because in some cases those passages provide context to the main topic. There are other cases where the passages contain key phrases, which the AKE method failed to identify. In these situations, eliminating those passages causes important information to be lost. Conversely, if the method in out-of-domain data, it is possible to observe a reduction in the performance of the AKE method, leading to more errors in key phrase extraction. In that case, CKP-Centrality might be preferable over the KP-Centrality and OKP-Centrality because it discounts the importance of passages containing key phrases associated with low confidence scores. Another advantage of the KP-Centrality method over the other two methods is that it supports privacy [133, 134] while offering state-of-the-art performance.

The description of the two-stage extractive summarization methods is not complete without describing the key phrase extraction method used. We used the state-of-the-art approach introduced by Marujo et al. [137, 131] (Section 3.1). Since the pre-processing steps methodology described in [131] could have impact on the outcome of our experiments, we opted for not including them. This fact led to the exclusion of the Freebase sub-categories which were only beneficial in combination with pre-processing steps. Unfortunately, the news articles topic information, such as Sport, Politics, were not available. Therefore, the inclusion of rhetorical device features is the main difference between the Portuguese/Spanish and English AKE.

In this work, we also explored the fact that Portuguese and Spanish are closely related lan-

guages to use the AKE system created for Portuguese to Spanish [137]. The unique adaptation of the method was the inclusion of the list of Spanish stopwords.

3.2.4 Experiments

In order to assess the quality of our methods, we analyzed its performance using two different datasets. To evaluate the detection of the most important sentences, we used ROUGE [114], namely ROUGE-1, which is the most widely used evaluation measure for this scenario.

3.2.4.1 English (EN) and Spanish (SP) Event Reports (ER) Datasets

To evaluate our method, we used the Concisus Corpus of Event Summaries [183]. The corpus is composed by 78 event reports and respective summaries, distributed across three different types of events: aviation accidents, earthquakes, and train accidents. This corpus also contains comparable data in Spanish. However, since our AKE system uses some language-dependent features, we opted for not using this part of the dataset in our previous work [178]. However, to show that the summarization model is robust enough to be applicable to other datasets/languages, we use our AKE system created for Portuguese with the list of Spanish stopwords in our experiments on the Spanish part of the dataset.

3.2.4.2 Setup

In the experiments using both the English Event Reports dataset (EN ER dataset) and Spanish Event Reports Dataset (SP ER dataset), we generate 3 sentence summaries, commonly found in online news-aggregating web sites, for example Google News[1], Yahoo!News[2], Sapo.pt[3] and Applications, such as News360[4], Flipboard[5], and Pulse[6].

We compare the new three different approaches described in Section 3.2 to the baseline (the centrality-as-relevance raw model), with the number of key phrases ranging from five to forty. The metric used to configure the centrality model was the cosine (using IDF). The heuristic used to compute the size of each support set was the one based on the selection of the sentences with less distance to the sentence under analysis [177]. LexRank performance was also included for a better understanding of the improvements.

[1] http://news.google.com/
[2] http://news.yahoo.com/
[3] http://www.sapo.pt/
[4] http://news360.com/
[5] https://flipboard.com/
[6] https://www.pulse.me/

3.2.4.3 Results

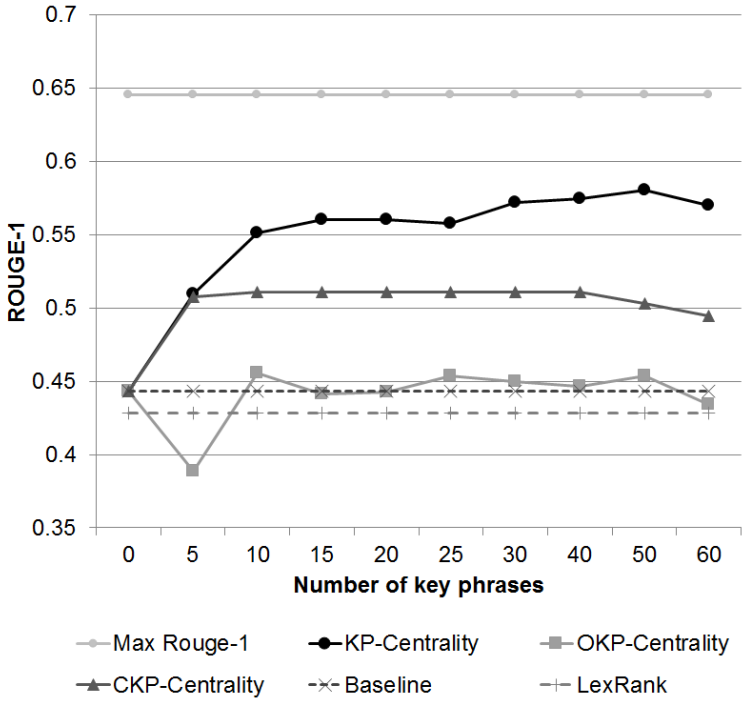


Figure 3.3: ROUGE-1 scores for the ENER.

Figure 3.3 shows the results for the EN ER dataset. As we can observe, the best performing approach is the KP-Centrality. It also reduces the distance to the optimal selection of important passages to 6.46 points. The optimal selection (Oracle) is identified as Max ROUGE-1 in Figure 3.3. This value is obtained by testing all summaries that can be generated and extracting the one with the highest score. In this dataset, KP-Centrality, and CKP-Centrality methods have statistically significant better performance than the baselines (p-value < 0.01 under t-test). However, in this dataset, the performance improves directly with the number of key phrases (until 40), while in the previous dataset the best results are achieved with about 25 key phrases. The performance of CKP-Centrality does not vary with the number of key phrases. This happens because the quality of the key phrases that are extracted became worse as we forced the extraction of more key phrases, and also because our key phrase models were optimized on datasets where each document contains approximately 30 to 40 key phrases for English and 20 to 30 for Portuguese/Spanish. Method OKP-Centrality achieves a similar performance in both datasets, although in the English dataset it does not outperform the baseline.

Figure 3.4 shows an example of important passages/summary retrieved using both KP-Centrality. All examples of summaries generated by our methods are shown in tokenized form. The methods are configured with 40 key phrases, which is the best configuration found

for the EN ER dataset.

Document avion-12110e

American Airlines Flight 587

American Airlines Flight 587, an Airbus A300, crashed into the Belle Harbor neighborhood of Queens, a borough of New York City, New York, shortly after takeoff from John F. Kennedy International Airport on November 12, 2001.
With 260 fatalities on board and 5 on the ground, this accident has the third highest death toll of any accident involving an Airbus A300.

Flight 587, circled in white, can briefly be seen in this video still moving downward with a white streak behind the aircraft. This video, released by the NTSB, was recorded by a tollbooth camera located on the Marine Parkway-Gil Hodges Memorial Bridge.

On November 12, 2001, about 09:16 eastern standard time, American Airlines flight 587, an Airbus A300-605R delivered in 1987 and powered by two General Electric CF6-80C2A5, crashed into Belle Harbor, a New York City residential area, shortly after takeoff from John F. Kennedy International Airport, New York. Flight 587 was a regularly scheduled passenger flight to Las Américas International Airport, Santo Domingo, Dominican Republic, with 2 flight crew members, seven flight attendants, and 251 passengers aboard the plane. Ed States served as the captain, and Sten Molin served as the first officer.

The plane's vertical stabilizer and rudder separated in flight and fell into Jamaica Bay, about 1 mile north of the main wreckage site. The plane's engines subsequently separated in flight and fell several blocks north and east of the main wreckage site. All 260 people aboard the plane and 5 people on the ground died, and the impact forces and a post-crash fire destroyed the plane. Flight 587 operated under the provisions of 14 Code of Federal Regulations Part 121 on an instrument flight rules flight plan. Visual meteorological conditions (VMC) prevailed at the time of the accident.

The A300-600 flew into the larger jet's wake, an area of turbulent air. The first officer attempted to keep the plane upright with aggressive rudder inputs. The strength of the air flowing against the moving rudder stressed the aircraft's vertical stabilizer and eventually snapped it off entirely, causing the aircraft to lose control and crash. The National Transportation Safety Board (NTSB) concluded that the enormous stress on the rudder was due to the first officer's "unnecessary and excessive" rudder inputs, and not the wake turbulence caused by the 747. The NTSB further stated "if the first officer had stopped making additional inputs, the aircraft would have stabilized". Contributing to these rudder pedal inputs were characteristics of the Airbus A300-600 sensitive rudder system design and elements of the American Airlines Advanced Aircraft Maneuvering Training Program.

Top 40 Key Phrases extracted ordered by rank

New York; died; Federal; mile; stopped; recorded; rules; accident; people; area; served; fell; ground; shortly; site; white; north; main; passenger; rules flight; take-off; separated; wreckage; wreckage site; main wreckage; two; keep; post; released; attempted; plan; seven; powered; involving; delivered; design; system; captain; camera; impact

---

Three-passages summary using KP-Centrality

American Airlines Flight 587 , an Airbus A300 , crashed into the Belle Harbor neighborhood of Queens , a borough of New York City , New York , shortly after takeoff from John F. Kennedy International Airport on November 12 , 2001 .

With 260 fatalities on board and 5 on the ground , this accident has the third highest death toll of any accident involving an Airbus A300 .

Flight 587 , circled in white , can briefly be seen in this video still moving downward with a white streak behind the aircraft .

Reference

2001 November 12 - American Airlines Flight 587, an Airbus A300, crashes into a Queens neighborhood in New York City when the plane's vertical tail fin snaps just after takeoff.

All 251 passengers and 9 crew members on board are killed as well as 5 people on the ground.

Figure 3.4: Example of important passage retrieval using KP-Centrality. Both methods use 40 key phrases and a document from the ENER.

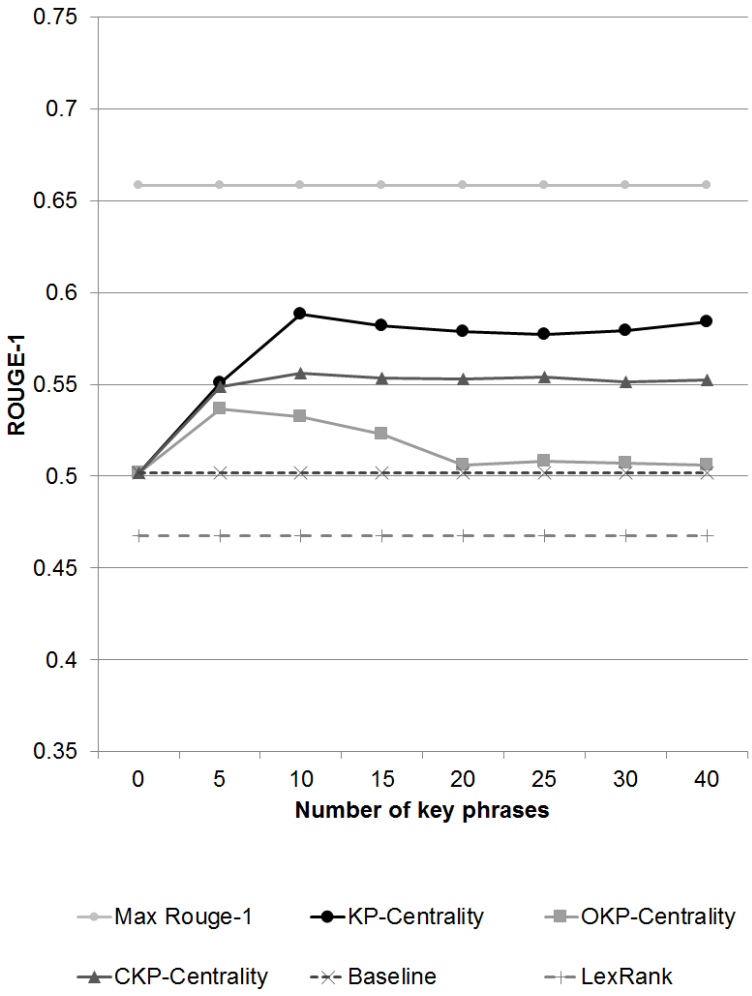


Figure 3.5: ROUGE-1 scores for the SPER.

Figure 3.5 shows the results for the SP ER dataset. The best summarization results are obtained with KP-Centrality method (58.8%). This method obtained this result using ten key phrases and degrades slightly as we increase this number. This result allow us to

draw the conclusion that the list of Spanish key phrases extracted is still flawed and we suspect that better results could be obtained by training a AKE for Spanish. Nevertheless, the KP-Centrality results are statistically significant better than the baseline (p-value $<$ 0.01 under t-test)

Method	ENER	SPER
KP-Centrality	28.403	37.217
Centrality	30.766	39.439

Table 3.2: Average number of words per sentence/SU in the summaries generated by the main approaches for all datasets.

Table 3.2 presents an interesting aspect concerning the content of the generated summaries by two and three-stage approaches: the number of words per sentence decreases (as the number of stages increases), showing that these more complex methods select more informative sentences while reducing the actual length (in terms of words) of the summaries.

3.2.5 Conclusions

In this chapter, we introduced new two-stage methods for single document summarization. Popular centrality-based models treat all elements of the retrieval space equally, impacting the retrieval task negatively. In line of recent work [102, 177], we show that our methods can improve the performance of a retrieval model that already addresses this issue. The methods we propose start by extracting a collection of key phrases that will be used to bias a centrality-as-relevance retrieval model. We explore three different methods for the integration of key phrases.

Using ROUGE-1 as evaluation metric, one method (KP-Centrality) improves the baseline model, with statistically significance, for English documents by 31% (with 50 key phrases). The improvements of the new approaches lowers for the Spanish text to 17% (with 10 key phrases), but the difference is also statistically significant. The best trade-off configuration in terms of number of key phrases across both datasets is 40 key phrases. Under this configuration, the KP-Centrality method obtained 0.5745 ROUGE-1 value (30% relative improvement over the baseline model) for the English documents and 0.5839 ROUGE-1 value (16% relative improvement) for the Spanish documents. The observed performance gain oscillations are directly correlated with the accuracy of the automatic key phrase extraction (AKE) method (both identification and ranking of the key phrases). The performance of the English AKE is shown on Section 3.1 and for Spanish/Portuguese is available in [137]. The best performing variant of the two-stage approaches, KP-Centrality, was also better,

with statistical significance, than the single-stage baselines. Furthermore, the approach where passages that do not contain key phrases are removed does not achieve as good results, which means different aspects are captured in the two stages of our methods. Key phrases and the centrality model seem to complement each other.

The summarization results benefited from the improvements in AKE. We investigated the use of additional semantic features and pre-processing steps to improve automatic key phrase extraction. These features include the use of signal words and Freebase categories. Some of these features lead to significant improvements in the accuracy of the results. We also experimented with two forms of document pre-processing that we call light filtering and co-reference normalization. Light filtering removes sentences from the document, which are judged peripheral to its main content. Co-reference normalization unifies several written forms of the same named entity into a unique form. We also needed a gold standard – a set of labeled documents for training and evaluation. While the subjective nature of key phrase selection precludes a true gold standard, we used AMT service to obtain a useful approximation for English.

Our data indicates that the biggest improvements in performance in automatic key phrase extraction were due to shallow semantic features, news categories, and rhetorical signals (NDCG 77.02% vs. 70.45%). The inclusion of deeper semantic features such as Freebase sub-categories was not beneficial in itself, but in combination with pre-processing, did cause slight improvements in the NDCG scores.

In the next chapter, we will explore the combination of the KP-Centrality method with event information.

4 Event-based Single-Document Summarization

> *Every significant event that takes place in our lives is set to some kind of music.*

Peabo Bryson, *American R&B and soul singer-songwriter (1951)*

A thorough analysis of the bibliography of the summarization research area clearly shows that the main focus of automatic single-document summarization are news stories. Organizations need to have quick access to the information that affects them, people want to be informed about the environment where they act. The bulk of this information is disseminated either by written text, such as newspaper articles, or speech as broadcast news. Interestingly, although this type of documents is characterized by conveying information about events, most of the work concentrates on approaches that do not take into account this aspect.

For the reasons above, in this chapter, we focus on how to improve the key phrase-guided centrality-as-relevance summarization model (KP-Centrality) using event information. Within this framework, we explore different ways of incorporating event information, attaining state-of-the-art results in both written and spoken language documents. The resulting summaries consist of a sequence of extracts (sentences, paragraphs, or, in some cases, sentence-like units when summarizing automatic transcriptions of spoken documents) that are selected according to a relevance rank influenced by event information.

We introduced a new event detection method based on Fuzzy Fingerprints [86, 181, 138] which is able to detect all types of events in the ACE 2005 Multilingual Corpus [208]. The Fuzzy Fingerprint method outperformed, in the event-detection task, traditional machine learning supervised classifiers, such as Support Vector Machines and Random Forests.

This chapter is structured as follows: the next section describes the event detection methods and experiments; our event-based summarization model is presented in Section 4.2, Section 4.3 describes the experiments with our event-based summarization model, and Section 4.4 draws the conclusions.

4.1 Event Detection

Automatic event detection is an important Information Extraction task that can be used to help finding specific content of interest to a user. By event detection we refer to the ability to properly classify text excerpts or passages according to specific categories, such as "Meet", "Transport", "Attack". For example, the following news excerpt "The construction of the high speed train line from Madrid to Lisbon, scheduled to start operation in 2017 has been canceled" should be automatically detected as a "Transport" event. Event detection has been largely explored in the context of Question Answering [188], Topic Detection and Tracking (TDT) [46], and Summarization [64]. Here we specifically address the problem of single event detection when in the presence of a large number of classes.

In our experiments, we use the ACE 2005 Multilingual Corpus [208]. Even through this corpus was manually annotated for 27 different single label event types, usually only a few event types are used due to the (arguably) insufficient number of instances necessary to train more traditional classifiers. For example, in [155], only 6 events types are used, due to the difficulties in obtaining results with more classes when using Support Vector Machines (SVMs) [167] or Random Forests [43], i.e., less than 20% of the possible event types are used.

In this work we propose and use Fuzzy Fingerprints [86] as a mechanism to improve automatic event detection for a large number of classes. When applying Fuzzy Fingerprints to the ACE 2005 Multilingual Corpus, it was possible to detect up to 100% of the event types, obtaining a much higher G–mean [98, 196], an assessment measure especially adequate to imbalanced multiclass classification problems, when comparing to the best prior event detection method, SVMs [155]. In order to obtain the results, we started by replicating and confirming the work described by Naughton, and then improved their results by adding several new features to the used machine learning algorithms. We achieved a 4.6% points improvement in F-scores over the results reported by [155]. Then, we created a Fuzzy Fingerprints Event library and adapted the similarity score proposed in [181] to retrieve events, in order to improve the results when using all the event types in the ACE 2005 corpus.

4.1.1 ACE 2005 Corpus

The ACE2005 Corpus was created for the ACE evaluations, where an event is defined as a specific occurrence involving participants described in a single sentence. The corpus has a total of 12,298 sentences (or 11,691 excluding sentences belonging to more than one event, see Fig. 4.1 do get more details about the distribution). Each sentence is identified with event types or off event/null (N) when the sentence does not belong to any of the event types. There are 33 event types: *Be-Born, Marry, Divorce, Injure, Die, Transport, Transfer-*

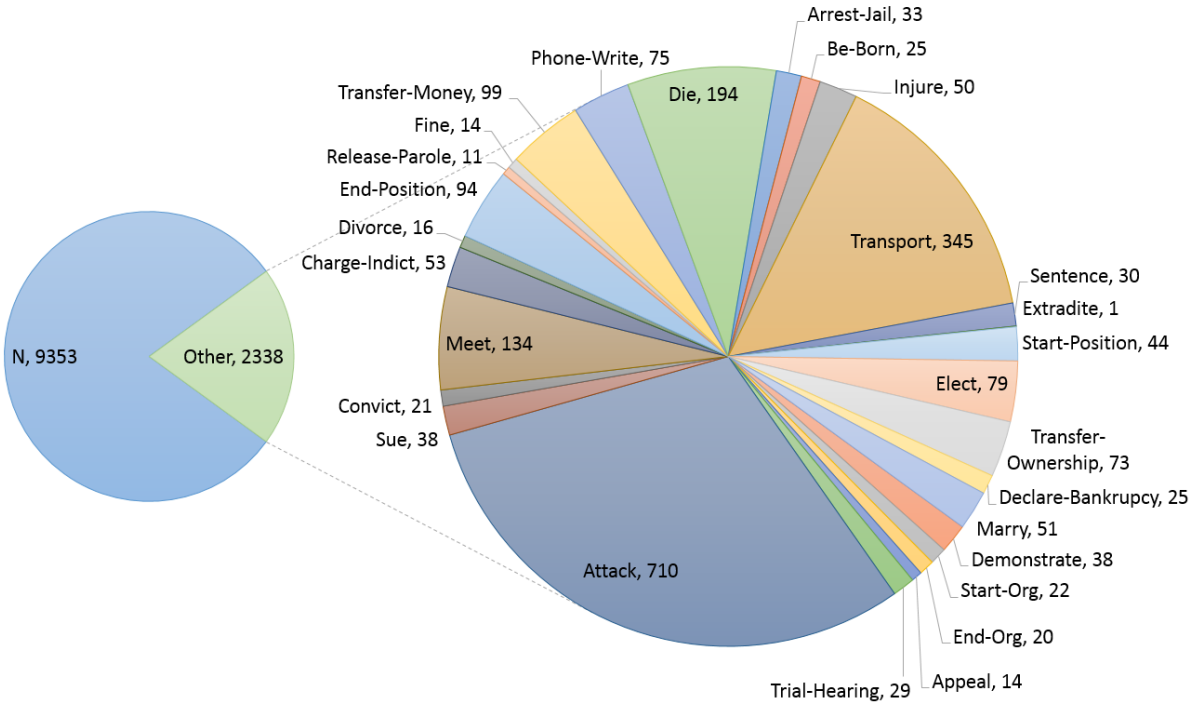


Figure 4.1: Distribution of the number of sentences per event type in the ACE 2005 Multilingual Corpus excluding multi-event sentences.

Ownership, Transfer-Money, Start-Org, Merge-Org, Declare-Bankruptcy, End-Org, Attack, Demonstrate, Meet, Phone-Write, Start-Position, End-Position, Nominate, Elect, Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon. From these 33 events, only the following 6 have a high number of instances or sentences in the corpus: *Die, Attack, Transport, Meet, Injure, and Charge-Indict.* These are the only ones used in the previous referred works. About 24% of the sentences contain at least 1 event, and 21% of those sentences are classified as multi-event (or multi-label); for example, the sentence "three people died and two were injured when their vehicle was attacked" involves 4 event types (or one event with 4 event type labels). This means that multi-event sentences correspond to only 5% of the corpus, and were removed since we are only addressing single-label classification. Also, six event types only occur in multi-event sentences, which means they were excluded from the dataset. Finally, event *Extradite* only occurs once in the corpus, and was removed, since it is not possible to separate 1 instance into test and training sets. As a result, the dataset used in this work contains 26 different event types.

```
<!DOCTYPE source_file SYSTEM "apf.v5.1.1.dtd">
<source_file URI="CNN_ENG_20030630_085848.18.sgm"
SOURCE="broadcast news"
TYPE="text"
AUTHOR="LDC"
ENCODING="UTF-8">
<document DOCID="CNN_ENG_20030630_085848.18">

[...]

<event ID="CNN_ENG_20030630_085848.18-EV1"
TYPE="Life"
SUBTYPE="Injure"
MODALITY="Asserted"
POLARITY="Negative"
GENERICITY="Specific"
TENSE="Unspecified">
<event_mention ID="CNN_ENG_20030630_085848.18-EV1-1">
<extent><charseq START="337" END="344">injuries</charseq></extent>
<ldc_scope><charseq START="334" END="388">no injuries have been
reported thankfully hat this time</charseq></ldc_scope>
<anchor><charseq START="337" END="344">injuries</charseq></anchor>
</event_mention>
</event>
</document>
</source_file>
```

Figure 4.2: ACE 2005 Multilingual Corpus event example.

4.1.2 Machine Learning Event detection

A state-of-the-art way to solve a text multi-class problem, like single-event detection, is to use Support Vector Machines (SVM) techniques [225]. Random Forests (RF) [43] are also seen as an alternative to SVM because they are considered one of the most accurate classifiers [57]. RF has also advantages on datasets where the number of features is larger than the number of observations [57]. In our dataset, the number of features extracted is between two and three times greater than the total number of instances of events.

4.1.2.1 Features

The spoken transcripts documents found in the ACE 2005 corpus contain raw Automatic Speech Recognition (ASR) single-case words with punctuation. This means that the transcriptions were either manually produced or were generated by a standard ASR with minimal

manual post-processing. Absence of capitalization is known to negatively influence the performance of parsing, sentence boundary identification, and NLP tasks in general. Recovery of capitalization entails determining the proper capitalization of words from the context. This task was performed using a discriminative approach described [34]. We capitalized every first letter of a word after a full stop, exclamation, or question mark. After true-casing, that is assigning the correct capitalization to words, we automatically populate three lists for each article: list K of key phrases, list V of verbs, and list E of named entities. Key phrase extraction is performed using our supervised AKE method (Chapter 3). The method extracted 40 key phrases per document. Verbs are identified using the Stanford POS tagger [201], and named-entities using the Stanford NER [65]. This extraction is performed over all English documents of the corpus. The K, V, and E lists are used in the extraction of lexical features and dependency parsing-based features. The lists K and V were also augmented using WordNet [152] synsets to include less frequent synonyms. Furthermore, we manually created list M of modal verbs, and list N of negation terms.

The feature space for the classification of sentences consists of all entries in the lists V, E, K, M, and N, which are corpus-specific. The value of each feature is the number of its occurrence in the sentence. These numbers indicate the description of events by numbering the number of participants, actions, locations, and temporal information. We have also explored other uncommon types of features: Rhetorical Signals (Chapter 3) and Sentiment Scores [198]. Finally, we removed all features with constant values across classes. This process reduced by half the number of features and improved the classification results.

4.1.3 Event Detection based on Fuzzy Fingerprint classification

Another alternative to the machine learning methods to perform *Even detection* is using the Fuzzy Fingerprints classification method [86, 181, 138]. Homem and Carvalho [86] approached the problem of authorship identification by using the crime scene fingerprint analogy to claim that a given text has its authors writing style embedded in it. The algorithm works as follows:

1. Gather the top-k most frequent words (and their frequencies) in all known texts of each known author;

2. Build the fingerprint by applying a fuzzifying function to the top-k list. The fuzzified fingerprint is based on the word order and not on the frequency value;

3. Perform the same calculations for the text being identified and then compare the obtained text fuzzy fingerprint with all available author fuzzy fingerprints. The most similar fingerprint is chosen and the text is assigned to the author of the fingerprint.

The method, when used for event detection [138], is similar in intention and form, but differs in a few crucial steps. First, it is important to establish the parallel between authorship identification and event detection. Instead of author fingerprints, in this context, we are looking for fingerprints of events and to classify each passage (sentence in text or Sentence Unit (SU) in speech) according to the encompassed event. The process starts with the creation of an event fingerprint library. Then, each unclassified passage can be processed and compared to the fingerprints existing in the event library. Second, a different criterion was used in ordering the top-k words for the fingerprint. While Homem and Carvalho [86] use word frequency as the main feature to create and order the top-k list, here we use an adaptation of the Inverse Document Frequency (Inverse Document Frequency (IDF)) technique, aiming at reducing the influence of frequent terms that are common across several events. We adapted the original IDF because it reduces the influence of frequent terms that are common across several documents ignoring event information.

4.1.3.1 Building the Event Fingerprint Library

The first step of the event fingerprint library creation stage is computing word frequencies for each event type. We use as training corpus the ACE 2005 [208] corpus. Only the top-k most frequent words are considered. The main difference between the original method and the one used here is due to the small size of each sentence: in order to make the different event fingerprints as unique as possible, its words should also be as unique as possible. Therefore, in addition to counting each word occurrence, we also account for its Inverse Topic Frequency (Inverse Topic Frequency (ITF)), an adaptation of IDF (TF-ITF): $\text{itf}_v = \frac{N}{n_v}$, where N is the cardinality of the event fingerprint library (i.e., the total number of events), and n_v becomes the number of fingerprint events where the specific word v is present. After obtaining the top-k list for a given event, we follow the original method and apply a fuzzy membership function to build the fingerprint. The selected membership function (Eq. 4.1) is a Pareto-based linear function, where 20% of the top-k elements assume 80% of the membership degree.

$$\mu(i) = \begin{cases} 1 - (1-b)\frac{i}{k}, & \text{if} \quad i \leq ak \\ & \text{with } a, b = 0.2 \\ a\Big(1 - \frac{i-a}{k-a}\Big), & \text{if} \quad i > ak \end{cases} \tag{4.1}$$

The fingerprint is a k-sized bi-dimensional array, where the first column contains the list of the top-k words ordered by their TF-ITF scores, and the second column contains the membership value of word i, $\mu(i)$, obtained by the application of Eq. 4.1. Table 4.1 shows two examples of event fingerprints ordered by $\mu(i)$ values for the event types *Start-organization* and *Meet*.

The table does not include the complete fingerprints to increase the readability. In the table, we show the top 10 entries, the bottom 3 and some intermediate entries. Each entry contains the rank based on $\mu(i)$, calculated setting $k = 600$, word i (where i value is the TF-ITF rank), and $\mu(i)$ the membership value.

Table 4.1: Event fingerprints of the *Start-Organization* (Left) and *Meet* (Right) event types, ordered by $\mu(i)$.

Rank	Word	$\mu(i)$	Rank	Word	$\mu(i)$
1	founded	34.3897	1	meet	33.8333
2	committee	32.7875	2	summit	31.9200
3	collectors	32.4060	3	ended	31.5000
4	sheik	32.0245	4	discussed	30.9400
5	films	31.4142	5	meetings	30.8933
6	budget	31.3379	6	eu	30.2400
7	reformist	30.8038	7	meeting	28.8200
8	hamshahri	30.6512	8	discuss	29.4933
9	cinema	30.4223	9	saint	29.4467
10	forging	30.1935	10	talks	17.1967
24	launched	10.7511	28	meets	5.2634
34	opening	5.5725	38	talk	5.0322
67	business	4.4114	119	contacting	2.6025
100	empire	3.1298	199	talked	1.8089
101	contract	3.1107	240	resolution	1.5405
178	acquired	1.4312	360	reunited	0.8986
195	launching	1.2786	394	organization	0.7352
265	make	0.5324	414	met	0.6652
365	then	0.0310	598	off	0.0117
366	year	0.0057	599	eased	0.0117
367	been	7.3399E-4	600	knows	0.0078

4.1.3.2 Classifying Sentences

The method for authorship identification has 3 steps: build the document fingerprint (using the previously described algorithm); compare the document fingerprint with every fingerprint present in the library; and, choose the match with highest score. However, for event detection, where a document is a sentence, such approach would not be feasible due to the small number of words comprised in one sentence. By using the S2E (Sentence to Event) function that tests the fitness of a sentence to a given event fingerprint. The S2E function (Eq. 4.2) provides a normalized value ranging between 0 and 1, that takes into the number of features in the preprocessed sentence, which in our case is the total number of words since every word is used as a feature.

account the size of the (preprocessed) sentence (i.e., its number of features). In the present work, we use as features the words of the sentences/SUs. We do not remove stop-words (empirical results show that the best results are obtained without removing stop-words or by imposing a minimum word size).

$$S2E(\Phi, S) = \frac{\sum_{v \in \Phi \cap S} \mu_\Phi(v)}{\sum_{i=0}^{j} \mu_\Phi(w_i)} \tag{4.2}$$

In Eq. 4.2, Φ is the event fingerprint, S is the set of words of the sentence, $\mu_\Phi(v)$ is the membership degree of word v in the event fingerprint, and j is the number of features of the sentence. Essentially, S2E divides the sum of the membership values $\mu_\Phi(v)$ of every word v that is common between the sentence, and the event fingerprint, by the sum of the top-j membership values in $\mu_\Phi(w_i)$ where $w_i \in \Phi$. Eq. 4.2 will tend to 1 when most or all words of the sentence belong to the top words of the fingerprint, and tends to 0 when none or very few words of the sentence belong to the bottom words of the fingerprint.

4.1.3.3 Evaluation and Results

Our evaluation compares SVM, Random Forest, and Fuzzy Fingerprints to detect events. The SVM performed better than the Random Forest to detect events in low to medium number of classes. For these reasons, we chose SVM to investigate the inclusion of the additional features over the baseline set proposed by Naughton [155]. We have also investigated the influence of the new features introduced in this work by using all features except for the ones under test. These novel features raised the G-Mean scores by 16.2% relative percentage (Table 4.2) when detecting six events. The average F-value was also improved to 0.5097.

In term of relative percentage, the inclusion of the domain-Id features raised the G-Mean score by 12.01%, which is the highest contribution among the new features. The second best

result, using dependency parse based features, is 6.72%. The relevance-based features, such as the sentiment analysis and rhetorical features had the lowest contribution with respectively 3.62% and 1.42%. As expected, the introduction of new features reduced the recall of the majority class (no-event or off event) between -0.75% and -0.32%, but improved the recall of the remaining labels. The exception to this fact is the detection of "Die" events that was also penalized. This can be explained in part by the imbalanced distribution of the event types, which biased the classifier towards more frequent event types. In this case, the classifier is biased towards "Attack" events, which are three times more frequent than "Die" and share similar new features values. When increasing the number of event types to cover all the 26 events present in the ACE 2005 database, the SVM performed very poorly, failing to detect 11 of the 26 events. This implied a G-Mean = 0, and the F-score decreased to 0.381. The SG-Mean was also a rather poor (0.160).

Table 4.2: Feature Extraction analysis in terms of recall (R_i) in ACE 2005 using SVM with improved features.

Labels	All Features	All - Rhetorical Signals	All - Sentiment Analysis	All - Dependency Parsing	All - Domain Id	Baseline Features
Injure	**0.280**	0.260	0.240	0.180	0.220	0.200
Transport	**0.328**	0.325	0.316	0.319	0.316	0.270
Attack	0.428	**0.431**	0.410	0.420	0.407	0.356
N	0.938	0.937	0.940	0.941	0.937	**0.944**
Meet	0.373	0.366	0.366	**0.381**	0.328	0.336
Charge-Indict	0.415	0.415	0.415	0.415	0.321	0.321
Die	0.469	0.464	0.469	**0.474**	0.433	**0.474**
G-Mean	**0.429**	0.423	0.414	0.402	0.383	0.369

Several tests were done in order to find the best Fuzzy Fingerprint parameters. The best empirical results led to the inclusion of all words in the fingerprints (i.e., include stop words and small sized words). The fingerprint size K was optimized for the best SG-mean. Figures 4.3 through 4.5 show the obtained SG-Mean and number of undetected event classes for 6 and 26 events for several values of K. With 6 events, the best G-Mean=0.808, SG-Mean=0.809, and WSG-Mean=0.021, were obtained for K=2500, and represent an improvement of around 88% when compared to the best SVM result. However, F-score=0.323 was lower. The best F-score=0.421 was obtained for K=10000, but the SG-Mean and WSG-Mean dropped to 0.498 and 0.013.

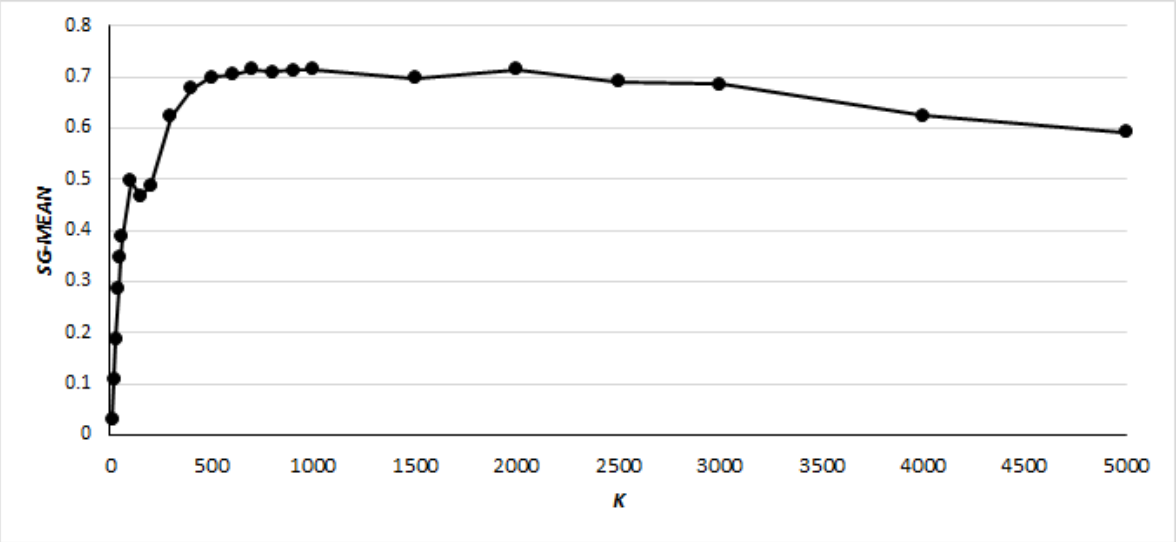


Figure 4.3: SG-Means results in the ACE 2005 with 6 events using Fuzzy Fingerprints with K values.

When tested for the 26 events database, the Fuzzy Fingerprints method is able to detect all the classes for the values of K between 400 and 1000, while the SVM only detects 22 out of 26 classes. The best SG-Mean is 0.714, was obtained with K=700, and represents a 4.5x improvement over the best SVM result. However, F-score=0.192 is 50 percentage points lower than SVM. For K=20000, the F-score=0.448 outperformed the F-score for SVM. In addition, the SG-Mean=0.326 is 2x improvement over the best SVM score (0.160).

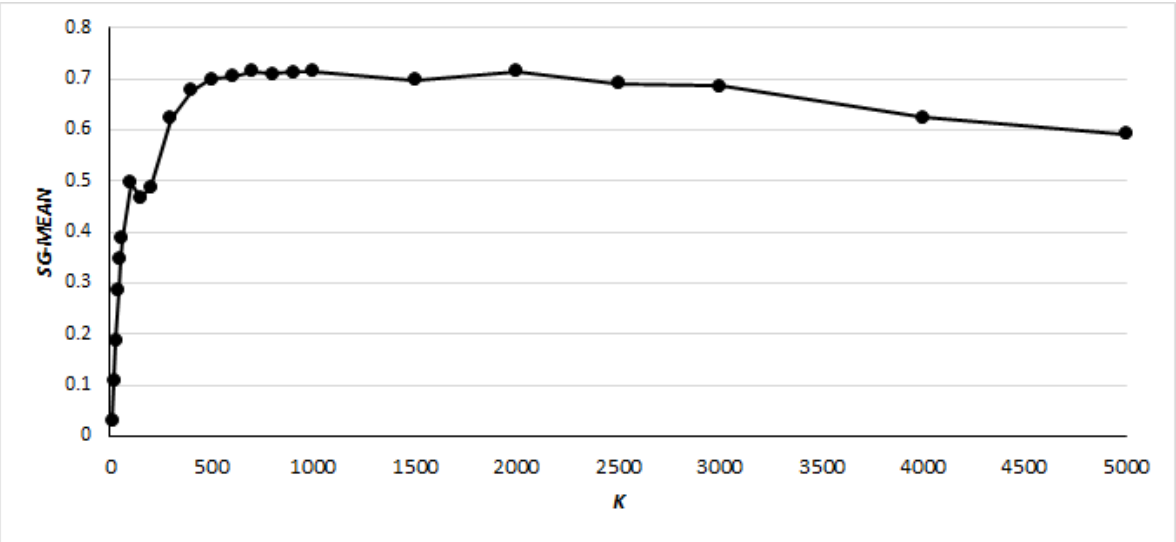


Figure 4.4: SG-Means results in the ACE 2005 with 26 events using Fuzzy Fingerprints with K values.

Finally, Tables 4.3 and 4.4 show the comparative results (including RF) for the 6 and 26 events

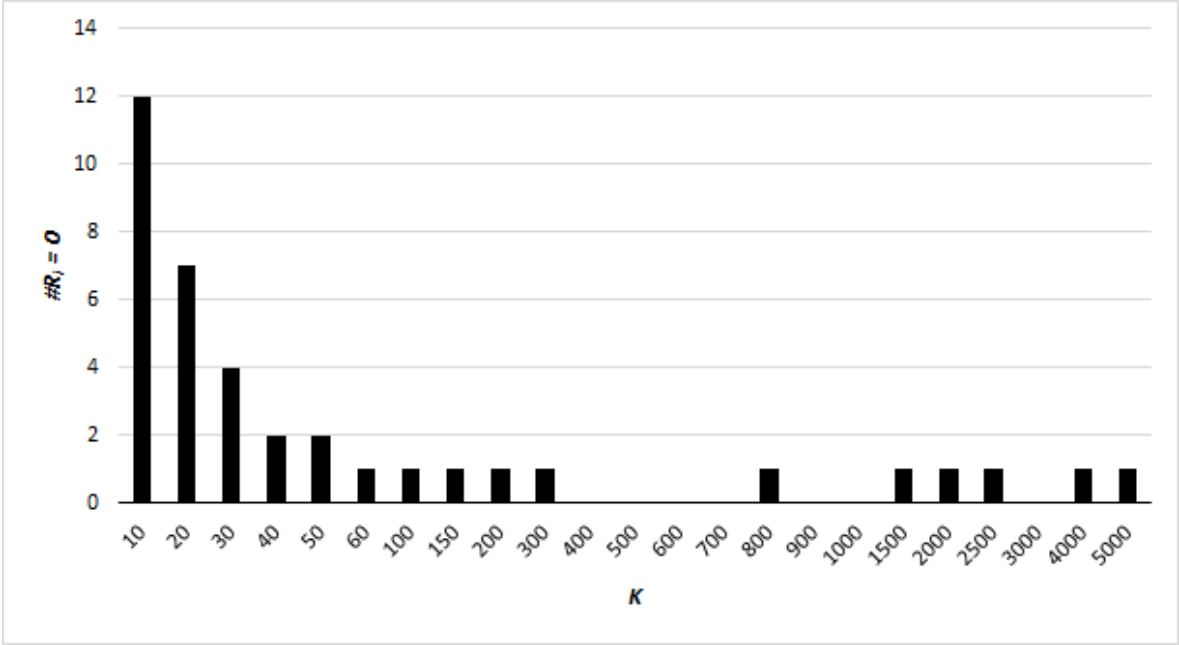


Figure 4.5: $\#R_i = 0$ (number of event classes missed) results in the ACE 2005 with 26 events using Fuzzy Fingerprints with several K values.

ACE2005 database. The best results are shown in bold. Since we used the Fuzzy Fingerprint method to detect and summarize events in another corpus (Section 4.2), we believe that the Fuzzy Fingerprint method is not significantly over-fitting the training data.

Table 4.3: Results in the ACE 2005 corpus with 6 events.

Measure	Random Forest	SVM (SMO)	Fuzzy Fingerprints (best)
F1 (avg.)	0.142	**0.510**	0.421
G-Mean	0	0.429	**0.808**
SG-Mean	0.006	0.430	**0.809**
WSG-Mean	1.607×10^{-4}	0.011	**0.021**
$\#R_i = 0$	4	**0**	**0**

4.1.3.4 Discussion

In this Section, we approached the problem of detecting events at sentence level, a single-label classification procedure whose results can be used to improve several NLP tasks such as personalization, recommendation, question answering, and/or summarization. We are

Table 4.4: Results in the ACE 2005 corpus with 26 events.

Measure	Random Forest	SVM (SMO)	Fuzzy Fingerprints (best)
F1 (avg.)	0.037	0.381	**0.448**
G-Mean	0	0	**0.644**
SG-Mean	0.002	0.160	**0.714**
WSG-Mean	4.365×10^{-6}	4.068×10^{-4}	**0.003**
$\#R_i = 0$	22	4	**0**

specifically interested in the cases where a large number of event classes must be detected, since more traditional classifiers, such as SVM or RF, usually loose a large number of classes. We started by improving the best previously known approaches, and then proposed the use of Fuzzy Fingerprints. The ACE 2005 corpus, which contains 26 different single event classes was used throughout the experiments. The results show that it is possible to detect all 26 different event types when using the Fuzzy fingerprints approach, while the best competitor, an SVM with enhanced features, only detects roughly 85% of the different types of events. This leads to a large increase in the G-Mean results when using the Fuzzy Fingerprints method.

The Fuzzy Fingerprints method also has the advantage of being much more efficient in computational terms. In our test conditions, it is more than 20x faster than SVM when classifying the 26 event types.

The application of the Fuzzy Fingerprints is still in an early development phase. Future work includes using advanced features such as key phrases to build the fingerprints, and also the fuzzification of key phrases. It is also in our plans to apply the method to the detection of multiple events in single sentences.

4.2 Event-based single-document summarization

Naturally, one way to identify important events is to select sentences describing events, filter the others, and then rank the sentences using a centrality-as-relevance method, such as KP-Centrality. This alternative has the disadvantage of not providing any event information about the sentences to the ranking algorithm, which might not be able to detect that two sentences are about the same event because they are written using different words. The simplest, but efficient, way to avoid this limitation is to use the event descriptors obtained using a event classifier as additional features of the ranking algorithm. Since these two alternatives are not mutually exclusive, we also explored their combination.

4.2.1 Event-Enhanced KP-Centrality (EE-KPC)

As previously mentioned, the KP-Centrality method consists of two steps: first, it extracts key phrases using a supervised approach, and, then, it combines them with a bag-of-words model, represented by a terms by passages matrix, to compute the most important content.

The EE-KPC method includes event information in the KP-Centrality-based summarization process at the Important Passage Retrieval module level. This is accomplished by expanding the bag-of-words matrix representation of passages with event descriptors—vectors of $S2E$ values describing each event type obtained using the event fingerprint method for each sentence/SU and key phrase. Fig. 4.6 shows the complete architecture.

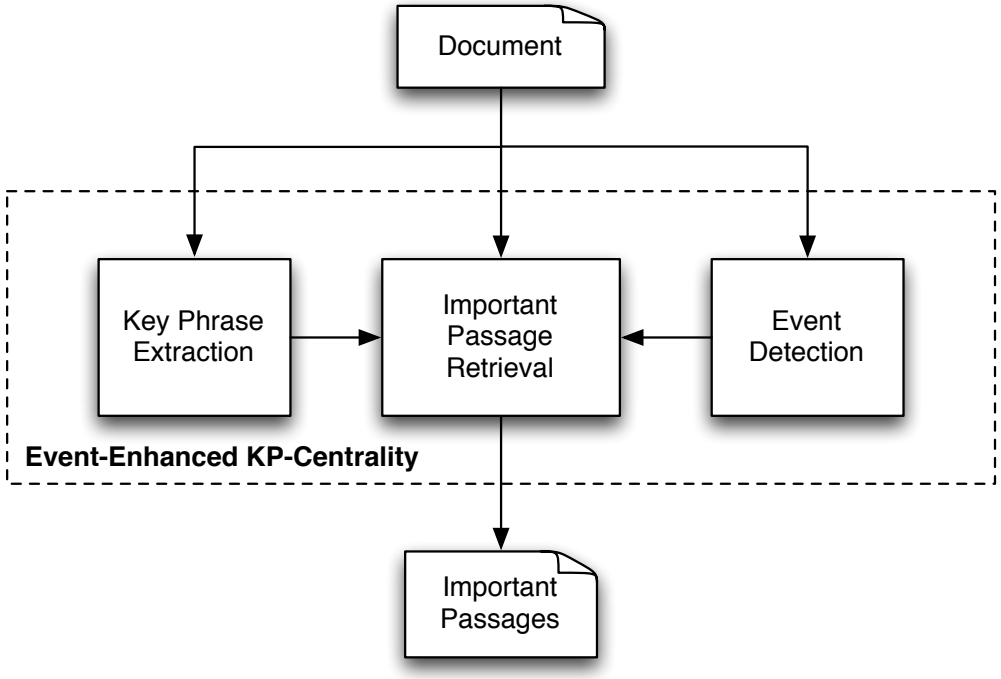


Figure 4.6: EE-KPC architecture.

Eq. 4.3 defines the new matrix representation, where w is a function of the number of occurrences of term t_i in extract e_j or key phrase k_l, T is the number of terms, M is the number of key phrases, c is a function of the $S2E$ score of each extract e_j or key phrase k_l for each event type ev_m.

$$\begin{bmatrix} w(t_1,e_1) & \dots & w(t_1,e_N) & w(t_1,k_1) & \dots & w(t_1,k_M) \\ \vdots & & & & & \vdots \\ w(t_T,e_1) & \dots & w(t_T,e_N) & w(t_T,k_1) & \dots & w(t_T,k_M) \\ \mathbf{c(ev_1,e_1)} & \dots & \mathbf{c(ev_1,e_N)} & \mathbf{c(ev_1,k_1)} & \dots & \mathbf{c(ev_1,k_M)} \\ \vdots & & & & & \vdots \\ \mathbf{c(ev_E,e_1)} & \dots & \mathbf{c(ev_E,e_N)} & \mathbf{c(ev_E,k_1)} & \dots & \mathbf{c(ev_E,k_M)} \end{bmatrix} \tag{4.3}$$

Each column represents an extract p_i. The extracts are ranked to produce a summary according to Eq. 2.19 and 2.20.

4.2.2 Event Filtering-based KP-Centrality (EF-KPC)

The EF-KPC method includes the same stages of the previous method, EE-KPC, but it uses event information in a different manner. Instead of expanding the bag-of-words matrix representation, it discards sentences that do not contain events—the Event Filtering stage shown in Fig. 4.7 includes an event detection step. All passages that the event fingerprint method classifies as not containing any event are removed. The exception to this simple rule occurs when the method is not confident about the classification result ($\max S2E < 0.0001$). Then, the KP-Centrality is used to produce a summary. Note that although pattern matching-based methods could adopt this strategy, our event detection method is more robust and allows us to correctly identify a larger number of event types. In fact, we are able to detect events in passages with different syntactic structures, something which is more difficult to pattern matching-based methods.

For example, the passage "Four marines died in the crash of an osprey in North Carolina." that appear in the Concisus dataset is not detected by pattern matching-based methods because it does not have a named entity followed by a verb or action noun, and another named entity. Within the passage, there is only one named entity, the location North Carolina. Our event detector is able to detect that the passage describes a Die event type ($S2E$ score = 0.1553).

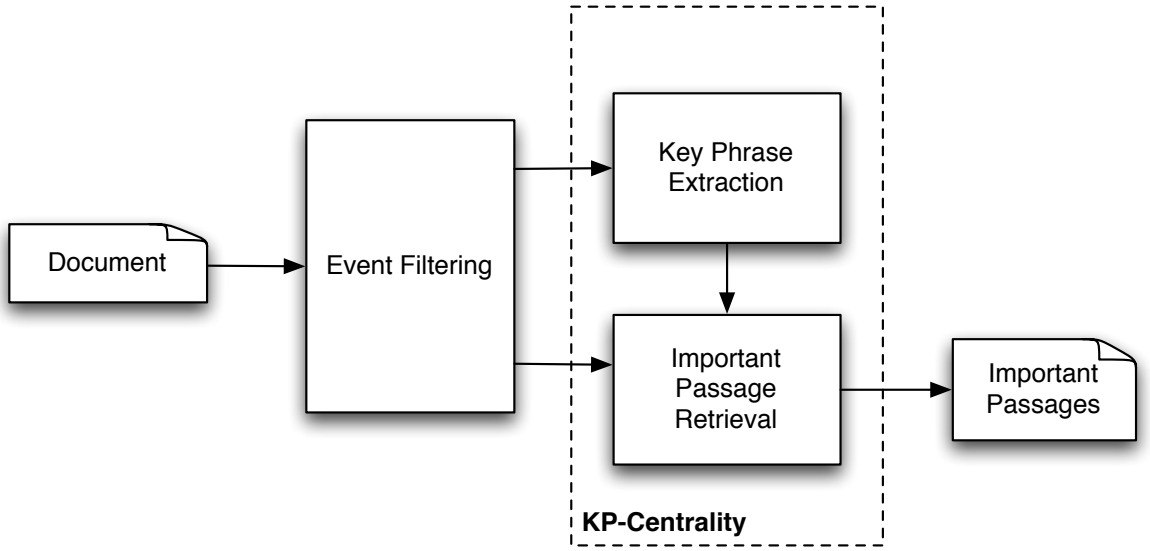


Figure 4.7: EF-KPC architecture.

4.2.3 Combination of Event Filtering-based and Event-Enhanced KP-Centrality (CE-KPC)

The CE-KPC method combines the two previous methods as shown in Fig. 4.8. It starts by filtering the passages without events, as in the EF-KPC method, and includes the $S2E$-based event descriptors in the bag-of-words matrix representation, as in the EE-KPC method.

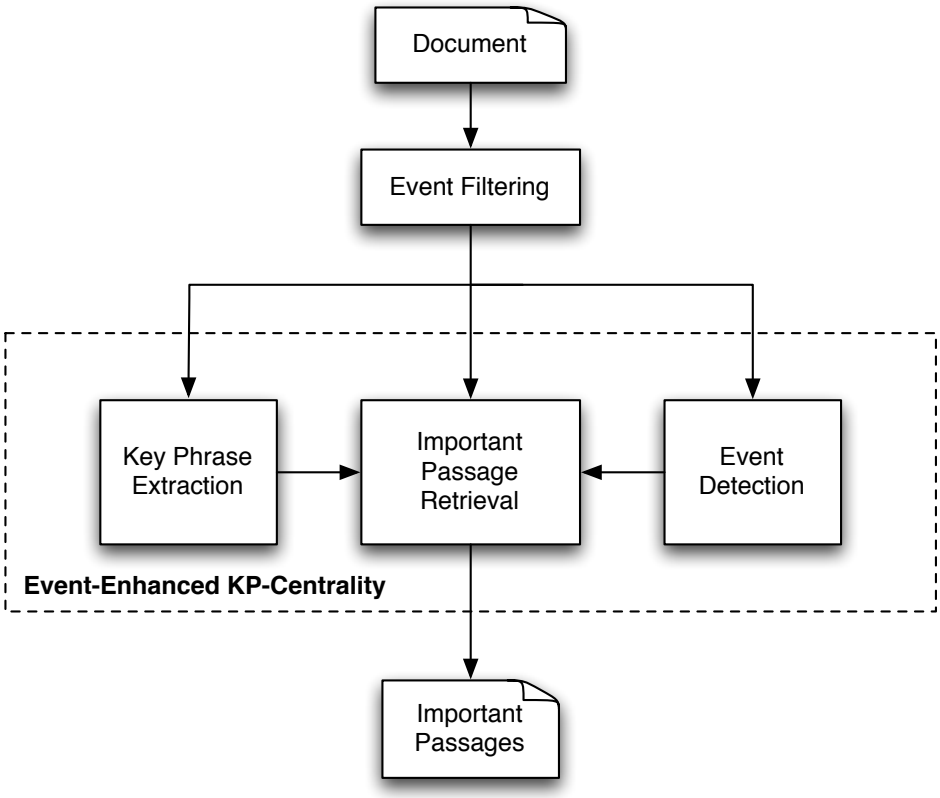


Figure 4.8: CE-KPC architecture.

4.3 Experiments

To assess the influence of using event information in our extractive summarization method, we use two datasets: the Concisus Corpus of Event Summaries [183]; and a subcorpus of the Columbia BN Speech Summarization Corpus [141]. The first dataset is composed by event reports (written text) and can be seen as an ideal dataset for this kind of approach. The second dataset is composed by broadcast news stories. The use of these two datasets provides different experimental conditions, helping to understand the real impact of the method.

To evaluate the detection of the most important sentences/SUs, we used ROUGE [114], namely ROUGE-1, which is the most widely used evaluation measure for this scenario. In the following experiments, we generate 3 sentence summaries, commonly found in online news web sites, like Google News.

4.3.1 Datasets

Concisus Corpus of Event Summaries. The corpus is composed by 78 event reports and respective summaries, distributed across three different types of events: aviation accidents, earthquakes, and train accidents. Table 4.5 has statistics about the size of corpus, namely the number of documents, average number of sentences, and average number of words.

Table 4.5: Statistics of the Concisus Corpus of Event Summaries.

	#Docs	Avg. #Sentences	Avg. #Words
Input Documents	78	4,286	224,922
Reference Summaries	78	2,154	65,820

Columbia BN Speech Summarization Corpus test set. The corpus consists of a random sample of 16 broadcast news stories from the test subcorpus of the Columbia BN Speech Summarization Corpus II. The CBNSCII is composed by 20 CNN Headlines News shows from the TDT-4 corpus. For each news story, there is a human summary that is used as reference. Table 4.6 provides some statistics about the corpus.

Table 4.6: Statistics of the Columbia BN Speech Summarization Corpus test set.

	#Docs	Avg. #Sentences	Avg. #Words
Input Documents	16	11,313	203,313
Reference Summaries	16	3,563	70,688

4.3.2 Results

Table 4.7 shows the ROUGE-1 results for the Concisus dataset and Table 4.9, for the Columbia BN dataset. As previously mentioned, it is possible to use different metrics to compute semantic similarity in the centrality-as-relevance summarization model. In these experiments, we explored the best performing metrics (for clean and noisy data), as presented by Ribeiro and de Matos [177]: cosine similarity and distance frac133 (generic Minkowski distance, Eq. 4.4, with $N = 1.(3)$). Since the result for the Concisus dataset do not show improvement over the baseline using metric frac133, we opted for not presenting them.

$$dist_{minkowski}(\mathbf{x}, \mathbf{y}) = \Big(\sum_{i=1}^{n} |x_i - y_i|^N \Big)^{\frac{1}{N}} \tag{4.4}$$

Table 4.7: ROUGE-1 scores for the Concisus dataset. $\diamond$ indicates statistical significance difference under macro t-test after rank transformation (p-value < 0.09) [223].

#Key Phrases	30	40	50	60
LexRank		0.428		
Centrality		0.443		
KP-Centrality	0.572	0.575$^{\diamond}$	0.581	0.570
EE-KPC	0.574	**0.586**$^{\diamond}$	0.585	0.581
EF-KPC	0.574	0.584	0.583	0.581
CE-KPC	0.574	0.584	0.583	0.579

Table 4.8: Percentage of number of sentences that are different between KP-Centrality and the event-based generated summaries in the Concisus corpus using 40 key phrases.

	KP-Centrality	EE-KPC	EF-KPC	CE-KPC
KP-Centrality	0.00	22.41	18.97	19.40
EE-KPC		0.00	16.38	11.64
EF-KPC			0.00	6.47
CE-KPC				0.00

For all the performed experiments, the use of event information clearly improves baselines: for the Concisus dataset, we observe differences between EE-KPC and KP-Centrality, using 40 key phrases, with statistical significance (p-value < 0.9) under the macro t-test after rank transformation [223]; in the broadcast news dataset, both EF-KPC (frac133, 30 and 50 key phrases) and CE-KPC (frac133, 50 key phrases) are significantly better than KP-Centrality using the same statistical test (p-value < 0.04).

The best result in the Concisus dataset is 0.586, achieved by EE-KPC, while in the Columbia Broadcast News dataset the best result was 0.752 (frac133 distance), achieved by EF-KPC and CE-KPC.

Another interesting aspect is that, even though not all variations achieve statistical significance, the resulting summaries are still different. As we can see in Table 4.8, for the Concisus

Table 4.9: ROUGE-1 scores for the Columbia Broadcast News dataset. Compared pairs of systems are marked with the same symbol (‡, †, ⋄); differences are statistically significant under the macro t-test after rank transformation (p-value < 0.04) [223].

Similarity metric: cosine				
# Keys Phrases	30	40	50	60
Centrality		0.564		
KP-Centrality	0.673	0.697	0.656	0.684
EE-KPC	0.673	0.697	0.656	0.684
EF-KPC	0.678	**0.710**	0.668	0.692
CE-KPC	0.678	**0.710**	0.668	0.692

Similarity metric: frac133				
# Keys Phrases	30	40	50	60
Centrality		0.700		
KP-Centrality	$0.695^{\ddagger}$	0.702	$0.738^{\dagger,\diamond}$	0.698
EE-KPC	0.714	0.693	0.743	0.700
EF-KPC	$0.729^{\ddagger}$	0.699	$\mathbf{0.752}^{\dagger}$	0.702
CE-KPC	0.712	0.691	$\mathbf{0.752}^{\diamond}$	0.702

LexRank	0.653

Table 4.10: Differences (in percentage) in terms of number of sentences between KP-Centrality and the event-based generated summaries (50 key phrases) in the Broadcast News corpus.

	KP-Centrality	EE-KPC	EF-KPC	CE-KPC
KP-Centrality	0.00	6.38	10.64	23.40
EE-KPC		0.00	19.15	19.15
EF-KPC			0.00	12.77
CE-KPC				0.00

dataset differences between the variants and the KP-Centrality baseline range from 18.97% to 22.41%. Table 4.10 shows the same information for the broadcast news dataset, with differences ranging from 6.38% to 23.40%. In fact, having different summarizations approaches generating different summaries with similar performances is in line with the possibility of having different good summaries for the same document.

As we can see, the best results on the Concisus dataset are obtained integrating event information as a feature in the summarization model (EE-KPC); followed by the use of event information to filter out unimportant content (EF-KPC) and the combination of both strategies (CE-KPC). Contrarily to what was observed in the Concisus dataset, the EF-KPC and CE-KPC methods outperform the EE-KPC method in the broadcast news corpus. The justification for this difference of performance is threefold. The first reason is the noisy nature of the speech data, where removing passages without events helps discarding the unimportant content. The second reason is that the percentage of sentences/SUs filtered out is higher in the broadcast news corpus than in the Concisus corpus, as shown in Table 4.11. The third reason is the nature of the corpus: since the Concisus corpus is composed of event reports, the filtering approach does not have the same impact as in the broadcast news corpus, and the use of event information as a feature helps distinguishing the most important facts or events within each report. In addition, the length of the documents to be summarized can influence the performance of the EE-KPC, since the number of event features is constant and the number of term features increases according to the Zipf's law. Concerning the better performance of the semantic relatedness metric frac133 over cosine similarity, which in general is the best performing metric, it might be related to the influence of the $S2E$ values. These values vary inversely proportional to the length of the sentence/SU. The high average sentence length of the Concisus corpus makes $S2E$ lower and the use frac133 makes it more difficult to distinguish close passages. On the other hand, on the broadcast news corpus, the average SU length is lower and, inversely, $S2E$ values are higher, which makes frac133 more effective.

Table 4.11 shows the specific effects of using event information: as we can see, the number of sentences/SUs classified as containing, at least, one event type is high ($\approx 90\%$). However, the number of sentences/SUs kept after filtering is even higher ($\approx 95\%$). The reason for this is to cope with the classifier errors.

Figure 4.9 shows an example of important passages retrieved using KP-Centrality, EE-KPC, EF-KPC, CE-KPC methods. The methods are configured with 40 key phrases, which is the best configuration found for the Concisus corpus. We also included the event label and respective $S2E$ values for each sentence/SU of the original document. The event detector identified two sentences that do not cover any of the event types (N - no event or null event). However, only one of the sentences obtained a "high"

Table 4.11: Statistics about event classification and filtering on the Concisus and the CBNSCII corpora.

	Concisus	CBNSCII
#Sentences/SUs	815	181
Avg. #Sentences/SUs	10	11
#Sentences/SUs after Filtering	786 (96%)	172 (95%)
Avg. #Sentences/SUs after Filtering	10	10
#Event-Classified Sentences/SUs	734 (90%)	159 (88%)
Avg. #Event-Classified Sentences/SUs	9	9

$S2E$ score ($S2E > 0.0001$). This kind of sentences usually describe secondary topics and details. The EF-KPC filter explores this information to filter irrelevant sentences to improve the quality of the summaries.

Document terremoto-31011906

The 1906 Ecuador-Colombia earthquake occurred at 15:36 UTC on January 31, off the coast of Ecuador, near Esmeraldas. The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia.

The earthquake occurred along the boundary between the Nazca Plate and the South American Plate. The earthquake is likely to be a result of thrust-faulting, caused by the subduction of the Nazca plate beneath the South American plate.

The coastal parts of Ecuador and Colombia have a history of great megathrust earthquakes originating from this plate boundary.

The greatest damage from the tsunami occurred on the coast between Río Verde, Ecuador and Micay, Colombia. Estimates of the number of deaths caused by the tsunami vary between 500 and 1,500.

Event classification of the sentences in the document

(Event=Charge-Indict, S2E=0.011) - The 1906 Ecuador-Colombia earthquake occurred at 15:36 UTC on January 31 , off the coast of Ecuador , near Esmeraldas .

(Event=Injure, S2E=0.020) - The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia .

(Event=N, S2E=1.4E-45) - The earthquake occurred along the boundary between the Nazca Plate and the South American Plate .

(Event=Injure, S2E=0.021) - The earthquake is likely to be a result of thrust-faulting , caused by the subduction of the Nazca plate beneath the South American plate .

(Event=N, S2E=0.016) - The coastal parts of Ecuador and Colombia have a history of great megathrust earthquakes originating from this plate boundary .

(Event=Charge-Indict, S2E=0.012) - The greatest damage from the tsunami occurred on the coast between Río Verde , Ecuador and Micay , Colombia .

(Event=Die, S2E=0.041) - Estimates of the number of deaths caused by the tsunami vary between 500 and 1,500 .

Three-passages summary using KP-Centrality

The earthquake is likely to be a result of thrust-faulting , caused by the subduction of the Nazca plate beneath the South American plate .

The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia .

The coastal parts of Ecuador and Colombia have a history of great megathrust earthquakes originating from this plate boundary .

Three-passages summary using EE-KPC

The earthquake is likely to be a result of thrust-faulting , caused by the subduction of the Nazca plate beneath the South American plate .

The greatest damage from the tsunami occurred on the coast between Río Verde , Ecuador and Micay , Colombia .

The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia .

Three-passages summary using EF-KPC

The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia .

The earthquake is likely to be a result of thrust-faulting , caused by the subduction of the Nazca plate beneath the South American plate .

The 1906 Ecuador-Colombia earthquake occurred at 15:36 UTC on January 31 , off the coast of Ecuador , near Esmeraldas .

Three-passages summary using CE-KPC

The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia .

The earthquake is likely to be a result of thrust-faulting , caused by the subduction of the Nazca plate beneath the South American plate .

The greatest damage from the tsunami occurred on the coast between Río Verde , Ecuador and Micay , Colombia .

Reference

January 31, 1906

The 1906 Ecuador-Colombia earthquake occurred at 15:36 UTC on January 31, off the coast of Ecuador, near Esmeraldas.

The earthquake had a magnitude of 8.8 and triggered a destructive tsunami that caused at least 500 casualties on the coast of Colombia.

Figure 4.9: Example of important passage retrieval using KP-Centrality, EE-KPC, EF-KPC, and CE-KPC. All methods use 40 key phrases and a document from the Concisus Corpus of Event Summaries.

According to ROUGE, the summary produced by EF-KPC, shown in Figure 4.9 should be the best, but for the human reader, it looks the most incoherent. One aspect that was not given attention in this work was the order in which the sentence appear in the summary. One simple solution is to order the sentences in the order they appear in the original document. Yet, existing metrics do not take into account the order that the sentence occurs, making it difficult to evaluate the sentence ranking. As it is not trivial to define and test a metric for this problem, this will be left as future work.

Notice, that the reference summary, in Figure 4.9, was composed of the first two sentences or the original article, which is consistent with the standard model of news articles, namely that the first paragraph gives the summary.

4.4 Conclusions

In this chapter, we introduced three event-based summarization techniques that perform above the state-of-the-art methods. Our event detection method is based on the Fuzzy Fingerprints classification method and trained on the ACE 2005 Multilingual Corpus. The obtained event information is integrated in a two-stage summarization method in three ways: one approach consists in expanding the feature representation of sentences/SUs with event information (EE-KPC); another technique is to filter out sentences/SUs without events (EF-KPC); and, finally, we also explore the combination of both techniques (CE-KPC). The approach that yielded the best results in the written text dataset (the Concisus corpus of event reports) was EE-KPC. The use of event information to filter out unimportant passages was the best performing approach in the speech dataset, the Columbia Broadcast News corpus. Still, EE-KPC also achieved better results than the baselines. In general, CE-KPC had a similar or worse performance than the EF-KPC because this method accumulates errors from both stages. Since the filtering stage discards all sentences/SUs, the available segments to be selected is similar to EF-KPC. Inherently, the next stage cannot overcome the errors made in the filtering step and, possibly, introduces additional errors. Given the experimental results, we believe that there is a relation between the performance of the EE-KPC and the length of the input. This might be mitigated by increasing the weight of the event features. Another aspect that influences the results is the performance of the classifier. In our experiments, we gave preference to recall, which maximized the number of sentences/SUs containing events.

In the next chapter, we will explore the adaptation of our single-document event-based summarization to multi-document summarization which is a much more complex task than single-document summarization. There are several challenges, such as redundancy between documents, and cohesion.

5 Event-based Multi-Document Summarization

> *History is the depository of great actions, the witness of what is past, the example and instructor of the present, and monitor to the future.*
>
> Miguel de Cervantes, *Spanish Novelist, poet, playwright, and soldier (1547-1616)*

As seen in Chapter 2, the use of the Internet to fulfill generic information needs motivated pioneer multi-document summarization efforts such as, NewsInEssence [174] or Newsblaster [144], online since 2001. Many other automatic multi-document summarization systems have been proposed in order to cope with the growing number of news stories published online. The main goal of these systems is to convey the important ideas in these stories, by eliminating less crucial and redundant pieces of information. In particular, most of the work in summarization has been focused on the news domain, which is strongly tied to events, as each news article generally describes an event or a series of events. However, only few attempts have focused on the use of event information for summarization systems for the news domain [71]. In fact, most of the work on multi-document summarization are either based on Centrality-based [173, 61, 212, 177], Maximal Marginal Relevance (MMR) [45, 78, 187, 112], and Coverage-base methods [116, 195, 64, 108, 122, 230, 29, 71]. Generally, centrality-based models are used to generate generic summaries, the MMR family generates query-oriented ones, and coverage-based models produce summaries driven by topics or events.

The use of event information in multi-document summarization can be arranged in the following categories: early **hand-based experiments** [54]; **pattern-based approaches** based on enriched representations of sentences, such as the cases of the work presented by Zhang et al. [232] and by Wenjie Li et al. [108], which define events using an event key term and a set of related entities, or centrality-based approaches working over an event-driven representation of the input [71], where events are also pattern-based; and, **clustering-based** event definition [107].

The major problem of these approaches is the difficulty to relate different descriptions of the same event due to different lexical realizations. In our work, we address this problem by

using an event classification-based approach and including event information supported by two different distributed representations of text—the skip-ngram and continuous bag-of-words models [151]. As seen in Chapter 4, our event detection and classification framework is based on vector-valued fuzzy sets [86, 138] introduced in the Section 4.1.

Previous chapters discussed single-document summarization methods. In this chapter, we are extending KP-Centrality to perform multi-document summarization. This extension requires us to overcome more challenges beyond greater compression factor and more overall redundancy, than single-document summarization needs. Multi-document summaries need to begin with an introductory sentence that provide some context about their main topic. They also need to be coherent, that means that good multi-document summaries can not contain contradictory information, such as causalities death update numbers. Another aspect that affects summaries coherence is anaphora resolution. The inherent complexity of anaphora resolution in multi-document summarization is higher than in single-document summarization because anaphoras (e.g., pronouns) need to be resolved at two levels: document level and set of documents level. Another major problem that multi-document summarization needs to address is focus, the summary should not contain extraneous information. Event-based information helps to solve or at least reduce some of these additional problems. In Filatova et al. [64], one of the initial summarization methods using event information, by only keeping sentences with at least two named-entities and a verb, the method filters out most anaphoric sentences according to our observations. Another advantage of using event information in multi-document summarization is in the fact that event-based summaries are a sequence of events, which increases the likelihood of the summary being coherent. Our adaptation of single-document to multi-document summarization explores two hierarchical strategies to perform this extension. One of them, namely the Waterfall method described in the next section, provides a better initial context sentence to summary than our default single-document summarization method KP-Centrality or the other multi-document extension (single-layer hierarchical). The underlying rationale for the advantage of the Waterfall KP-Centrality over the other methods is that it explores the fact that more recent news tend to containing a good introductory sentence that summarizes previous events. This fact is accomplished by increasing the relevance to of the content of the most recent documents. The use of event information also have a positive impact in the selection of a good introductory sentence and summarization in general as we will show in Section 5.2.

We evaluated both the generic and event-based multi-document summarization methods using the standard summarization evaluation metric, ROUGE [114]. Moreover, to better understand the impact of using event information, we also performed a human evaluation using the AMT.

The rest of this chapter is organized as follows: Section 5.1 presents the adaptation of KP-

CENTRALITY to generic multi-document summarization. Section 5.2 extends the generic multi-document summarization to multi-document summarization based on events. The conclusions close the document.

5.1 Generic Multi-Document Summarization

Our goal is to extend the KP-CENTRALITY method for multi-document summarization. The simplest method would be to concatenate all documents and use the single-document method to produce the summary. We shall use this approach as a baseline. This baseline works quite well for a small number of documents, but the performance decreases as the number of documents increases. This means that KP-CENTRALITY has limitations identifying redundant content, such as events, when it is written with different words. Another limitation of the baseline method is to ignore temporal information as more recent news documents tend to contain more relevant information and sometimes include brief references to the past events to provide some context.

To overcome the first limitation, we consider two simple but effective alternative approaches for improving the baseline method. The first approach is a two-step method where we summarize each document individually in such a way that each of the summaries has the size of the final multi-document summary. This is followed by the concatenation of all the resulting summaries, which is then summarized again into the final summary. In both steps, we use the KP-CENTRALITY method to generate the summaries. The advantage of this approach is to reduce the redundancy of information at document level (intra-document). This means that we also need to reduce the redundancy of information between document (inter-documents). The second method also reduces the redundancy inter-documents, but in a different way. Rather than considering all summaries simultaneously, it takes one summary s_1, concatenate with another summary s_2, summarize the result to obtain a summary of documents s_1 and s_2, which we denote as $s_{1...2}$. Next, it takes $s_{1...2}$ and performs the same operation with s_3, obtaining $s_{1...3}$. This is done recursively for all the N documents in the from the input, and the final summary is the one obtained in $s_{1...N}$.

We will denote these methods as hierarchical single-layer and waterfall. These are illustrated in Figures 5.1 and 5.2, respectively.

We also modified the first stage of the KP-CENTRALITY method, where a set of key phrases are predicted and used in subsequent stages. Since we are working with multiple documents, we rank the key phrases according to the number of documents containing them, and only keep the set of top k ranking key phrases, where k is a hyper-parameter of the model. Our main motivation is the fact that important key phrases for the topic must occur consistently

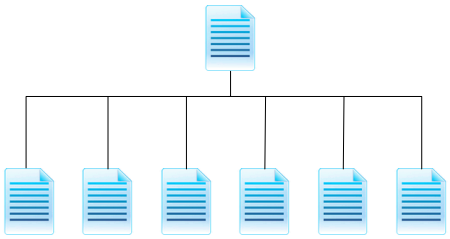


Figure 5.1: Single-layer architecture.

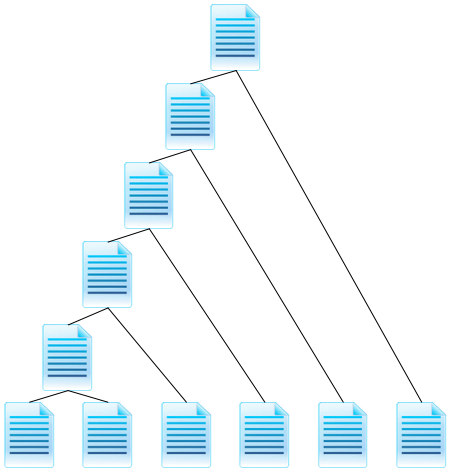


Figure 5.2: Waterfall architecture.

across different documents about that topic. For instance, a set of documents about "Barack Obama" moving to the "White House", should include the key phrases "Barack Obama" and "White House" in most of them.

5.1.1 Experiments

We compare the performance of our methods against other representative models, namely MEAD, MMR, Expected n-call@k [112], and the Portfolio Theory [213]. The MEAD is a centroid-based method and one of the most popular centrality-based methods. The MMR is one of the most used query-based methods. The Expected n-call@k adapts and extends the MMR as a probabilistic model (Probabilistic Latent MMR). The Portfolio Theory also extends the MMR based on the idea of ranking under uncertainty. As baseline, we used the straightforward idea of combining all input documents into a single one, and then submit the document to the single-document summarization method. Considering that coverage-based systems explore event information, we opted for not including them in this comparative analysis.

To assess the informativeness of the summaries generated by our methods, we used ROUGE-

1 [114] on DUC 2007 and TAC 2009 datasets.

5.1.1.1 DUC 2007

The main summarization task in DUC 2007[1] is the generation of 250-word summaries of 45 clusters of 25 newswire documents and 4 human reference summaries. Each document set has 25 news documents obtained from the AQUAINT corpus [73].

5.1.1.2 TAC 2009

The TAC 2009 Summarization task[2] has 44 topic clusters. Each topic has 2 sets of 10 news documents obtained from the AQUAINT 2 corpus [207]. There are 4 human 100-word reference summaries for each set, where the reference summaries for the first set are query-oriented multi-document summaries, and for the second set are update summaries. In this work, we used the first set of reference summaries.

5.1.1.3 Results

The used features include the bag-of-words model representation of the sentences (TF-IDF), the key phrases and the query (obtained from the topics descriptions). Including the query is a new extension to the KP-Centrality method, which, in general, improved the results. We experimented with different numbers of key phrases, obtaining the best results with 40 key phrases. To compare and rank the sentences, we used several distance metrics, namely: Frac133 (generic Minkowski distance, with $N = 1.(3)$), Euclidean, Chebyshev, Manhattan, Minkowski, the Jensen-Shannon Divergence, and the cosine similarity. Table 5.1 shows that the best results were obtained by the proposed hierarchical models, in both datasets. The best performing distance metrics for the centrality-based method were the frac133 for the single-layer method (DUC 2007, although the difference for cosine is hardly noticeable), and the cosine for the waterfall method (TAC 2009). Single-layer with frac133 shows a performance improvement of 0.0180 ROUGE-1 points (relative performance improvement of 5.0%) over the best of the other systems, Portfolio, in DUC 2007 and of 0.0845 ROUGE-1 points (19.7% relative performance improvement) in TAC 2009. Overall, the best trade-off configuration is waterfall using cosine, since it also achieves a performance improvement over Portfolio of 0.0106 ROUGE-1 points (relative performance improvement of 3%). Note that our baseline obtained results similar to the best reference system in DUC 2007 and better results than all reference systems in TAC 2009 (0.0454 ROUGE-1 points; 10.6% relative performance

[1] http://www-nlpir.nist.gov/projects/duc/duc2007/tasks.html
[2] http://www.nist.gov/tac/2009/Summarization/

Distance	Model	DUC 2007	TAC 2009
frac133	baseline	0.3565	0.4706
cosine		0.3406	0.4746
frac133	single-layer	**0.3775**	0.4983
cosine	waterfall	0.3701	**0.5137**
MEAD		0.3282	0.4153
MMR		0.3269	0.3917
E.n-call@k		0.3209	0.3873
Portfolio		0.3595	0.4292

Table 5.1: ROUGE-1 scores.

improvement). The better results obtained on the TAC 2009 dataset are due to the small size of the reference summaries and to the fact that the documents sets to be summarized contain topics with higher diversity of subtopics.

5.1.2 Discussion

In this Section, we presented a multi-document summarization framework that extends a single-document summarization method, KP-Centrality, in two hierarchical ways: single-layer and waterfall. Overall, the best trade-off configuration for the summarizer is waterfall using cosine similarity.

5.2 Event-based Multi-Document Summarization

The waterfall method [136], introduced in the previous section, is sensitive to the order of the input documents. Since at each iteration the summaries of the documents are merged with the summary of the previous documents, the content of the initial documents is more likely to be removed than the content in the last documents. Thus, it is important to consider the order of the documents. We chose to organize the documents chronologically where the older documents are summarized and merged in the first iteration of the waterfall method. The waterfall method has two drawbacks. One limitation is the size of the intermediate summaries. Once we decided the size of the final summary, we obtain the intermediate summaries with the size of the final summary. In practice, this work well, but in some cases the size of the intermediate summary is not enough to contain all necessary information for the summarization process. From this limitation also emerges the second, which is the identification of redundant content between documents when written with different words.

Our solution to the first limitation of the waterfall method is as we merge more documents recursively, the intermediate summaries that contains the information of the documents so far, will grow in size to avoid losing important information. For that reason, we increased the number of sentences in the intermediate summary as a function of the number of documents that have been covered. More formally, the size of the summary at a given time or document t is defined as:

$$L = \delta \times K \times log(t + \phi) \tag{5.1}$$

where K is the maximum number of words in the final summary, ϕ is a constant to avoid zeros ($\phi = 2$). δ is a scale factor that is 1 for the generation of the initial documents summaries and 200 for the remaining cases.

Since the more recent documents contain more updated content, we also increased the size of initial documents summaries created by the hierarchical single-layer based on Eq. 5.1 to not give an unfair advantage to the waterfall method.

The identification of redundant sentences written in different ways is not an easy task. For instance, the sentence "The Starbucks coffee co. plan to acquire Pasqua coffee is leaving a bitter aftertaste in the mouths of some patrons of the San Francisco-based coffeehouse." and "Starbucks , the nation 's largest coffee retailer , announced Tuesday that it would buy Pasqua for an undisclosed amount." have essentially the same meaning: a company plans to buy another. Nevertheless, the only common content between the two sentences are the company names. For this purpose, we propose two alternatives that complement each other. On the one hand, news documents describe events (e.g., Company acquisitions), thus sentences that cover the same event are good candidates to contain redundant information. On the other hand, different lexical realizations with the same meaning can be addressed using distributed word representations.

From this point, we present the two extensions to our multi-document summarization framework. Figure 5.3 gives an overview of the architecture of our event-based multi-document methods (the block Hierarchical Important Passage Retrieval corresponds to the Waterfall and Single-layer methods).

5.2.1 Supervised Event Classification

Our event detection method is based on the Fuzzy Fingerprints classification method [138], which was introduced in Section 4.1. This work approaches the problem of authorship identification by using the crime scene fingerprint analogy that leverages the fact that different authors have different writing styles. The algorithm is computed as follows: (1) Gather the top-k word frequencies in all known texts/sentences of each known author/event; (2) Build

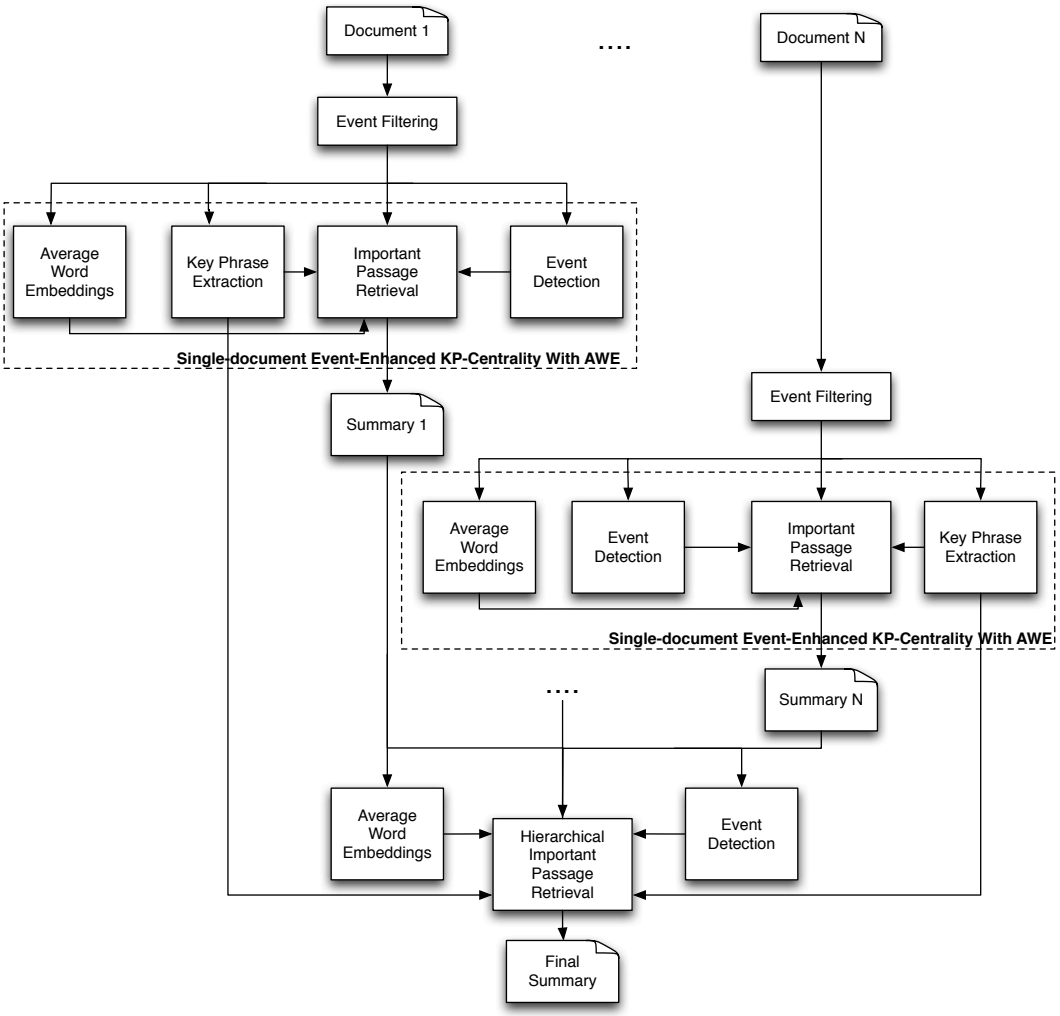


Figure 5.3: Architecture of the Event-based Multi-document summarization methods.

the fingerprint by applying a fuzzifying function to the top-k list. The fuzzified fingerprint is based on the word order and not on the frequency value; (3) For each document, perform the same computations to obtain a fingerprint and assign the author/event with the most similar fingerprint.

Our motivation for the use of event information is the existence of secondary events that are not relevant to the main event of the documents, which need to be excluded from the summary. We already observed in the previous Chapter the importance of event information to single-document summarization, but it becomes even more important for multi-document summarization where the number of secondary events is much higher.

To obtain event information, we use the event fingerprint method to identify sentences that describe events. Since we needed training data to build the event fingerprint of each event type, we used again the ACE 2005 Multilingual Corpus [208]. These event fingerprints are used to generate each sentence fingerprint. For example, the fingerprint of the sentence "ETA, whose name stands for Basque Homeland Freedom, has killed nearly 800 people since 1968 in its campaign for Basque independence" considering, for example, only four event types would be the following vector: [Die = 0.1061, Attack = 0.0078, Divorce = 0.0, Null or No-event = 0.01907]. All sentences that the event fingerprint method classified as not containing any event are removed (F.E. - filtering events). The exception to this simple rule occurs when the method is not confident in the classification result (confidence less than 0.0001, obtained when we compute the fingerprint of the sentence). This event filtering is an optional pre-processing step of the multi-document summarization.

After filtering out the sentences that do not describe events, we also need to identify similar events. This is accomplished by using the sentence event fingerprints as features in the summarization process. This means that each sentence has 27 new features, each corresponding to one of the 27 different event types: Appeal, Arrest-Jail, Attack, Be-Born, Charge-Indict, Convict, Declare-Bankrupcy, Demonstrate, Die, Divorce, Elect, End-Org, End-Position, Fine, Injure, Marry, Meet, N (Null/No Event), Phone-Write, Release-Parole, Sentence, Start-Org, Start-Position, Sue, Transfer-Money, Transfer-Ownership, Transport, Trial-Hearing.

Our approach to the extraction of event information does not fall into any of the previously known categories (exploratory hand-based experiments; pattern-based approaches; and, clustering-based), since it is a supervised classification method.

5.2.2 Unsupervised Word Vectors

Although the event detection method described above is supervised, where features are extracted from annotated data, we also need to leverage the large amount of raw text (without

annotation) in an unsupervised setup. The small size of the annotated data is insufficient to cover also possible ways of describing events. Large amounts of raw text without event annotations are easy to obtain and contain different descriptions about the same event. Thus, we need a method to relate the event descriptions. For this purpose, we use the method recently introduced in Mikolov et al. [151], which uses raw text to build a representation for each word, consisting of a d-dimensional vector. Two models were proposed in this work, the skip-ngram model and the continuous bag-of-words model, which we shall denote as SKIP and CBOW, respectively. While both models optimize their parameters by predicting contextual words, the models differ in terms of architecture and objective function. SKIP iterates through each word w_i at index i, and predicts each of the neighboring words up to a distance c. More formally, given a document of T words, the model optimizes its parameters by maximizing the log likelihood function:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-c \leq j \leq c, \\ j \neq 0}} \log p(w_{t+j} \mid w_t) \tag{5.2}$$

where the probability $p(w_{t+j} \mid w_t)$ is the output probability given by the network. The log likelihood function is optimized using gradient descend.

CBOW is similar to SKIP, in the sense that it uses word vectors to predict surrounding words, but predicts each word w_i conditioned on all surrounding words up to a distance of c. That is, we estimate the parameters that maximize the probability $p(w_t \mid w_{t-c}, ..., w_{t+c})$.

To use this information as features in our summarization model, we added to the representation of each sentence a vector consisting in the average of the vectors representing each word in that sentence. Each word is described by 50-features vector.

We have also experimented using a distributed representation of sentences [104], but the results were worse than averaging word vectors due to overfitting.

To create the models, we used articles from the New York Times covering a 16-year period from January of 1994 to December of 2010, included in the English Gigaword Fifth Edition [165]. Since the results obtained with both CBOW and SKIP models were very similar, we opted to present only the results with the SKIP model.

5.2.3 Experiments

We evaluate our work in two distinct ways: through the automatic estimation, using ROUGE; and through a human study, designed according to two previous reference studies [156, 143], using the Amazon Mechanical Turk.

To empirically analyse the performance of our event-based multi-document summarization methods, we use two standard evaluation datasets already used in Section 5.1: DUC 2007 and TAC 2009. However, the set of events types occurring in evaluation datasets only partially overlaps with the event types detected by our event detector. Hence, we created a subset for each of the evaluation datasets. Tables 5.2 and 5.3 identify the selected topics.

Table 5.2: Subset of DUC 2007 topics containing several event types in the ACE 2005 list.

Topic	Description
D0705A	Basque separatism.
D0706B	Burma government change 1988.
D0712C	"Death sentence" on Salman Rushdie.
D0718D	Starbucks Coffee attempted to expand and diversify through joint ventures, acquisitions or subsidiaries.
D0721E	Matthew Sheppard's death.
D0741I	Day trader killing spree.
D0742J	John Kennedy Jr. Dies in plane crash.

Table 5.3: Subset of TAC2009 topics containing several event types in the ACE 2005 list.

Topic	Description
D0904A	Widespread activities of white supremacists and the efforts of those opposed to them to prevent violence.
D0910B	Struggle between Tamil rebels and the government of Sri Lanka.
D0912C	Anti-war protest efforts of Cindy Sheehan.
D0914C	Attacks on Egypt's Sinai Penninsula resorts targetting Israeli tourists.
D0915C	Attacks on Iraqi voting stations.
D0922D	US Patriot Act, passed shortly after the September 11, 2001 terrorist attacks.
D0934G	Death of Yassar Arafat.
D0938G	Preparations and planning for World Trade Center Memorial
D0939H	Glendale train crash.
D0943H	Trial for two suspects in Air India bombings.

5.2.3.1 Evaluation Setup

To assess the performance of our methods, we compare them against other representative models as we did in the previous section: namely MEAD, MMR, Expected n-call@k [112], and the Portfolio Theory [213]. Since none of these baselines included an event-based summarizer nor a topic-based summarizer, we also compare our results against Filatova's event-based summarizer [64] (our implementation), and TopicSum [80]. We opted to provide the results of LexRank [61] to give a better perception over the single-document results presented in the previous chapters.

Filatova's event-based summarizer is a summarization method that also explores event information in a pattern-based way. TopicSum models topics in documents and uses them for content selection, making it close to event-based summarization. LexRank is well-known PageRank-based summarization method often used as baseline. As our baseline method, we used the straightforward idea of combining all input documents into a single one and then submit the resulting document to the single-document summarization method.

Internally, our event-based method is built upon the KP-Centrality method which uses a distance metric to compute semantic similarity between the sentences. In these experiments, we explored the several metrics presented by Ribeiro and de Matos [177], but only present the results using the Euclidean distance in this chapter, as it was the best-performing one in this context. Appendix B includes the results using other distance metrics.

To evaluate informativeness of the summaries, we used ROUGE-1. For the human evaluation, we used the Amazon Mechanical Turk. We assess the performance of the various models by generating summaries with 250 words.

In the next sections, we analyze the results of the automatic evaluation and of the human study. Although we have experimented both the single-layer and waterfall architectures in both datasets, we only present the best performing model for each dataset, but we included the complete set of results in Appendix B.

5.2.3.2 Automatic Evaluation

The second column on the right Table 5.4 provides the results on the DUC 2007 dataset using the waterfall summarization model. Our first observation is that our proposed approach, even without using any event information, filtering or the temporal dilation of the size of the initial and intermediate summaries, achieves better results than the baseline. Note that, although the presented results are for the waterfall architecture, the single-layer approach using all features (event information and filtering in addition to average word embeddings of sentences

and temporal dilation) also achieved better results than the baseline (0.3522 ROUGE-1 score). The same does not happen for other summarization models: MEAD and Portfolio achieved better results than the baseline, but Filatova's event-based summarizer, MMR ($\lambda = 0.3$ was the best performing configuration), Expected n-call@k, TopicSum, and LexRank did not.

Another important aspect is that, in the DUC 2007, all variants (except for the ones using event information without including event filtering, word embeddings, and temporal dilatation) obtain better results than the one not using event information or temporal dilation.

After we observed the summaries, we find out that the intermediate summaries were not large enough to keep all important events till the generation of the final summary. At the same time, the sentences describing the same event types were not exactly the same events, but follow up events (which are semantic similar), such as a new strike, or another company acquisition.

The best performing baseline was MEAD and only achieved a performance similar to the default model without event information or the temporal dilation. The best results in the DUC 2007 were obtained when using the average word embeddings of the sentences (SKIP model) combined with the event distribution scores and using event filtering and temporal dilation.

Figure 5.4 shows an example of a summary produced by our best method on the DUC 2007 dataset and the corresponding reference summary.

The right column of Table 5.4 presents the obtained results on the TAC 2009 dataset. Note that, in this dataset, our best results were achieved using the single-layer architecture instead of the waterfall architecture. Nonetheless, the best result achieved by the waterfall approach (using all features) was better than our baseline (0.5163 ROUGE-1 score). On the other hand, all other approaches, achieved worse results than the baseline. The results in the TAC 2009 results exhibit the same behavior in term of features and temporal dilation observed in the DUC 2007 dataset: the best results use all features and temporal dilation of the size of the initial and intermediate summaries.

The event filtering consistently lower the results in the TAC 2009. The smaller number of documents to summarize 10 vs. 25 suggest that there is less redundant content in the TAC 2009 than in the DUC 2007. Some of the topics in the TAC 2009 are more complex, in the sense, that there are more relevant events, but with distributed lower relevance of those events making the distinction between primary and secondary events hard even for humans as topic D0910B exemplifies. Under this conditions, an event classification error have more impact in the final outcome and should be avoided. Our event filtering results were also in line with Filatova's event-based summarizer, which had worse performance than Expected n-call@k and MMR on the TAC 2009.

Table 5.4: ROUGE-1 results in the DUC 2007 (waterfall) and TAC 2009 (single-layer).

Features	F.E.	T.D.	DUC2007	TAC2009
default + AWE + events info.	yes	yes	**0.3811**	0.5231
default + AWE + events info.	yes	no	0.3531	0.5298
default + AWE + events info.	no	yes	0.3612	**0.5501**
default + AWE + events info.	no	no	0.3520	0.5075
default + events info.	yes	yes	0.3715	0.5334
default + events info.	yes	no	0.3527	0.5284
default + events info.	no	yes	0.3642	0.5333
default + events info.	no	no	0.3488	0.5125
default + AWE	yes	yes	0.3794	0.5261
default + AWE	yes	no	0.3534	0.5381
default + AWE	no	yes	0.3669	0.5401
default + AWE	no	no	0.3505	0.5224
default	yes	yes	0.3680	0.5251
default	yes	no	0.3516	0.5229
default	no	yes	0.3606	0.5251
default	no	no	0.3516	0.5201
baseline			0.3255	0.4749
MEAD			0.3519	0.4690
Portfolio			0.3492	0.4223
Filatova's event-based summarizer			0.3008	0.3794
MMR			0.2994	0.3697
E.n-call@k			0.2800	0.3638
TopicSum			0.1713	0.2711
LexRank			0.1704	0.2620

Table 5.5: Results of maximum ROUGE-1 scores and of our best performing methods.

#Sent.	Corpus	Oracle	Summarizer
1	TAC 2009	0.2420	0.1929
2		0.4099	0.3100
3		0.5283	0.3870
1	DUC 2007	0.1182	0.0898
2		0.2151	0.1674
3		0.3962	0.2292

We have also observed that when the connection between news documents covering a topic is weak, the cascade method performs worse than the singe-layer. This fact also helps to explain the performance differences between the hierarchical methods and datasets.

In order to give a better perspective over the results shown in Tables 5.4, we need to know the ROUGE-1 of the perfect summary. This result corresponds to the optimal selection of important sentences achievable in the evaluation datasets (oracle) and it is shown in Table 5.5. We also included the results obtained using our best summarizer configuration. These values are obtained by testing all summaries that can be generated and extracting the one with the highest score. The precise calculation of this exponential combination problem is, in the most cases, unfeasible. As a result, we restricted the size of the oracle to 3 sentences. The comparison of results of the oracle and our summarizer's show that our best methods are in the 70-80% range of the oracle summaries.

Another interesting aspect that we observed is related to the representation of dates and numbers when using word embeddings. Since the frequency of this information is low in the used training data, it is not well captured by these models. The result is that this type of information is not well represented in the summaries generated by our methods, when using word embeddings. For example, Figure 5.4 shows an example of a summary produced by our best method on the DUC 2007 dataset and the corresponding reference summary. The reference summary contains four date entities and two money entities and in the automatic summary only one date entity appears.

5.2.3.3 User Study

The initial informativeness evaluation of our multi-document summarization framework was performed using the ROUGE evaluation metric.

The ROUGE metric does not measure how useful the summaries are for humans. To evaluate

Event-based Multi-document Summary

Iranian Foreign Minister Kamal Kharrazi , who made the announcement in New York , and his British counterpart , Robin Cook , had portrayed the move as a way to improve ties that have remained strained over the issue and agreed to exchange ambassadors . LONDON _ The British government said Wednesday that it would continue to press Iran to lift the death sentence against the author Salman Rushdie when its foreign secretary , Robin Cook , meets the Iranian foreign minister in New York on Thursday . VIENNA, Austria (AP) – The European Union on Monday welcomed a move by the Iranian government to distance itself from an Islamic edict calling for British author Salman Rushdie's death even as two senior Iranian clerics said the ruling was irrevocable . The move follows the Iranian government's distancing itself last month from bounties offered for the death of Rushdie and a strong reaction by hard-liners who support the killing of the Booker Prize-winning author . He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair . On February 14, 1989, late Iranian leader Ayatollah Khomeini issued a religious edict, pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel " The Satanic Verses " , which was believed by Moslems as defaming Islam , and exhorting all Moslems to carry out the sentence .

Reference

In 1989, Ayatollah Khomeini of Iran issued a death sentence on British author Salman Rushdie because his book "Satanic Verses" insulted Islamic sanctities. Rushdie was born in India, but his book was banned and his application for a visit was denied. British Airways would not permit Rushdie to fly on its airplanes. Reacting to diplomatic pressures by Britain and other European Nations, Iran announced in 1996 that the death sentence was dropped. President Rafsanjani said there was a difference between a fatwa (ruling) and a hokm (command) and that Khomeini did not mean the sentence to be a command. Despite official retraction of the death sentence, Iranian Islamic fundamentalists continue to demand Rushdie's death. The Khordad Foundation raised the reward for Rushdie's death to 2.5 million dollars and announced, "There is nothing more important to the foundation than seeing Imam Khomeini's decree executed." In 1998, Grand Ayatollah Lankarani and Grand Ayatolla Hamedani said the fatwa must be enforced and no one can reverse it. More than half of Iran's parliament signed a letter saying the death sentence against Rushdie still stands. A hard-line student group offered $333K to anyone who kills Salman Rushdie; residents of a village in northern Iran offered land and carpets to anyone who kills him and thousands of Iranian clerics and students pledged a month's salary toward a bounty. In February 2000, the Islamic Revolutionary Guard said in a radio report that the death sentence was still in force and nothing will change it.

Figure 5.4: Example of summary produced by our summarizer and the reference summary from the Topic D0712C DUC 2007 - "Death sentence" on Salman Rushdie.

usefulness, we needed a set of summaries from our event-based summarizer with the corresponding evaluation scores. We also needed a similar set for the baseline system to establish a proper comparison. Obtaining such sets presents both conceptual and practical difficulties.

Defining usefulness or relevance of summaries are subjective decisions of each reader that can be influenced by their background.

Our solution was to use multiple judges for the same news story and provide a Likert scale to assign a score to each question. We used a five-level Likert scale, ranging from strongly disagree (1) to strongly agree (5).

We used the AMT service to recruit and manage our judges. To the best of our knowledge, this has not been done before for this purpose. Each HIT consisted of answering 9 evaluation questions. Evaluating one summary was a HIT and it paid $0.05 if accepted. We selected the reference summaries from each topic of the subsets of the TAC 2009 and DUC 2007 datasets.

We obtained 8 summaries for each topic: one using our event-based summarizer, another using the reference summary, and 7 using the baseline systems. Then, we created 5 HITs for each of the 17 topics. An individual judge could only do one HIT per summary of a topic and summarizer.

The use of Mechanical Turk created the practical problem of uneven quality of the judges: some of the judges used shortcuts to accomplish a HIT, producing meaningless results. We used several rules to weed out bad HITs. One of the most widely strategies used to filter bad HITs is to limit the access to bad judges. We only accepted judges having a HIT approval rate for all HITs executed greater than or equal to 95%. This approval rate is not very meaningful for a low number of accomplished HITs; however, for a large number of completed tasks it is indicative of consistently good workers. Judges could only get access to our HITs if they had a number of HITs approved greater than or equal to one thousand. Tools for this purpose are available in the Mechanical Turks HIT design interface. Despite this strict requirement to be allowed to judge our HITs, there were still some judges taking shortcuts. Very fast work completion time is usually an indicator of a bad HIT. We filtered work completed in less than thirty seconds. This number corresponds to 25% of the average completion time of the HITs. Another very simple rule that we used was to filter HITs with missing answers. While these requirements were sufficient to filter bad judges in more than 90% of the cases, we also wanted to detect HITs that could contain random answers. To this end, we included two additional rules. When a judge submitted 5 or more HITs with same answer to all questions (e.g., 1 or 5), we opted to filter out their work. Another indicator of bad HITs is the lack of consistency between the overall quality of the summary answer and the answers to the other questions. We opted to also exclude HITs where the answer to the overall quality of the summary question was higher or lower than all other answers to the HITs by 2 or more points. For instance, if the overall quality of the summary equals 5 and the values for the remaining answers were either 1, 2, or 3, this was considered a bad HIT. After applying all these rules, we filtered 29.9% of the submitted HITs and asked another judges to perform

them. As a result, we were able to keep 99.9% of HITs.

We created a "Gold Standard" set of 680 annotated summaries. For each summary, we used the 5 questions' quality description developed by Nenkova [156] to assess the linguistic quality of the summaries. In addition, we developed an additional set of questions to evaluate the usefulness of the summaries based on the work of McKeown et al. [143] and we included a question to measure the overall quality of the summary.

To be more precise, each HIT had a description of the task. It indicated that we were conducting a survey about computer-generated summaries. The evaluation was performed without reference to the original texts. We did not distinguish the reference summaries from the automatically generated summaries.

Each HIT contains the following questions:

1. To which degree do you agree with the following information:

   (a) *Background* - Familiarity with the main topic before reading it, that is: "I was familiar with the main topic of the summary before reading it".

2. Please indicate to which degree do you agree that the summary possessed the following qualities:

   (a) *Usefulness* - The summary informs you about the <TopicDescription> (variable replaced by the description of the topic included in Tables 5.2 and 5.3)

   (b) *Coherence* - The summary is well-structured and organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

   (c) *Referential clarity* - It should be easy to identify in the summary to whom or what the pronouns and noun phrases are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. Thus, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.

   (d) *Non-redundancy* - There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Barack Obama") when a pronoun ("he") would suffice.

   (e) *Focus* - The summary should not have extraneous information.

   (f) *Context Coverage* - The summary should cover all main events of a story and give a brief context about them.

(g) *Grammaticality* - The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments and missing components) that make the text difficult to read.

(h) *Overall* - What is the overall quality of the summary?

Table 5.6: DUC 2007 human results.

Question	Reference	MEAD	MMR	E.ncall	Portf.	EventSum	Filatova et al.	TopicSum	LexRank
Background	3.0000	2.7420	2.9260	2.6818	3.1250	**3.1430**	2.7647	2.7273	3.0882
Usefulness	3.9655	3.4194	3.5556	3.5000	3.7500	**4.0000**	3.4706	2.9697	3.2059
Coherence	3.7586	2.9032	3.5185	3.3636	3.3750	**3.8571**	3.6176	3.2424	2.7059
Referential Clarity	**3.9655**	3.4194	3.4815	3.3636	3.5833	3.8214	3.6471	2.9091	3.1176
Non-redundancy	3.6552	2.9032	3.4815	3.1363	3.4583	**3.8571**	3.4706	2.9697	3.0588
Focus	3.8276	3.7742	3.7407	3.6818	3.7500	**3.9286**	3.4706	2.8485	2.8235
Context Coverage	4.0344	3.4516	3.6667	3.4545	3.7083	**4.1071**	3.5882	2.8788	3.0882
Grammaticality	**4.1379**	3.7097	3.8889	3.7727	4.0000	3.8929	3.5294	2.9091	3.3235
Overall	**4.0000**	3.2258	3.6667	3.4091	3.5833	3.8929	3.6176	2.8788	2.8824

Table 5.7: TAC 2009 human results.

Question	Reference	MEAD	MMR	E.ncall	Portf.	EventSum	Filatova et al.	TopicSum	LexRank
Background	2.7368	2.9250	2.8485	2.9189	3.0000	**3.0625**	2.7234	2.6596	2.6458
Usefulness	3.6842	3.9750	3.6970	3.5946	3.7368	**4.0313**	3.6595	3.0638	3.5417
Coherence	**3.7895**	3.6500	3.6667	3.4865	3.5000	3.7812	3.6383	3.4894	2.9375
Referential Clarity	**3.9737**	3.8750	3.6667	3.5946	3.3947	3.9688	3.5957	3.1489	3.3333
Non-redundancy	**4.1053**	3.5500	3.7879	3.3243	3.4210	3.7188	3.8085	3.2766	3.6250
Focus	3.8158	**4.0750**	3.6667	3.8378	3.8684	4.0000	3.6596	2.8510	3.2500
Context Coverage	3.4737	3.8500	3.6364	3.5946	3.7368	**3.9688**	3.8085	3.1702	3.4792
Grammaticality	**4.0789**	3.9750	3.8485	3.8649	3.8684	4.0313	3.8298	3.1064	3.5833
Overall	3.6842	3.7750	3.6970	3.6486	3.7105	**3.8125**	3.8085	3.1915	3.4167

Tables 5.6 and 5.7 show the average scores obtained in the user study. As we can observe in both tables, the judges rated our event-based multi-document summaries as more useful than reference summaries and the baseline systems. They also reported that they better recognize the topic of the summaries using our summarization method.

In terms of coherence of the summaries, event-based summaries were perceived as more coherent than the references for DUC 2007. While on TAC 2009, the judges judged the coherence of our event-based summaries to be nearly the same. We empirically observed that the waterfall method produces more coherent summaries than the single-layer method, which is explained in part by the fact that most of the extracted sentences belong to few documents (in general, the most recent ones).

The reference summaries clearly outperformed our summaries in the Referential Clarity and Grammaticality categories. These are expected results because the reference summaries do not contain news source names (possibly motivated by the presence in the generated summaries of extracts like "VIENNA, Austria (AP)") and because all pronoun references can be resolved.

The evaluation scores for the Focus category highlight an important difference in the topics of the datasets. While in TAC 2009 most topics describe several equal-importance sub-topics/events spread in time, there is a single main topic center on a date in several topics of DUC 2007. One implication is that our event-based multi-document summaries do not discard the sub-topics, which penalizes the Focus score in the TAC 2009 dataset when compared to the centroid-based method (MEAD) that selected the sentences for the summary using a single topic (centroid). Another implication is that increasing the focus in a single sub-topic can reduce the Context Coverage. However the results are not conclusive.

Even though the overall results are higher for our event-based multi-document summaries in TAC 2009, we cannot conclude that our method is better than the reference. The reason lies in the smaller size of reference summaries when compared to the remaining summaries (100 vs. 250 words).

Among the event-based and topic-based baselines, the human evaluation clearly shows that the Filatova et al. event-based method performed better than the topic based summarizer (TopicSum). More interesting is the fact that the overall human score of the Filatova et al. event-based were either the best or second best baseline.

5.3 Conclusions

In this work, we explore a multi-document summarization framework based on event information and word embeddings that achieves performance above the state-of-the-art.

The multi-document summarization framework was developed by extending a single-document summarization method, KP-Centrality, in two hierarchical ways: single-layer and waterfall. The single-layer approach combines the summaries of each input document to produce the final summary. The waterfall approach combines the summaries of the input documents in a cascade fashion, in accordance with the temporal sequence of the documents. Event information is used in two different ways: in a filtering stage and to improve sentence representation as features of the summarization model. Related to event information, we also explored the temporal sequence of the input documents by increasing the size of the initial and intermediate summaries, used by our framework. To better capture content/event information expressed using different terms, we use two distributed representations of text: the skip-ngram model, the continuous bag-of-words model, and the distributed representation of

sentences. Event detection is based on the Fuzzy Fingerprint method and trained on the ACE 2005 Corpus.

To evaluate this multi-document summarization framework, we used two different setups: an automatic evaluation of the informativeness of the summaries using ROUGE-1, and a user study.

Concerning the automatic evaluation of the informativeness, results show that the proposed framework achieves better results than previous models. To this contributed, not only the single-document summarization method on which our multi-document approach is based, but also the use of event information and the better representation of text. Note that a simple baseline that combines all input documents and summarizes the resulting meta-document achieves better results than all other approaches in the TAC 2009 dataset and also achieves better results than five of the reference methods in the DUC 2007 dataset. Nevertheless, our best performing configurations relative improvement in ROUGE-1 scores of 16% for TAC 2009 and of 17% for DUC 2007 (8% for TAC 2009 and DUC 2007 over the performance of the reference systems).

In what concerns the human study, the judges preferred our event-based summaries over all automatically generated summaries, which included other event-based summaries produced by our own implementation of Filatova et al. [64] method. Moreover, in the TAC 2009 dataset, the summaries generated by the proposed methods were even preferred over the reference summaries. In terms of usefulness, our event-based summaries were again preferred over all other summaries, including the reference summaries in both datasets. This is related to the scores obtained for context coverage, where our event-based summaries obtained the highest scores. It is also interesting to observe that, although being extractive summaries, as it happens in all other approaches, our summaries obtained high scores on readability aspects such as grammaticality, referential clarity, and coherence. In fact, they were better than all other automatically generated summaries (except for Portfolio, on grammaticality, in DUC 2007). The best coherence score achieved in DUC 2007 might be related to the use of the waterfall architecture, that boosted the number of sentences selected from the last documents (the most recent ones). Concerning grammaticality, we believe that our event-based method could be improved by the inclusion of a pre-filtering step to remove news sources and datelines.

Our experiments showed that the use of event information combined with a distributed text representation (the SKIP model) further improved a generic multi-document summarization approach above state-of-the-art. Although we propose two different strategies for developing our multi-document methods, single-layer and waterfall, the best results were not achieved by the same architecture in the evaluation datasets because waterfall approach seems to be preferable to summarize large number of documents (e.g., 25 documents) and the single-layer

seems more suitable for small number of documents (e.g., 10 documents). We confirmed this tendency by reducing the number of documents in the DUC 2007 to for example 10. In this situation the single-layer architecture (ROUGE-1 score: 0.3427) was better than the waterfall architecture (ROUGE-1 score: 0.3412). Nevertheless, both architectures achieved better results than the baseline and the reference systems. Analysis of the results also suggests that the waterfall model offers the best trade-off between performance and redundancy.

In the next chapter we will present the overall conclusions of this thesis. We will also propose directions for future work.

6 Conclusions and Future Work

> *The press, the machine, the railway, the telegraph are premises whose thousand-year conclusion no one has yet dared to draw.*

Friedrich Nietzsche , *German Philosopher (1844-1900)*

The large amount of news published every day exceeds the human reading capacity and motivated pioneer multi-document summarization efforts. The majority of multi-document summarization efforts are extractive, that is, the summarization methods select the sentences regarded as the most important from a set of input documents.

Despite the large body of work on summarization in the news domain, most works do not explore the fact that news documents describe events [71] (e.g., marriage, election). The few research works that explore event information in summarization, have either pattern-based (rules) or clustering-based methods to include event information. The major problem of these approaches is twofold: the inherent difficulty to relate different descriptions of the same event and identify different events covering the same event type (e.g., meeting, election) due to different lexical realizations.

We show that identifying sentences containing detectable events is useful to improve summarization. Since news documents' main focus is on describing events, sentences not containing (detectable) events are good candidates to be excluded from the summaries. At the same time, duplicate events should not appear in summaries. Our multi-document summarization framework improves the summarization of documents containing detectable events, achieving state-of-the-art results in both single and multi-document summarization. We explored three different ways to integrate event information. Filtering sentences without events before feeding them to the summarizer is one way. Another way is to include a list of values measuring how close is a sentence to the event types as features of the summarizer. The third way is to combine filtering of sentences and event-based features on the classifier. As a result, we observed significantly improvements between using event information on summarization (event-based summarization) and multi-document summarization without event-based information. How-

ever, the main limitation of our event-based summarization framework is summarizing sets of news documents covering events that do not occur in the training data of our event detector.

Our event detector, based on fuzzy fingerprint method, seems to be more robust to the addition of new event types than standard machine learning classifiers including SVM and Random Forests. These results are justified by the properties of the training data, in particular, low number of examples of sentences describing events and, at the same time, a dominant number of sentences that do not contain events (reaching more than 76% of the sentences).

We demonstrated that event information improves generic summarization. Moreover, the inclusion of key phrases, as new passages to guide a centrality-based summarizer, lead to the largest improvements in the summarization results. The additional semantic features proposed for the key phrase extraction, such as rhetorical signals (e.g., words that give emphasis), also had a clear impact on the summarization results. In our work, we show that AKE and event detection have a direct impact on the informativeness of the summarization, which we measured automatically using ROUGE. The user study complemented the automatic evaluation with other evaluation metrics, including usefulness, coherence, context coverage, and referential clarity. Despite the good results on these metrics, in particular coherence, our summarization methods do not have any re-ordering mechanism of the sentences to improve coherence. This result is in line with the Christensen et al. results' [52], which claims that simple re-ordering of sentences has little impact on the summary coherence. Nevertheless, recent work using event-based information to sentence reordering [230] could be adapted to further improve our summarization framework.

6.1 Extensibility of Event-based Multi-document Summarization to other domains

In this work, we applied event-based summarization methods to the news domain. We believe that the same methods are applicable to other domains, including business/technical reports, short stories, and research articles.

Most domains are action or event oriented, because the documents from these domains are usually created to report human activities. Our method can be adapted to domains focused on object description. The rationale is the following: it is possible to define a set of "anchoring concepts", which are equivalent to events. These concepts can be in most cases obtained from domain ontologies. Of course, there are some domains where an ontology is not readily available, such as in the biomedical domain. In those cases, we believe it is possible to use the list of keywords or key phrases included in the documents, as the list of concepts. Each concept would have associated a fuzzy fingerprint, which would be based on the sentences

containing the corresponding concept. After building the fuzzy fingerprints of each concept, the concept classifier is ready to be used for summarization. Then, it is straightforward to apply our event-based summarization methods proposed to the new domain.

While in our work we have not considered the internal structure of events (e.g., sub-events [32]), such structure could be inferred from ontologies such as Wordnet [62]. By taking advantage of the structure of events/concepts, it becomes feasible to detect both main event/concept and sub-events/sub-concepts using the fuzzy fingerprints method . The sub-events information could be propagated to help the detection of the main events. This propagation can be accomplished in multiple ways. One possibility is to expand the fuzzy fingerprints of the events with new features and update the existing ones. Another possibility is to add an additional weight parameter to the fuzzifying function and rebuild the fuzzy fingerprints. Although our work did not explore new methods to learn new and update existing fuzzy fingerprints, we still think that it would be possible to perform this learning task in the news domain and in other domains. For this purpose, we suggest to combine active learning methods [59] with tools that perform co-reference resolution of events and named-entities. The reason behind this suggestion is the following: since co-reference resolution systems detect the same the events/concepts, it is possible to update the fuzzy fingerprints or in some cases create new fuzzy fingerprints for new events/concepts.

In our preliminary experiments to adapt our methods to other domains, we observed that for long documents with multiple sections or chapters, it is preferable to treat each section or chapter as a individual document and apply our multi-document summarization methods rather than applying the single-document summarization methods directly.

6.2 Future work

A possible future research direction is compression of the sentences selected by our extractive summarizer. The process of compressing sentences should use event information to delete irrelevant words and to shorten long phrases. A solution to adequately compress sentences using event information entails solving multiple subproblems. For example, the identification of the relation between named entities (relationship extraction), identification of sentences mentioning the same event (event co-reference), and extract when the events take place (temporal information extraction), among other problems. Solving each of these problems is also useful for building a knowledge base of events. Deciding on how to build, update a knowledge base of events is not an easy task, because it is necessary to model time, location, and the likelihood of the information being correct. Still, a knowledge base of events should be useful to provide a more compressed representation formats for summarization, such as summarizing temporal spans of information. Another complementary source of useful information to

compress sentences and events would be to use additional event information. For example, it would be useful to perform event co-reference analysis to detect if two event mentions or sentences describe the same event. At the same time, it would also be very useful to look at the internal structure of events (e.g., sub-events) to further improve the retrieval of important content.

Another interesting research direction for future work is to adapt our event-based multi-document methods to new domains, such as social media, reports (e.g., accidents, military), transcription of meetings. These domains do not follow the inverted pyramid structure used by journalists, where the most important content appear in the top of the documents, and the less relevant information appear in the bottom. The adaptation entails obtaining new training data for the key phrase extraction and event detector tools. For some domains, such as social media, the adaptation might also require the replacement of the POS tagger and NER. Another alternative would be to normalize the text, which has large proportion of abbreviations, alternative spellings, novel words, and other non-canonical language; to more readable "clean" text, and use the same AKE, POS tagger, and NER adopted in our work.

Bibliography

[1] Automatic Content Extraction (ACE) Evaluation. http://www.itl.nist.gov/iad/mig/tests/ace/, 1993. [Online; accessed 17-July-2013].

[2] TIPSTER Text Summarization Evaluation (SUMMAC). http://www-nlpir.nist.gov/related_projects/tipster_summac/, 1998. [Online; accessed 17-July-2013].

[3] Message Understanding Conference (MUC). http://www.itl.nist.gov/iaui/894.02/related_projects/muc/, 2001. [Online; accessed 10-July-2013].

[4] DUC. http://duc.nist.gov/, 2008. [Online; accessed 10-July-2013].

[5] The New York times article: "Can't Grasp Credit Crisis? Join the Club". http://www.nytimes.com/2008/03/19/business/19leonhardt.html?pagewanted=all, 2008. [Online; accessed 10-July-2013].

[6] Death of Muammar Gaddafi reported by BBC. http://www.bbc.co.uk/news/world-africa-15397812, 2011. [Online; accessed 25-July-2013].

[7] MorphAdorner. http://morphadorner.northwestern.edu/documentation/, 2011. [Online; accessed 25-July-2013].

[8] Smooth-nlp Toolkit. https://code.google.com/p/smooth-nlp/, 2011. [Online; accessed 17-July-2013].

[9] Croatia becomes 28^{th} European Union member. http://web.archive.org/web/20130702204911/http://www.bbc.co.uk/news/world-europe-23118035, 2013. [Online; accessed 25-July-2013].

[10] DUC 2001. http://www-nlpir.nist.gov/projects/duc/data/2001_data.html, 2013. [Online; accessed 25-July-2013].

[11] Freebase Schema. http://www.freebase.com/schema, 2013. [Online; accessed 17-July-2013].

[12] MEAD. http://www.summarization.com/mead/, 2013. [Online; accessed 10-July-2013].

[13] Repository of Automatic Key Phrase Extraction Corpora. https://github.com/snkim/AutomaticKeyphraseExtraction, 2013. [Online; accessed 10-July-2013].

[14] TAC. http://www.nist.gov/tac/, 2013. [Online; accessed 10-July-2013].

[15] Text REtrieval Conference (TREC). http://trec.nist.gov/, 2013. [Online; accessed 10-July-2013].

[16] Twitter. https://twitter.com/, 2013. [Online; accessed 17-July-2013].

[17] UN Food and Agriculture Organization documents. http://www.fao.org/documents/, 2013. [Online; accessed 10-July-2013].

[18] Weibo. http://www.weibo.com, 2013. [Online; accessed 17-July-2013].

[19] Wikipedia. http://en.wikipedia.org/wiki/Wikipedia, 2013. [Online; accessed 10-July-2013].

[20] Number of Ebola infections in west Africa passes 16,000. http://web.archive.org/web/20141208152036/http://www.theguardian.com/world/2014/nov/29/ebola-infections-west-africa-16000, 2014. [Online; accessed 8-December-2014].

[21] Amr Ahmed. *Modeling Users and Content: Structured Probabilistic Representation, and Scalable Online Inference Algorithms.* PhD thesis, CMU, 2011.

[22] Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J Smola, and Choon Hui Teo. Unified analysis of streaming news. In *Proceedings of the* 20^{th} *international conference on World Wide Web*, pages 267–276. ACM, 2011.

[23] Amr Ahmed and Eric P. Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the* 26^{th} *International Conference on Uncertainty in Artificial Intelligence*, 2010.

[24] James Allan. *Automatic hypertext construction.* PhD thesis, Cornell University, Ithaca, NY, 1995.

[25] James Allan. Automatic hypertext link typing. In *Proceedings of the ACM Hypertext Conference*, pages 42–52, Washington DC, 1996.

[26] James Allan. Building hypertexts using information retrieval. *Information Processing and Management*, 33(2):145–159, 1997.

[27] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, Brian Archibald, and Mike Scudder. Topic Detection and Tracking Pilot Study Final Report. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, 1998.

[28] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR Conference on Research and development in Information Retrieval*, SIGIR '01, pages 10–18, New York, NY, USA, 2001. ACM.

[29] Miguel Almeida and Andre Martins. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 196–206, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[30] Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. Exploratory search using timelines. In *SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop*, number 1. ACM, 2007.

[31] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 97, New York, New York, USA, 2009. ACM Press.

[32] Jun Araki, Zhengzhong Liu, Eduard Hovy, and Teruko Mitamura. Detecting subevent structure for event coreference resolution. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[33] Ricardo Baeza-Yates, Paolo Boldi, and Carlos Castillo. Generalizing PageRank: damping functions for link-based ranking algorithms. In *SIGIR'06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 308–315, New York, NY, USA, 2006. ACM.

[34] Fernando Batista, Isabel Trancoso, and Nuno Mamede. Automatic recovery of punctuation marks and capitalization information for iberian languages. In *I Joint SIGIL/Microsoft Workshop on Speech An Language Technologies for Iberian Languages*, pages 99–102, 2009.

[35] P. B. Baxendale. Machine-made index for technical literature - an experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

[36] Giang Binh Tran. Structured summarization for news events. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 343–348, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[37] Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. Predicting relevant news events for timeline summaries. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 91–92, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[38] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the* 23rd *International Conference on Machine learning*, pages 113–120. ACM, 2006.

[39] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[40] Rodrigo A. Botafogo. Cluster analysis for hypertext systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, pages 116–125, New York, NY, USA, 1993. ACM.

[41] Rodrigo A. Botafogo, Ehud Rivlin, and Ben Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems (TOIS)*, 10(2):142–180, April 1992.

[42] Leo Breiman. Bagging Predictors. *Machine learning*, 24(2):123–140, 1996.

[43] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[44] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[45] Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the* 21^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, 1998.

[46] Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D. Brown, Tom Pierce, and Xin Liu. CMU Approach to TDT: Segmentation, Detection, and Tracking. In *Proceedings of the 1999 Darpa Broadcast News Conference*, 1998.

[47] Angel X. Chang and Christopher D. Manning. SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation (LREC 2012)*, May 2012.

[48] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the* 34^{th} *Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[49] Lee-Feng Chien. Pat-tree-based keyword extraction for chinese information retrieval. In *Proceedings of the* 20^{th} *annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'97, pages 50–58, New York, NY, USA, 1997. ACM.

[50] Hai Leong Chieu and Yoong Keok Lee. Query based event extraction along a timeline. In *Proceedings of the* 27^{th} *annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 425–432, New York, NY, USA, 2004. ACM.

[51] Heidi Christensen, Yoshihiko Gotoh, BalaKrishna Kolluru, and Steve Renals. Are Extractive Text Summarisation Techniques Portable To Broadcast News? In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pages 489–494. IEEE, 2003.

[52] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Towards coherent multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013)*, 2013.

[53] Thomas M. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[54] Naomi Daniel, Dragomir Radev, and Timothy Allison. Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL 03 on Text Summarization Workshop - Volume 5*, HLT-NAACL-DUC '03, pages 9–16, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[55] Dipanjan Das and André FT Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007.

[56] Ernesto D'Avanzo and Bernardo Magnini. A keyphrase-based approach to summarization: the lake system at duc-2005. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, 2005.

[57] Ramón Díaz-Uriarte and Sara Alvarez De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[58] Robert L. Donaway, Kevin W. Drummey, and Laura A. Mather. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 69–78, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[59] Pinar Donmez, Jaime Carbonell, and Paul N. Bennett. Dual strategy active learning. In JoostN. Kok, Jacek Koronacki, RaomonLopezde Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Machine Learning: ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 116–127. Springer Berlin Heidelberg, 2007.

[60] Brigitte Endres-Niggemeyer. *Summarizing Information*. Springer, 1998.

[61] Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[62] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[63] Ao Feng and James Allan. Finding and linking incidents in news. In *Proceedings of the* 16^{th} *ACM Conference on information and knowledge management*, CIKM '07, page 821, New York, New York, USA, 2007. ACM Press.

[64] Elena Filatova and Vasileios Hatzivassiloglou. Event-based extractive summarization. In *Proceedings of ACL Workshop on Summarization*, pages 104–111, 2004.

[65] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the* 43^{rd} *Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[66] Massimo Franceschet. PageRank: Standing on the Shoulders of Giants. *Communications of the ACM*, 54(6), 2011.

[67] Edward B. Fry and Jacqueline E. Kress. *The New Reading Teacher's Book of Lists*, volume 55. Prentice Hall, 1990.

[68] Michel Galley. A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 364–372. Association for Computational Linguistics, 2006.

[69] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tür. ClusterRank: A Graph Based Method for Meeting Summarization. In *Proceedings of the* 10^{th} *Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 1499–1502. ISCA, 2009.

[70] Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 710–720, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[71] Goran Glavaš and Jan Šnajder. Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications*, 2014.

[72] Yihong Gong and Xin Liu. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proceedings of the* 24^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'01, pages 19–25, New York, NY, USA, 2001. ACM.

[73] David Graff. The aquaint corpus of english news text. *Linguistic Data Consortium*, 2002.

[74] David Graff, John Garofolo, Jonathan Fiscus, William Fisher, and David Pallett. HUB4 Broadcast News Speech, Linguistic Data Consortium, Philadelphia. *Linguistic Data Consortium*, 1996.

[75] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *Proceedings of the* 16^{th} *conference on Computational linguistics - Volume 1*, COLING '96, pages 466–471, Stroudsburg, PA, USA, 1996. Association for Computational Linguistics.

[76] Robert Gunning. *The technique of clear writing.* Brookline Books, New York, USA, 1952.

[77] Robert Gunning. The fog index after twenty years. *Journal of Business Communication*, 1969.

[78] Shengbo Guo and Scott Sanner. Probabilistic latent maximal marginal relevance. In *Proceedings of the* 33^{rd} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 833–834, New York, NY, USA, 2010. ACM.

[79] David Guthrie, Mark Hepple, and Wei Liu. Efficient minimal perfect hash language models. In *Proceedings of the* 7^{th} *International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[80] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 362–370, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[81] Donna Harman and Mark Liberman. TIPSTER Complete. `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC93T3A`, 1993. [Online; accessed 17-July-2013].

[82] Kazi Saidul Hasan and Vincent Ng. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the* 23^{rd} *International Conference on Computational Linguistics: Posters*, pages 365–373, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[83] Philip J. Hayes, Laura E. Knecht, and Monica J. Cellio. A news story categorization system. In *Proceedings of the* 2^{nd} *conference on Applied Natural Language Processing*, ANLC '88, pages 9–17, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics.

[84] Makoto Hirohata, Yosuke Shinnaka, Koji Iwano, and Sadaoki Furui. Sentence-extractive automatic speech summarization and evaluation techniques. *Speech Communication*, 48:1151–1161, 2006.

[85] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the* 22^{nd} *annual international ACM SIGIR Conference on Research and development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[86] Nuno Homem and Joao Paulo Carvalho. Authorship identification and author fuzzy "fingerprints". In *Proceedings of 2011 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*, pages 1–6. IEEE, 2011.

[87] Yu Hong, Jianfeng Zhang, Bin Ma, and Jianmin Yao. Using cross-entity inference to improve event extraction. In *Proceedings of the* 49^{th} *Annual Meeting of the Association for Computational Linguistcs*, number 4 in ACL 2011, pages 1127–1136, Portland, Oregon, USA, 2011.

[88] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223, 2003.

[89] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. The icsi meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 1, pages I–364. IEEE, 2003.

[90] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the* 23^{rd} *Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '00, pages 41–48. ACM, 2000.

[91] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In *Proceedings of the* 49^{th} *Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 1148–1158, Portland, Oregon, USA, 2011.

[92] Igor Jurišica. How to Retrieve Relevant Information? In Russel Greiner, editor, *Proceedings of the AAAI Fall Symposium Series on Relevance*, pages 101–104, 1994.

[93] Remy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. Finding salient dates for building thematic timelines. In *Proceedings of the* 50^{th} *Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 730–739, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[94] Dongwoo Kim and Alice Oh. Topic chains for understanding a news corpus. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 163–176. Springer Berlin Heidelberg, 2011.

[95] Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26. Association for Computational Linguistics, 2010.

[96] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[97] Mikalai Krapivin, Aliaksandr Autaeu, and Maurizio Marchese. Large dataset for keyphrases extraction. Technical report, University of Trento, 2009.

[98] Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of the* 14^{th} *International Conference on Machine Learning*, ICML'97, pages 179–186, 1997.

[99] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[100] Byron Y-L Kuo, Thomas Hentrich, Benjamin M . Good, and Mark D. Wilkinson. Tag clouds for summarizing web search results. In *Proceedings of the* 16^{th} *international conference on World Wide Web*, WWW '07, pages 1203–1204, New York, USA, 2007. ACM Press.

[101] Oren Kurland and Lillian Lee. PageRank without Hyperlinks: Structural Re-Ranking using Links Induced by Language Models. In *Proceedings of the* 28^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2005, pages 306–313, New York, NY, USA, 2005. ACM.

[102] Oren Kurland and Lillian Lee. PageRank without Hyperlinks: Structural Reranking using Links Induced by Language Models. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[103] Martha Larson and Gareth J. F. Jones. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, pages 235–422, 2012.

[104] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. JMLR Workshop and Conference Proceedings, 2014.

[105] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the* 15^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 497–506, New York, NY, USA, 2009. ACM.

[106] David D Lewis and Kimberly A Knowles. Threading electronic mail: A preliminary study. *Information processing & management*, 33(2):209–217, 1997.

[107] Peng Li, Yinglin Wang, Wei Gao, and Jing Jiang. Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1137–1146, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[108] Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. Extractive summarization using inter- and intra- event relevance. In *Proceedings of the* 21^{st} *International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 369–376, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[109] Shasha Liao and Ralph Grishman. Filtered ranking for bootstrapping in event extraction. In *Proceedings of the* 23^{rd} *International Conference on Computational Linguistics (Coling 2010)*, number August, pages 680–688, Beijing, 2010.

[110] Shasha Liao and Ralph Grishman. Using document level cross-event inference to improve event extraction. In *Proceedings of the* 48^{th} *Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, number July, pages 789–797, Uppsala, Sweden, 2010.

[111] George James Lidstone. Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8(182-192):13, 1920.

[112] Kar Wai Lim, Scott Sanner, and Shengbo Guo. On the mathematical relationship between expected n-call@k and the relevance vs. diversity trade-off. In *Proceedings of the* 35^{th} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 1117–1118, New York, NY, USA, 2012. ACM.

[113] Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. Generating Event Storylines from Microblogs. In *Proceedings of the* 21^{st} *ACM International Conference on Information and Knowledge Management*, CIKM '12, page 175, New York, New York, USA, 2012. ACM Press.

[114] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.

[115] Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. An Information-Theoretic Approach to Automatic Evaluation of Summaries. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 463–470. Association for Computational Linguistics, 2006.

[116] Chin-Yew Lin and Eduard Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the* 18^{th} *International Conference on Computational Linguistics*, volume 1 of *Coling 2000*, pages 495–501. Association for Computational Linguistics, 2000.

[117] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 71–78, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[118] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[119] Jimmy Lin, Nitin Madnani, and Bonnie J. Dorr. Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 305–308, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[120] Marina Litvak and Mark Last. Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pages 17–24, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[121] Feifan Liu, Deana Pennell, Fei Liu, and Yang Liu. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 620–628, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[122] Maofu Liu, Wenjie Li, Mingli Wu, and Qin Lu. Extractive summarization based on event term clustering. In *Proceedings of the* 45^{th} *Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 185–188, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[123] Yang Liu and Dilek Hakkani-Tur. Speech summarization. In Gokhan Tur and Renato De Mori, editors, *Spoken language understanding: Systems for extracting semantic information from speech*. Wiley, 2011.

[124] Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 257–266, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[125] Annie Louis and Ani Nenkova. Automatically Evaluating Content Selection in Summarization without Human Models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 306–314. Association for Computational Linguistics, 2009.

[126] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1:309–317, October 1957.

[127] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[128] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[129] Luís Marujo. Reap em português. Master's thesis, Instituto Superior Técnico, 2009.

[130] Luís Marujo, Miguel Bugalho, João P. Neto, Anatole Gershman, and Jaime Carbonell. Hourly traffic prediction of news stories. In *Proceedings of the* 3^{rd} *International Workshop on Context- Aware Recommender Systems held as part of the* 5^{th} *ACM RecSys Conference*, October 2011.

[131] Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking, and João P. Neto. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization. In *Proceedings of the* 8^{th} *International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[132] Luís Marujo, Wang Ling, Anatole Gershman, Jaime Carbonell, João P. Neto, and David Matos. Recognition of named-event passages in news articles. In *Proceedings of the COLING 2012: Demonstration Papers*, pages 329–336, Mumbai, India, December 2012. ACL.

[133] Luís Marujo, José Portêlo, David Martins de Matos, João P. Neto, Anatole Gershman, Jaime Carbonell, Isabel Trancoso, and Bhiksha Raj. Privacy-preserving important passage retrieval. In *PIR14: Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security (SIGIR 2014 Workshop)*. ACM, 2014.

[134] Luís Marujo, José Portêlo, Wang Ling, David Martins de Matos, João P Neto, Anatole Gershman, Jaime Carbonell, Isabel Trancoso, and Bhiksha Raj. Privacy-preserving multi-document summarization. In *PIR15: Privacy-Preserving IR (SIGIR 2015 Workshop)*. ACM, 2015.

[135] Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João P. Neto, Anatole Gershman, and Jaime Carbonell. Key phrase extraction of lightly filtered broadcast news. In *Proceedings of* 15^{th} *International Conference on Text, Speech and Dialogue (TSD 2012)*. Springer, September 2012.

[136] Luís Marujo, Ricardo Ribeiro, David Martins de Matos, João Neto, Anatole Gershman, and Jaime Carbonell. Extending a single-document summarizer to multi-document: a hierarchical approach. In *Proceedings of the* 4^{th} *Joint Conference on Lexical and Computational Semantics*, pages 176–181, Denver, Colorado, June 2015. Association for Computational Linguistics.

[137] Luís Marujo, Márcio Viveiros, and João P. Neto. Keyphrase Cloud Generation of Broadcast News. In *Proceedings of the* 12^{th} *Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*. ISCA, September 2011.

[138] Luís Marujo, João Paulo Carvalho, Anatole Gershman, Jaime Carbonell, João P. Neto, and David Martins de Matos. Textual event detection using fuzzy fingerprints. In P. Angelov, K.T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, and S. Zadrożny, editors, *Intelligent Systems'2014*, volume 322 of *Advances in Intelligent Systems and Computing*, pages 825–836. Springer International Publishing, 2015.

[139] Brij Masand, Gordon Linoff, and David Waltz. Classifying news stories using memory based reasoning. In *Proceedings of the* 15^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pages 59–65, New York, NY, USA, 1992. ACM.

[140] Sameer R. Maskey and Julia Hirschberg. Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization. In *Proceedings of the* 9^{th} *EUROSPEECH - Interspeech 2005*, 2005.

[141] Sameer R. Maskey, Andrew Rosenberg, and Julia Hirschberg. Intonational Phrases for Speech Summarization. In *Proceedings of the* 9^{th} *Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, pages 2430–2433. ISCA, 2008.

[142] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[143] Kathleen McKeown, Rebecca J. Passonneau, David K. Elson, Ani Nenkova, and Julia Hirschberg. Do summaries help? In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 210–217, New York, NY, USA, 2005. ACM.

[144] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In Mitchell Marcus, editor, *Proceedings of the* 2^{nd} *International Conference on Human Language Technology Research (HLT 2002)*, pages 280–285. Morgan Kaufmann, 2002.

[145] Olena Medelyan, Eibe Frank, and Ian H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, EMNLP '09, pages 1318–1327, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[146] Olena Medelyan, Vye Perrone, and Ian H. Witten. Subject metadata support powered by Maui. In *Proceedings of the* 10^{th} *annual joint conference on Digital Libraries*, JCDL '10, pages 407–408, New York, NY, USA, 2010. ACM.

[147] Olena Medelyan and Ian H. Witten. Thesaurus based automatic keyphrase indexing. In *Proceedings of the* 6^{th} *ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '06, page 296, New York, New York, USA, 2006. ACM Press.

[148] Margot Mieskes, Christoph Müller, and Michael Strube. Improving extractive dialogue summarization by utilizing human feedback. In *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*, pages 627–632, 2007.

[149] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411. Barcelona, Spain, 2004.

[150] Rada Mihalcea and Paul Tarau. A Language Independent Algorithm for Single and Multiple Document Summarization. In *Proceedings of the* 2^{nd} *International Joint Conference on Natural Language Processing: Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts*, pages 19–24. Asian Federation of Natural Language Processing, 2005.

[151] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[152] George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, November 1995.

[153] Gabriel Murray, Steve Renals, Jean Carletta, and Johana D. Moore. Evaluating Automatic Summaries of Meeting Recordings. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 33–40. Association for Computational Linguistics, 2005.

[154] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *Proceedings of the* 13^{th} *ACM International Conference on Information and Knowledge Management*, CIKM '04, pages 446–453, New York, NY, USA, 2004. ACM.

[155] Martina Naughton, Nicola Stokes, and Joe Carthy. Investigating statistical techniques for sentence-level event classification. In *Proceedings of the* 22^{nd} *International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 617–624, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[156] Ani Nenkova. Summarization Evaluation for Text and Speech: Issues and Approaches. In *Proceedings of the* 9^{th} *International Conference on Spoken Language Processing (INTERSPEECH 2006)*, pages 1527–1530. ISCA, 2006.

[157] Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.

[158] Ani Nenkova and Kathleen McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012.

[159] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions Speech Language Processing (TSLP)*, 4(2), May 2007.

[160] Ani Nenkova and Rebecca J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, HLT-NAACL'04, pages 145–152, 2004.

[161] ThuyDung Nguyen and Min-Yen Kan. Keyphrase extraction in scientific publications. In DionHoe-Lian Goh, TruHoang Cao, IngeborgTorvik Sølvberg, and Edie Rasmussen, editors, *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, volume 4822 of *Lecture Notes in Computer Science*, pages 317–326. Springer Berlin Heidelberg, 2007.

[162] Paul Over. Introduction to duc-2001: an intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2001 Document Understanding Workshop*, 2001.

[163] Paul Over, Hoa Dang, and Donna Harman. DUC in context. *Information Processing & Management*, 43(6):1506 – 1520, 2007.

[164] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the* 40^{th} *Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[165] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword fifth edition. *Linguistic Data Consortium*, 2011.

[166] Jakub Piskorski, Hristo Tanev, Martin Atkinson, and Erik Van Der Goot. Cluster-centric approach to news event extraction. In *Proceedings of the 2008 conference on New Trends in Multimedia and Network Information Systems*, pages 276–290, Amsterdam, The Netherlands, 2008. IOS Press.

[167] John C Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods*, 1999.

[168] James Pustejovsky, José Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. Timeml: Robust specification of event and temporal expressions in text. *Proceedings of the* 5^{th} *International Workshop on Computational Semantics (IWCS-5)*, 3:28–34, 2003.

[169] J. R. Quinlan. Bagging, boosting, and c4.5. In *Proceedings of the* 13^{th} *National Conference on Artificial Intelligence - Volume 1*, AAAI'96, pages 725–730. AAAI Press, 1996.

[170] John Ross Quinlan. *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[171] Dragomir R. Radev, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. A Description of the CIDR System as Used for TDT-2. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[172] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and

user studies. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*, pages 21–30. Association for Computational Linguistics, 2000.

[173] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938, 2004.

[174] Dragomir R. Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. NewsInEssence: Summarizing Online News Topics. *Communications of the ACM*, 48(10):95–98, 2005.

[175] Kira Radinsky and Sagie Davidovich. Learning to Predict from Textual Data. *Journal of Artificial Intelligence Research*, 45:641–684, 2012.

[176] Kira Radinsky and Eric Horvitz. Mining the Web to Predict Future Events. In *Proceedings of the* 6^{th} *ACM International Conference on Web Search and Data Mining*, WSDM '13, 2013.

[177] Ricardo Ribeiro and David Martins de Matos. Revisiting Centrality-as-Relevance: Support Sets and Similarity as Geometric Proximity. *Journal of Artificial Intelligence Research (JAIR)*, 42:275–308, 2011.

[178] Ricardo Ribeiro, Luís Marujo, David Martins de Matos, João P. Neto, Anatole Gershman, and Jaime Carbonell. Self reinforcement for important passage retrieval. In *Proceedings of the* 36^{th} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 845–848, New York, NY, USA, 2013. ACM.

[179] Korbinian Riedhammer, Benoit Favre, and D Hakkani-Tur. A keyphrase based approach to interactive meeting summarization. In *IEEE Spoken Language Technology Workshop (SLT 2008)*, pages 153–156, 2008.

[180] Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. Long story short – Global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52:801–815, 2010.

[181] Hugo Rosa, Fernando Batista, and Joao P. Carvalho. Twitter Topic Fuzzy Fingerprints. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 776–783. IEEE, 2014.

[182] Timothy N. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[183] Horacio Saggion and Sandra Szasz. The concisus corpus of event summaries. In *Proceedings of the* 8^{th} *international conference on Language Resources and Evaluation (LREC)*, 2012.

[184] Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, and Eric SanJuan. Multilingual summarization evaluation without human models. In *Proceedings of the* 23^{rd} *International Conference on Computational Linguistics: Posters*, COLING '10, pages 1059–1067, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[185] Gerald Salton, Andew Wong, and Chungshu S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[186] Gerard Salton. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the* 16^{th} *annual international ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR'93, pages 49–58, New York, NY, USA, 1993. ACM.

[187] Scott Sanner, Shengbo Guo, Thore Graepel, Sadegh Kharazmi, and Sarvnaz Karimi. Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In *Proceedings of the* 20^{th} *ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1977–1980, New York, NY, USA, 2011. ACM.

[188] Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. Evita: a robust event recognizer for QA systems. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, number October, pages 700–707, Vancouver, Canada, 2005.

[189] Roger C. Schank and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures (Artificial Intelligence Series)*. Psychology Press, 1 edition, July 1977.

[190] Alexander Thorsten Schutz. Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods. Master's thesis, National University of Ireland, 2008.

[191] Rushin Shah, Bo Lin, Kevin Dela Rosa, Anatole Gershman, and Robert Frederking. Improving cross-document co-reference with semi-supervised information extraction models. In *Proceedings of the Symposium on Machine Learning in Speech and Language Processing (MLSLP 2011)*, 2011.

[192] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the* 16^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–632. ACM, 2010.

[193] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *Proceedings of the* 18^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1122–1130, New York, NY, USA, 2012. ACM.

[194] Rushdi Shams and Robert E. Mercer. Improving supervised keyphrase indexer classification of keyphrases with text denoising. In Hsin-Hsi Chen and Gobinda Chowdhury, editors, *The Outreach of Digital Libraries: A Globalized Resource Network*, volume 7634 of *Lecture Notes in Computer Science*, pages 77–86. Springer Berlin Heidelberg, 2012.

[195] Ruben Sipos, Adith Swaminathan, Pannaga Shivaswamy, and Thorsten Joachims. Temporal corpus summarization using submodular word coverage. In *Proceedings of* 21^{st} *ACM International Conference on Information and Knowledge Management*, CIKM 2012, 2012.

[196] Renxu Sun, Chai-Huat Ong, and Tat-Seng Chua. Mining dependency relations for query expansion in passage retrieval. In *Proceedings of the* 29^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'06, pages 382–389, New York, NY, USA, 2006. ACM.

[197] Russell Swan and James Allan. Automatic generation of overview timelines. In *Proceedings of the* 23^{rd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 49–56, New York, NY, USA, 2000. ACM.

[198] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, January 2012.

[199] Alvin Toffler. *Future shock*. Bantam, 1970.

[200] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 33–40, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[201] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[202] R.I. Tucker and Karen Spärck Jones. Between shallow and deep: an experiment in automatic summarising. Technical Report 632, University of Cambridge, 2005.

[203] Peter D. Turney. Learning to extract keyphrases from text. national research council. *Institute for Information Technology, Technical Report ERB-1057*, 1999.

[204] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond SumBasic: Task-focused summarization and lexical expansion. *Information Processing and Management*, 43:1606–1618, 2007.

[205] Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the* 5^{th} *International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[206] Ellen M. Voorhees. Overview of TREC 2003. In *Proceedings of TREC*, volume 2003, 2003.

[207] Ellen Vorhees and David Graff. AQUAINT-2 Information-Retrieval Text Research. LDC, 2008.

[208] Christopher Walker, Stephanie Strassel, and Julie Medero. ACE 2005 Multilingual training Corpus. *Linguistic Data Consortium, Philadelphia*, 2006.

[209] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *Proceedings of the* 23^{rd} *National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 855–860. AAAI Press, 2008.

[210] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the* 45^{th} *Annual Meeting of the Association of Computational Linguistics*, pages 552–559, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[211] Jun Wang and Jianhan Zhu. Portfolio theory of information retrieval. In *Proceedings of the* 32^{nd} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 115–122, New York, NY, USA, 2009. ACM.

[212] Mengqiu Wang and Luo Si. Discriminative probabilistic models for passage based retrieval. In *SIGIR'08: Proceedings of the* 31^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 419–426, New York, NY, USA, 2008. ACM.

[213] Shuo Wang and Xin Yao. Diversity analysis on imbalanced data sets by using ensemble models. In *IEEE Symposium CIDM '09*, pages 324–331, 2009.

[214] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the* 12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA, 2006. ACM.

[215] Furu Wei, Wenjie Li, Qin Lu, and Yanxiang He. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of* 31^{st} *the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'08, pages 283–290, New York, NY, USA, 2008. ACM.

[216] Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. Kea: practical automatic keyphrase extraction. In *Proceedings of the* 4^{th} *ACM conference on Digital libraries*, DL '99, pages 254–255, New York, NY, USA, 1999. ACM.

[217] Shasha Xie and Yang Liu. Improving supervised learning for meeting summarization using sampling and regression. *Computer, Speech and Language*, 24(3):495–514, 2010.

[218] Shasha Xie and Yang Liu. Using Confusion Networks for Speech Summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL (HLT-NAACL 2010)*, pages 46–54. Association for Computational Linguistics, 2010.

[219] Shize Xu, Liang Kong, and Yan Zhang. A picture paints a thousand words: a method of generating image-text timelines. In *Proceedings of the* 21^{st} *ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2511–2514, New York, NY, USA, 2012. ACM.

[220] Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 433–443, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[221] Rui Yan, Xiaojun Wan, Mirella Lapata, Wayne Xin Zhao, Pu-Jen Cheng, and Xiaoming Li. Visualizing timelines: evolutionary summarization via iterative reinforcement between text and image streams. In *Proceedings of the* 21^{st} *ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 275–284, New York, NY, USA, 2012. ACM.

[222] Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of the* 34^{th} *International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 745–754, New York, NY, USA, 2011. ACM.

[223] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, 14(4):32–43, July 1999.

[224] Yiming Yang, Abhimanyu Lad, Ni Lao, Abhay Harpale, Bryan Kisiel, Monica Rogati, Jian Zhang, and Jaime Carbonell. Utility-based information distillation over temporally sequenced documents. In *Proceedings of the* 30^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 31–38, New York, NY, USA, 2007. ACM.

[225] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the* 22^{nd} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 42–49, New York, NY, USA, 1999. ACM.

[226] Yiming Yang, Tom Pierce, and Carbonell. A Study of Retrospective and On-line Event Detection. In *Proceedings of the* 21^{st} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 28–36, New York, NY, USA, 1998. ACM.

[227] Klaus Zechner and Alex Waibel. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *Proceedings of the* 11^{th} *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010)*, pages 186–193, 2000.

[228] Hongyuan Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence. In *Proceedings of the* 25^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 113–120, New York, NY, USA, 2002. ACM.

[229] Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, and Bo Wang. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Information Systems*, 3, 2008.

[230] Justin Jian Zhang, Ricky Ho Yin Chan, and Pascale Fung. Extractive Speech Summarization Using Shallow Rhetorical Structure Modeling. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 18(6):1147–1157, 2010.

[231] Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword extraction using support vector machine. *Advances in Web-Age Information Management*, pages 85–96, 2006.

[232] Renxian Zhang, Wenjie Li, and Qin Lu. Sentence ordering with event-enriched semantics and two-layered clustering for multi-document news summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1489–1497, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[233] Liang Zhou and Eduard Hovy. On the summarization of dynamically introduced information: Online discussions and blogs. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 237–242, 2006.

[234] Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 447–454, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[235] Liang Zhou, Miruna Ticrea, and Eduard Hovy. Multi-document biography summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 434–441, 2004.

[236] Xianshu Zhu and Tim Oates. Finding story chains in newswire articles. In *IEEE* 13^{th} *International Conference on Information Reuse and Integration (IRI 2012)*, pages 93–100, 2012.

Appendices

A Extended Key Phrase Extraction Results

Condition	NDCG	Precision
Baseline	0.7045	0.4790
Baseline + SS1	0.7099	0.4934
Baseline + SS2	0.7169	0.5076
Baseline + SS3	0.7230	0.5158
Baseline + SS4	0.7062	0.4791
Baseline + SS5	0.7112	0.5040
Baseline + SS	0.7329	0.5301
Baseline + SS + TC	0.7504	0.5180
Baseline + SS + TC + RS1	0.7589	0.5235
Baseline + SS + TC + RS2	0.7504	0.5180
Baseline + SS + TC + RS3	0.7570	0.5212
Baseline + SS + TC + RS4	0.7560	0.5223
Baseline + SS + TC + RS5	0.7582	0.5298
Baseline + SS + TC + RS6	0.7504	0.5180
Baseline + SS + TC + RS7	0.7577	0.5259
Baseline + SS + TC + RS8	0.7504	0.5180
Baseline + SS + TC + RS9	0.7579	0.5267
Baseline + SS + TC + RS10	0.7520	0.5201
Baseline + SS + TC + RS11	0.7504	0.5180
Baseline + SS + TC + RS	0.7657	**0.5430**
Baseline + SS + TC + RS + SC	0.7356	0.5140
Baseline + SS + TC + RS + SC + CN	0.7577	0.5278
Baseline + SS + TC + RS + CN + LF	0.7560	0.5170
Baseline + SS + TC + RS + SC + CN + LF	**0.7702**	0.5401

Table A.1: Extended Results of our AKE system when extracting 10 key phrases ($p-value < 0.05$)
(SS - All Shallow Semantics, SS1 - number of Characters, SS2 - number of named entities, SS3 - number of capital letters, SS4 - Part-Of-Speech tags, SS5, TC - Top Categories, RS - All Rhetorical Signals, RS1 - Continuation signals, RS2 - change of direction signals, RS3 - sequence signals, RS4 - Illustration signals, RS5 - emphasis signals, RS6 - cause/condition/result signals, RS7 - spatial signals, RS8 - comparison/contrast signals, RS9 - conclusion signals, RS10 - Fuzz signals, RS11 - non-word emphasis signals, SC - Sub-Categories from Freebase, CN - Co-reference Normalization pre-processing, LF - Light Filtering pre-processing).

B Extended Event-based Multi-document Summarization Results

Our event-based multi-document summarization methods have multiple setups. In this appendix, we show some of the results omitted in the chapters. Tables B.1 and B.3 include the results varying the hierarchical architecture, features, filtering of events, and time dilation. Then, we varied the distance metric using the best setup. All the results in Tables B.1 and B.3 use 80 key phrases. We also tried using other number of key phrases as Tables B.2 and B.4 show. Graphs B.1 and B.2 complement Tables B.2 and B.4 by providing a graphical view of data.

Table B.1: Complete results of Event-based multi-document summarization, using 80 key phrases, in the DUC 2007.

Distance	Hier.	Features	F.E.	T.D.	ROUGE-1
Euclidean	cascade	default	no	no	0.3516
Euclidean	cascade	default	no	yes	0.3606
Euclidean	cascade	default	yes	no	0.3516
Euclidean	cascade	default	yes	yes	0.3680
Euclidean	cascade	default+events	no	no	0.3488
Euclidean	cascade	default+events	no	yes	0.3642
Euclidean	cascade	default+events	yes	no	0.3527
Euclidean	cascade	default+events	yes	yes	0.3715
Euclidean	cascade	default+AWE	no	no	0.3505
Euclidean	cascade	default+AWE	no	yes	0.3669
Euclidean	cascade	default+AWE	yes	no	0.3534
Euclidean	cascade	default+AWE	yes	yes	0.3794
Euclidean	cascade	default+events+AWE	no	no	0.3520
Euclidean	cascade	default+events+AWE	no	yes	0.3613
Euclidean	cascade	default+events+AWE	yes	no	0.3531
Euclidean	cascade	default+events+AWE	yes	yes	**0.3811**
Euclidean	single-layer	default	no	no	0.3327
Euclidean	single-layer	default	no	yes	0.3516
Euclidean	single-layer	default	yes	no	0.3719
Euclidean	single-layer	default	yes	yes	0.3584
Euclidean	single-layer	default+events	no	no	0.3634
Euclidean	single-layer	default+events	no	yes	0.3611
Euclidean	single-layer	default+events	yes	no	0.3737
Euclidean	single-layer	default+events	yes	yes	0.3499
Euclidean	single-layer	default+AWE	no	no	0.3343
Euclidean	single-layer	default+AWE	no	yes	0.3602

Euclidean	single-layer	default+AWE	yes	no	0.3641
Euclidean	single-layer	default+AWE	yes	yes	0.3642
Euclidean	single-layer	default+events+AWE	no	no	0.3576
Euclidean	single-layer	default+events+AWE	no	yes	0.3482
Euclidean	single-layer	default+events+AWE	yes	no	0.3559
Euclidean	single-layer	default+events+AWE	yes	yes	0.3522
Chebyshev	cascade	default+events+AWE	yes	yes	0.3466
Cosine	cascade	default+events+AWE	yes	yes	**0.3826**
Frac01	cascade	default+events+AWE	yes	yes	0.3018
Frac05	cascade	default+events+AWE	yes	yes	0.3284
Frac075	cascade	default+events+AWE	yes	yes	0.3451
Frac133	cascade	default+events+AWE	yes	yes	0.390
JSD	cascade	default+events+AWE	yes	yes	0.3519
Manhattan	cascade	default+events+AWE	yes	yes	0.3763
Minkowski	cascade	default+events+AWE	yes	yes	0.3362

Table B.2: ROUGE-1 scores for our Event-based multi-document summarization, in the DUC 2007, using the best configuration, but varying the number of key phrases.

Nr. Key Phrases	ROUGE-1	Nr. Key Phrases	ROUGE-1
10	0.3708	80	0.3811
20	0.3686	90	0.3670
30	0.3674	100	0.3743
40	0.3735	150	0.3612
50	0.3771	200	0.3692
60	0.3709	250	0.3574
70	0.3732	300	0.3552

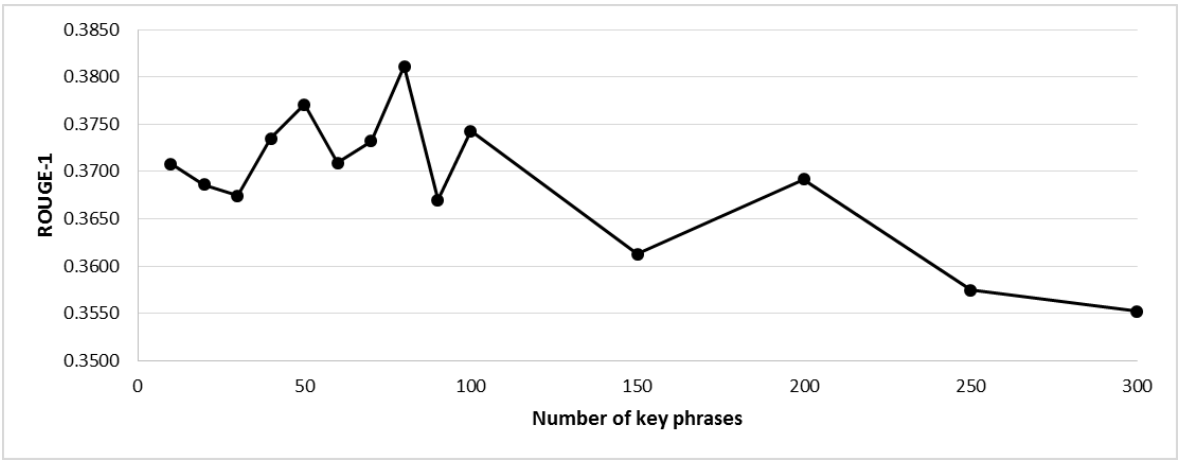


Figure B.1: Graphical visualization of Table B.2.

Table B.3: Complete results of Event-based multi-document summarization, using 80 key phrases, in the TAC 2009.

Distance	Hier.	Features	F.E.	T.D.	ROUGE-1
Euclidean	cascade	default	no	no	0.5264
Euclidean	cascade	default	no	yes	0.5037
Euclidean	cascade	default	yes	no	0.5274
Euclidean	cascade	default	yes	yes	0.5099
Euclidean	cascade	default+events	no	no	0.5345
Euclidean	cascade	default+events	no	yes	0.5114
Euclidean	cascade	default+events	yes	no	0.5413
Euclidean	cascade	default+events	yes	yes	0.5219
Euclidean	cascade	default+AWE	no	no	0.5181
Euclidean	cascade	default+AWE	no	yes	0.5058
Euclidean	cascade	default+AWE	yes	no	0.5101
Euclidean	cascade	default+AWE	yes	yes	0.5190
Euclidean	cascade	default+events+AWE	no	no	0.5153
Euclidean	cascade	default+events+AWE	no	yes	0.5105
Euclidean	cascade	default+events+AWE	yes	no	0.5163
Euclidean	cascade	default+events+AWE	yes	yes	0.5009
Euclidean	single-layer	default	no	no	0.5201
Euclidean	single-layer	default	no	yes	0.5251
Euclidean	single-layer	default	yes	no	0.5229
Euclidean	single-layer	default	yes	yes	0.5151
Euclidean	single-layer	default+events	no	no	0.5125
Euclidean	single-layer	default+events	no	yes	0.5333
Euclidean	single-layer	default+events	yes	no	0.5284
Euclidean	single-layer	default+events	yes	yes	0.5335
Euclidean	single-layer	default+AWE	no	no	0.5224
Euclidean	single-layer	default+AWE	no	yes	0.5401
Euclidean	single-layer	default+AWE	yes	no	0.5381
Euclidean	single-layer	default+AWE	yes	yes	0.5261
Euclidean	single-layer	default+events+AWE	no	no	0.5075
Euclidean	single-layer	default+events+AWE	no	yes	**0.5501**
Euclidean	single-layer	default+events+AWE	yes	no	0.5298
Euclidean	single-layer	default+events+AWE	yes	yes	0.5231
Chebyshev	single-layer	default+events+AWE	yes	yes	0.4746
Cosine	single-layer	default+events+AWE	yes	yes	0.5038
Frac01	single-layer	default+events+AWE	yes	yes	0.4210
Frac05	single-layer	default+events+AWE	yes	yes	0.4449
Frac075	single-layer	default+events+AWE	yes	yes	0.4572
Frac133	single-layer	default+events+AWE	yes	yes	0.5085
JSD	single-layer	default+events+AWE	yes	yes	0.5019
Manhattan	single-layer	default+events+AWE	yes	yes	0.4981
Minkowski	single-layer	default+events+AWE	yes	yes	0.4728

Table B.4: Results of Event-based multi-document summarization, in the TAC 2009, using the best configuration but varying the number of key phrases.

Nr. Key Phrases	ROUGE-1	Nr. Key Phrases	ROUGE-1
10	0.5288	80	0.5501
20	0.5095	90	0.5501
30	0.5255	100	0.5513
40	0.5373	150	0.5504
50	0.5403	200	0.5510
60	0.5454	250	0.5461
70	0.5570	300	0.5419

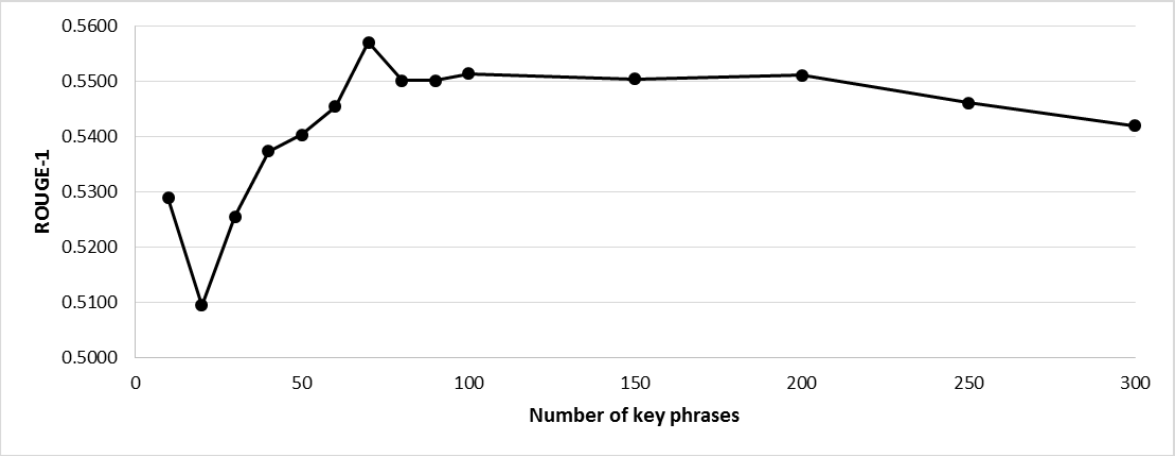


Figure B.2: Graphical visualization of Table B.4.

C Extended example of an Event-based Multi-document Summary

In this Appendix, we show an extended example of the event-based multi-document summary of Topic D0712C from the DUC 2007 using the best summarizer configuration (waterfall, all features, event filtering, and time dilation). We included the initial summaries of the first three documents, two intermediate summaries, and the final summary.

Table C.1: Extended example of Event-based multi-document summarization of Topic D0712C from the DUC 2007.

Document 1

TEHRAN, March 7 (Xinhua) – Iran declared here today that it will take the case of insulting religious sanctities like the Rushdie affair to the United Nations.
Mostafa Mirsalim, Minister of Culture and Islamic Guidance, said at a press conference here that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as Rushdie affairs.
"We think there is room for such an action at different international forums including the U.N.," he said.
Speaking through an interpreter to an international corps of reporters here to report on March 8th parliamentary elections, he stressed that the concept of freedom in the West led to violations of the rights of others.
While answering the question by Xinhua, Mirsalim said "Iran is planning to take the issue of insulting religious sanctities, which has resulted in the Rushdie issue, to the international bodies, including to the U.N.."
"We believe that there is a deviation inherent in the concept of freedom in some of the Western societies that tramples the rights of the most people in the world," he said.
"If the cultural community and the authors in the West had been committed to respecting the religious beliefs of the people, a problem such as Salman Rushdie would have never emerged," he added.
He added that his ministry was committed to defend Islamic culture and rejected allegations by critics that his ministry took sides with the religious conservatives.
"Being neutral does not mean anything. We have to defend the Islamic culture," he said.
However, the Iranian culture minister did not say whether Iran intends to drop late religious leader Ayatollah Khomeini's fatwa against Salman Rushdie.
Iranian Foreign Minister Ali Akbar Velayati said during his recent trip to central Asian states that Iran will resume talks with the European Union (EU) on the Rushdie issue.
On 14 February 1989, late Iranian leader Ayatollah Khomeini issued a religious edict, pronouncing a sentence of death on British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel The Satanic Verses, and exhorting all Moslems to carry out the sentence.
Liberal weekly Bahman reported in February that former deputy guidance minister, Ahmad Masjed Jame'ie, resigned over differences pertaining the new restrictions imposed on authors.
The Ministry of Culture and Islamic Guidance came under heavy attacks in 1995 by the religious conservatives that took the ministry for allowing the publication of romantic novels.
In a sudden about-face in its publication policy, the ministry declared that authors have to receive permission prior to publication of their books.

Summary of Document 1 (initial summary)

<Empty - The event-based detector didn't detect any event type, and filtered all sentences.>

Document 2

TEHRAN, March 11 (Xinhua) – Iran announced here today that it drops the death sentence on Salman Rushdie, a British writer accused of defaming Islam in his novel, under tremendous pressure from European countries.
Iranian President Akbar Hashemi Rafsanjani said at a press conference this morning that "the death sentence was never meant to be carried out."
Rafsanjani attacked Europe, especially Britain, for their lack of resolve to solve the issue.
He said: "The Europeans, especially the British, made hue and cry over it. Even now they do not allow it to be solved."
"The fatwa issued by Imam Khomeini can be found in all the books on Islamic law for the past 1,000 years. He only said something which existed in these books," Rafsanjani added.
"We really do not think it is prudent that we would dwell on this issue so much," he added.
Observers here said that the indirect Tehran-EU talks on Salman Rushdie is producing some understanding. The two sides have agreed that the case should be regarded merely as a religious and theoretical issue.
On February 14, 1989, late Iranian leader Ayatollah Khomeini issued a religious edict, pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel "The Satanic Verses," which was believed by Moslems as defaming Islam, and exhorting all Moslems to carry out the sentence.
In 1993, Rafsanjani told prominent Egyptian journalist Muhammad Hassanein Haykal that there was a difference between a fatwa (ruling) and hokm (command). He said that Khomeini did not mean the death sentence to be a command.
An observer here said that with U.S. pressure mounting on Iran, Iranian authorities have decided to solve the Rushdie issue to gain cooperation of the European Union.
"The Iranians no longer want any of the troubles they went through over the Rushdie issue," one observer said.
"The question is not whether but how to withdraw from previous position, in such a way that Iran will not lose face both at home and abroad," he said.
Iranian Foreign Minister Ali Akbar Velayati reiterated here last week that Iran would not send anyone to kill the author.
Mostafa Mirsalim, Minister of Culture and Islamic Guidance, said at a press conference held last week that Iran will take the case of insulting religious sanctities like the Rushdie affair to the United Nations.
He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair.

Summary of Document 2 (initial summary)

TEHRAN, March 11 (Xinhua) – Iran announced here today that it drops the death sentence on Salman Rushdie, a British writer accused of defaming Islam in his novel , under tremendous pressure from European countries .
Mostafa Mirsalim , Minister of Culture and Islamic Guidance , said at a press conference held last week that Iran will take the case of insulting religious sanctities like the Rushdie affair to the United Nations .
He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair .
Iranian President Akbar Hashemi Rafsanjani said at a press conference this morning that " the death sentence was never meant to be carried out .
On February 14 , 1989 , late Iranian leader Ayatollah Khomeini issued a religious edict , pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel " The Satanic Verses , " which was believed by Moslems as defaming Islam , and exhorting all Moslems to carry out the sentence .
He said that Khomeini did not mean the death sentence to be a command .
" The question is not whether but how to withdraw from previous position , in such a way that Iran will not lose face both at home and abroad , " he said.
Iranian Foreign Minister Ali Akbar Velayati reiterated here last week that Iran would not send anyone to kill the author . " The fatwa issued by Imam Khomeini can be found in all the books on Islamic law for the past 1,000 years.

Document 3

TEHRAN, March 13 (Xinhua) – Iran's Supreme Leader Ali Khamenei is the only person who has the authority to cancel the death sentence on Salman Rushdie, a high-ranking official at presidential office said here today.
"The decree [death sentence] has not been cancelled. No one, except for Ayatollah Khamenei, can revise the decree," the official at presidential office told Xinhua today.
The official at presidential office who spoke on condition of anonymity said that since the decree was issued by a religious scholar of the high stature of Khomeini it could not be changed by officials.
"Only Ayatollah Khamenei, as the Supreme Jurisprudent, can revise decrees issued by the late Imam," he stressed.
On February 14, 1989, the late Iranian leader Ayatollah Khomeini issued a religious edict, pronouncing a sentence of death on the British author Salman Rushdie and his publishers in protest at the publication of Rushdie's novel "The Satanic Verses", and exhorting all Moslems to carry out the sentence.
Since then, the Rushdie issue has turned into a big controversial problem that hinders the relations between Iran and European countries.
However, Iranian authorities have given into tremendous pressures from the European Union after the death of Khomeini, promising not to dispatch any commandos to kill Rushdie.
Foreign Minister Ali Akbar Velayati told a press conference here last week that Iran would not send anyone to kill Rushdie. Similar assurances have been given to European Union by President Rafsanjani and Parliament Speaker Ali Akbar Nateq-Nouri.
President Akbar Hashemi Rafsanjani said at a press conference here Monday that the sentence on Salman Rushdie was an issue of Islamic law and that "it was not essentially an Executive issue."
The official at presidential office said that Rafsanjani wanted to draw the attention to the religious importance of the decree.
"The president wanted to make a distinction between a religious, jurisprudential decree and an executive, administrative one," the official said.
"The British with their hue and cry wanted to make it sound that the decree was an executive one," he added.
"In fact it was the British themselves who did not want to have this issue resolved and even now their actions show they do not seek an end to it," he said.
Rafsanjani's Monday statement on Rushdie has drawn criticism from a national newspaper Jomhuri Eslami.
The paper issued an editorial Tuesday, saying that Rushdie still has to face death sentence because "no one has the right to change the divine command."
"It is the responsibility of all Moslems to kill anyone who would insult the Blessed Progeny of Prophet Muhammad, the Holy Koran and Islamic sanctities," the daily said.
"This punishment has to be done regardless of time, place and individuals involved. The decree is not a governmental or personal decree and hence cannot be changed to adjust to new political conditions," it warned.
Observers here said that the indirect talks on Salman Rushdie between EU and Tehran is producing some understanding. The two sides have agreed the case be regarded merely as a religious and theoretical issue.
They said that with U.S. pressures mounting on Iran, Iranian authorities have decided to solve the Rushdie issue to gain the cooperation of the European Union.
"The question is not whether, but how, to withdraw from previous position, in such a way that Iran will not lose face both at home and abroad," a political analyst here said.
Iranian authorities also face tremendous pressure and opposition from the religious extremists inside the country on the issue of Rushdie.
Iranian officials have stated recently that they want to adopt a protective stance with regard to issues such as Rushdie affair.
Mostafa Mirsalim, Minister of Culture and Islamic Guidance, said at a press conference held last week that Iran will take the case of insulting religious sanctities like the Rushdie affair to the United Nations.
He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as Rushdie affairs.

Summary of Document 3 (initial summary)

On February 14 , 1989 , the late Iranian leader Ayatollah Khomeini issued a religious edict , pronouncing a sentence of death on the British author Salman Rushdie and his publishers in protest at the publication of Rushdie's novel " The Satanic Verses " , and exhorting all Moslems to carry out the sentence .
Mostafa Mirsalim, Minister of Culture and Islamic Guidance , said at a press conference held last week that Iran will take the case of insulting religious sanctities like the Rushdie affair to the United Nations .
TEHRAN , March 13 (Xinhua) – Iran's Supreme Leader Ali Khamenei is the only person who has the authority to cancel the death sentence on Salman Rushdie , a high-ranking official at presidential office said here today .
The official at presidential office who spoke on condition of anonymity said that since the decree was issued by a religious scholar of the high stature of Khomeini it could not be changed by officials .
President Akbar Hashemi Rafsanjani said at a press conference here Monday that the sentence on Salman Rushdie was an issue of Islamic law and that " it was not essentially an Executive issue .
He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as Rushdie affairs .
However , Iranian authorities have given into tremendous pressures from the European Union after the death of Khomeini , promising not to dispatch any commandos to kill Rushdie .
" The British with their hue and cry wanted to make it sound that the decree was an executive one , " he added .
" In fact it was the British themselves who did not want to have this issue resolved and even now their actions show they do not seek an end to it , " he said .

Summary of Document 1 and 2 (intermediate summary)

<Is the same of Summary of Document 2 (initial summary) because the Summary of Document 1 was empty>

Summary of Document 1,2, and 3 (intermediate summary)

He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair .
On February 14 , 1989 , late Iranian leader Ayatollah Khomeini issued a religious edict , pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel " The Satanic Verses , " which was believed by Moslems as defaming Islam , and exhorting all Moslems to carry out the sentence .
TEHRAN, March 11 (Xinhua) – Iran announced here today that it drops the death sentence on Salman Rushdie , a British writer accused of defaming Islam in his novel , under tremendous pressure from European countries .
Mostafa Mirsalim , Minister of Culture and Islamic Guidance , said at a press conference held last week that Iran will take the case of insulting religious sanctities like the Rushdie affair to the United Nations .
He said that Khomeini did not mean the death sentence to be a command .
Iranian President Akbar Hashemi Rafsanjani said at a press conference this morning that " the death sentence was never meant to be carried out .
" The question is not whether but how to withdraw from previous position, in such a way that Iran will not lose face both at home and abroad , " he said .
Iranian Foreign Minister Ali Akbar Velayati reiterated here last week that Iran would not send anyone to kill the author .
" The fatwa issued by Imam Khomeini can be found in all the books on Islamic law for the past 1,000 years .

Summary of Document 1,2, ..., 25 (Final Summary)

Iranian Foreign Minister Kamal Kharrazi , who made the announcement in New York, and his British counterpart , Robin Cook , had portrayed the move as a way to improve ties that have remained strained over the issue and agreed to exchange ambassadors .
LONDON _ The British government said Wednesday that it would continue to press Iran to lift the death sentence against the author Salman Rushdie when its foreign secretary , Robin Cook , meets the Iranian foreign minister in New York on Thursday .
VIENNA, Austria (AP) – The European Union on Monday welcomed a move by the Iranian government to distance itself from an Islamic edict calling for British author Salman Rushdie's death even as two senior Iranian clerics said the ruling was irrevocable .
The move follows the Iranian government's distancing itself last month from bounties offered for the death of Rushdie and a strong reaction by hard-liners who support the killing of the Booker Prize-winning author .
He said that Iran will ask the United Nations to effectively put a ban on insulting religious sanctities in a bid to prevent disputes such as the Rushdie affair .
On February 14 , 1989 , late Iranian leader Ayatollah Khomeini issued a religious edict , pronouncing a death sentence on the Indian-born British author Salman Rushdie and his publishers in protest against the publication of Rushdie's novel " The Satanic Verses ", which was believed by Moslems as defaming Islam , and exhorting all Moslems to carry out the sentence .