Towards
Automated Related Work Summarization

by

© **HOANG Cong Duy Vu**

(BSc.(Hons), HCMUS-VNU, Vietnam)

A thesis submitted to the
School of Computing
in partial fulfilment of the
requirements for the degree of
Master of Science

Department of Computer Science
NATIONAL UNIVERSITY OF SINGAPORE

December 2010

Acknowledgments

I would like to show my gratitude to my advisor, Professor Min-Yen Kan, whose encouragement, guidance, and support from the beginning to the final level helped me develop an understanding of the research subject. This thesis would not have been possible without his help.

I owe my deepest gratitude to my parents, sisters, and brothers who always help me even in the most difficult circumstances.

Lastly, I would like to thank my friends in the Web Information Retrieval & Natural Language Processing Group (WING) who supported me in a number of ways with invaluable and insightful comments during the completion of my thesis. Also special thanks to the School of Computing (NUS) for assisting me during my study here.

HOANG Cong Duy Vu (December 2010)

Abstract

Towards Automated Related Work Summarization

HOANG Cong Duy Vu

"*This thesis introduces and describes the novel problem of automated related work summarization. Given multiple articles (e.g., conference or journal papers) as input, and a set of keywords that describes a target paper's topics of interest in a hierarchical fashion, a related work summarization system creates a topic-biased summary of related work specific to the target paper. This thesis has two main contributions. First, I conducted a deep manual analysis on various aspects of related work sections to identify their important characteristics in locating appropriate information for summarization and generation processes. Second, based on the observations from my manual analysis, I have developed my initial prototype* ***Re****lated* ***Wo****rk* ***S****ummarization system, namely* ***ReWoS****, which creates its extractive summaries using two different strategies for locating appropriate sentences for general topics as well as detailed ones. The proposed* ***ReWoS*** *system significantly outperforms baseline systems in terms of human evaluation measures designed specific to the task.*"

Preface

All work presented in this thesis is the original work of the author. A part of this thesis has been published in the following conference paper:

Cong Duy Vu Hoang, Min-Yen Kan. "Towards Automated Related Work Summarization". In the 23rd International Conference on Computational Linguistics (COLING'10), August 23-27, 2010, Beijing, China, pp. 427-435. (acceptance rate: 22%).

To my parents!

To my brothers and sisters!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In scientific research, scholars spend a significant amount of time determining which articles are relevant to their specific tasks. Getting up to speed on the comparative advantages and disadvantages of related work is crucial in positioning a scholar's current work for publication. The growing number of scholarly publications hampers this, as the ambiguity and diversity in expressing relevant techniques, datasets and tools is only limited by the authors' use of natural language.

In many fields, a scholar needs to show an understanding of the context of his problem and relate his work to prior community knowledge. A related work section is often the vehicle for this purpose; it contextualizes the scholar's contribution and helps the reader understand the critical aspects of the previous works that the current work addresses. Creating such a related work summary requires the scholar to understand the nuances of his own work, and to manipulate the contextual research to support the advantages of his method.

Imagine a scenario where scholars use a search engine to update or seek for certain research topics of interest. In this scenario, the search engine may return a long list of

results in different formats such as HTML web pages, PDF, MS Document as well as text files. The scholar then needs to check all the links one by one, to identify which are truly relevant. In such a situation, a natural question arises: "Is there any technique to generate a unified, thorough overview of these related results?".

Let me paint another scenario. Current research is increasingly cross-disciplinary. For example, a scholar in Natural Language Processing (NLP) is working on a research problem related to an another discipline, perhaps biology. Such research is also termed natural language processing in biology or bioinformatics[1]. A scholar new to this domain may not have the appropriate background knowledge in biology and needs to rapidly learn about this unfamiliar research domain without wasting a lot of time. Such a requirement can only be satisfied with the development of effective tools to help a scholar cover the necessary background as quickly as possible.

Currently, to my best knowledge, there are no existing tools that have such capabilities. To build such automatic intelligent systems is difficult, requiring the combination of different techniques in information retrieval (IR) and NLP. To partially address this difficulty, individuals and organizations have put efforts into building smart scholarly repositories that can limit the search scope, given (semi-)manually provided filtering criteria. Exemplars built for the domain of computer science include DBLP - the Computer Science Bibliography[2], CiteSeer - the Scientific Literature Digital Library[3], Google Scholar - a service by Google for scholarly literature search[4], and ArnetMiner - the online Academic Researcher Social Network Search built by Tsinghua University[5]. More specifically, there are also some systems built for specific domains such as: BioInformatics (e.g. PubMed[6]) or Computational Linguistics (e.g. ACL Anthology[7]), AAN

[1] http://en.wikipedia.org/wiki/Bioinformatics
[2] http://dblp.uni-trier.de/
[3] http://citeseer.ist.psu.edu/
[4] http://scholar.google.com
[5] http://www.arnetminer.org/
[6] http://www.ncbi.nlm.nih.gov/pubmed/
[7] http://www.aclweb.org/anthology-new/

- the ACL Anthology Network hosted by University of Michigan (2008)[8]). Such systems provide supporting tools such as advanced search by authors or topic keywords (e.g. DBLP, CiteSeer, Google Scholar, ACL Anthology), visualization and statistics (e.g. ArnetMiner, AAN) to facilitate the scholars' search requests.

Even though such repositories can perform limited-scope search, the problem of information overload still remains. For instance, using three systems (DBLP, AAN, CiteSeer), a keyword search for "multi-document summarization" retrieves over 200 hits – 87 from DBLP, 29 from AAN, and 127 from CiteSeer. To read through all of such retrieved results is still non-trivial and time-consuming. Moreover, scholars need to cover all the retrieved results to ensure comprehensive working knowledge of the relevant previous work. Thus, a demand for summarization of scientific articles is very necessary and important to accelerate and optimize the working hours for scholars.

I now envision an NLP application that assists the scholar in creating his related work summary. I propose *related work summarization* as a challenge to the automatic summarization community. In the full challenge, it is a topic-biased, multi-document summarization problem that takes as input a target scientific document for which a related work section needs to be generated. The output goal is to create a related work section that finds the relevant related works and contextually describes them in relationship to the scientific document at hand.

I dissect the full challenge as bringing together work of disparate interests; 1) in finding relevant documents; 2) in identifying the salient aspects of a relevant document worth mentioning in relation to the current work; and 3) generating the topic-biased final summary. While it's clear that current NLP technology does not let us build a complete solution for this task, I believe tackling the component problems will help bring us towards an eventual solution.

Also, unlike other summarization scenarios, a source of gold standard summaries

[8] `http://clair.si.umich.edu/clair/anthology/index.cgi`

is available in publications that feature an explicitly demarcated summary of the related literature. This makes the evaluation of such systems plausible and comparable. For example, a solution to the first citation prediction component task may use the actual identity of the cited papers for evaluation. In the final component of related work summarization task, I can use the gold standard summaries for comparison.

In fact, existing work in the NLP and recommendation systems communities have already begun work that fits towards the completion of the first two tasks. Citation prediction (Nallapati et al., 2008) is a growing research area that has aimed both at predicting citation growth over time within a community and at individual paper citation patterns. Also, automatic survey generation (Mohammad et al., 2009) is becoming a growing field within the summarization community.

However, to date, I have not yet seen any work that examines topic-biased summarization of multiple scientific articles. For these reasons, I work towards the final component in the current work – *the creation of a related work section, given a structured input of an appropriate topic for summary.*

The key contributions of my thesis consists of work towards this goal:

1. I conduct a study of the argumentative patterns used in related work sections, to describe the plausible summarization tactics for their creation in Chapter 2.

2. In Chapter 4, I describe in detail my approach to generate an extractive related work summary, given an input topic hierarchy tree. This approach uses two separate summarization processes to differentiate between summarizing shallow internal nodes from deep detailed leaf nodes of the topic tree.

1.2 Research Goals

Inspired from the situations described as the above, I propose the following novel research problem: to automatically generate a scientific summary, given multiple articles

(e.g. conference or journal papers) as input, and a set of keywords that describe the topics of interest presented in a hierarchical fashion. This query-biased summarization process is targeted at generating a related work section of a paper, and not a generic summary as would be the case in a survey paper. Such a related work summary is a text summary which describes briefly the main ideas of previous or recent works, particularly indicating important aspects in relationship to the current paper where the section is to be embedded. More importantly, a related work summary should clearly describe the similarities and differences among articles.

1.3 Overview of Thesis

The organization of this thesis is as follows:

In Chapter 2, I will discuss my manual analysis characterizing actual related work summaries. This analysis will help recognize the challenges when dealing with related work summarization.

Chapter 3 will give a literature review on previous works relevant to the proposed problem.

Chapter 4 firstly justifies the formulation of my proposed research problem, and then describes the proposed system that will implement the idea using two separate strategies for general topics and detailed topics, given a topic hierarchy tree. This idea is inspired from a rhetorical analysis on human-written related work summaries.

In Chapter 5, I will evaluate the proposed system against two baselines, using both objective automatic and subjective human evaluation methods.

Chapter 6 discusses future work and Chapter 7 concludes this thesis.

Chapter 2

Manual Analysis

In the first part of this chapter, I will discuss the construction of a new related work summarization dataset, namely **RWSData** (Data for Related Work Summaries) used for the analysis and evaluation in this thesis. I then deconstruct actual related work summaries from articles in **RWSData** to gain insight on how they are structured and authored, from both rhetorical and content levels as well as on the surface lexical levels. Based on this manual analysis, I identify key problems in composing a solution to related work summarization. I discuss these issues, namely – the topical structure of related work summaries, the decomposition and alignment problems, related work representation in the output summaries, and the evaluation metrics designed specific for evaluation – in second part of this chapter.

2.1 Data Construction

2.1.1 Annotation

The first challenge I encountered was the lack of a suitable dataset, designed specific to the evaluation process. Thus, I needed to manually construct such a dataset for my use. As the data preparation was very costly in terms of time, my aim in this goal was not only

to create a dataset for my own use, but also to further provide this dataset to assist other researchers in related work summarization and to allow them to verify my experimental results.

Most scientific articles contain a section presenting related works, often titled "Related Work", "Background", "Literature Review", "Previous Studies", "Prior Work". This observation led me to utilize such related work sections as gold standard related work summaries to aim to generate.

No.	Article ID	Conference
1	C08-1013	COLING
2	C08-1031	COLING
3	C08-1064	COLING
4	C08-1066	COLING
5	E09-1018	EMNLP
6	N09-1008	NAACL
7	N09-1019	NAACL
8	N09-1027	NAACL
9	N09-1034	NAACL
10	N09-1042	NAACL
11	P07-1034	ACL
12	P08-1001	ACL
13	P08-1006	ACL
14	P08-1027	ACL
15	P08-1032	ACL
16	P08-1052	ACL
17	p27-kalashnikov	SIGIR
18	p79-raghavan	SIGIR
19	p203-wu	SIGIR
20	p343-ko	SIGIR

Table 2.1: A list of 20 selected articles in the RWSData dataset and their associated conferences.

To construct the **RWSData**, I carefully selected twenty articles from well-respected venues in NLP and IR, namely SIGIR, ACL, NAACL, EMNLP and COLING. The de-

tails of these articles are shown in Table 2.1. I then painstakingly extracted the related work summaries directly from the PDF files by using manual copy-and-paste operations to ensure the cleanliness of the resultant text. References within each related work were identified, located and their text extracted in the same manner. Only references to books or Ph.D. theses were removed from these reference lists, as summarizing very long documents may cause problems as mentioned in (Mihalcea and Ceylan, 2007). The remaining references were conference/journal articles or technical reports. As a result, all the related work sections together with the references within them were then turned into the pre-processing steps.

Article ID	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
C08-1013	26	512	228	10	2194	219	47790	4779	4572	3	2
C08-1031	19	437	201	16	3347	209	68337	4271	5217	3	2
C08-1064	20	438	217	8	2108	263	48727	6090	4149	5	2
C08-1066	14	408	231	8	929	116	21734	2716	2679	3	2
E09-1018	25	837	359	8	1646	205	36539	4567	3851	3	2
N09-1008	10	296	176	2	348	174	8580	4290	1564	1	1
N09-1019	16	540	282	13	3370	259	71580	5506	6076	5	2
N09-1027	9	264	159	6	1039	173	22255	3709	2895	2	1
N09-1034	13	471	195	12	2107	175	42906	3575	4383	4	2
N09-1042	15	361	184	12	2470	205	56728	4727	4953	2	1
P07-1034	13	327	144	5	1035	207	22745	4549	2747	1	1
P08-1001	9	472	225	9	1461	162	30899	3433	3919	1	1
P08-1006	6	179	106	9	1862	206	45264	5029	4376	3	2
P08-1027	40	866	352	26	4400	169	94172	3622	6464	6	2
P08-1032	21	492	257	7	2287	326	45139	6448	4289	3	2
P08-1052	24	793	349	18	4422	245	91679	5093	6027	4	2
p27-kalashnikov	26	818	324	20	5549	277	112267	5613	6223	5	2
p79-raghavan	20	604	267	9	2978	330	71683	7964	5528	3	2
p203-wu	18	922	352	9	2017	224	51009	5667	4731	7	3
p343-ko	14	411	203	11	2151	195	44758	4068	4287	1	1

Table 2.2: **Detailed statistics of the RWSData dataset.** Legend: N1-4) No. of {sentences, words, distinct words, cited articles} in the related work section, N5-9) {total no. of sentences, average no. of sentences, total no. of words, average no. of words, total no. of distinct words} in the referenced articles, and N10-11) {no. of nodes, height} of the topic tree.

The pre-processing steps were as follows. First, an OCR package, OmniPage[1], was used to extract the raw text from the corresponding PDF files. OmniPage was nec-

[1] `http://www.nuance.com/imaging/products/omnipage.asp`

essary to extract the text with a very high accuracy. Next, a sentence segmentation tool[2] was used to segment raw text files into individual sentences.

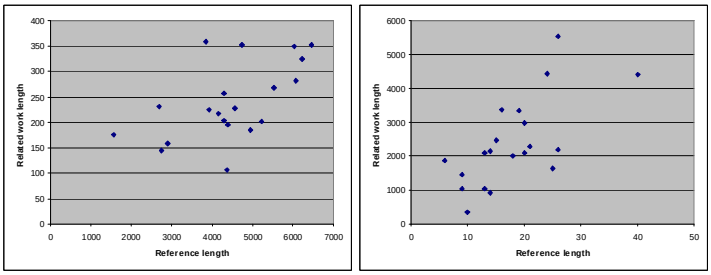
The OmniPage output required post-processing as the tool extracts all possible texts from the PDF, including non-body text, such as those associated with figures, tables or mathematical symbols/formulas. In a first pass, this caused problems where text are partially lost or uncorrected segmented. Subsequently, I solved this problem by manually correcting the text by: 1) proofreading extracted raw texts sentence by sentence, 2) identifying sentences including errors mentioned above and 3) removing them. This step was overly time-consuming, taking almost a month. Finally, tokenization and lowercase steps were performed.

2.1.2 Data Statistics

The detailed statistics of the **RWSData** dataset is shown in Table 2.2. This dataset includes 20 articles with one related work section for each article. Based on this table, the correlation between the word- and sentence- based length of related work summaries and the original referring articles (ORAs) is shown in Figure 2.1. The word-based length of related work summaries and ORAs is in range of $100-350$ and $1500-6500$ distinct words, respectively, referring to a word-based compression rate of approximately $0.05-0.07\%$. Meanwhile, their sentence-based length is in range of $6-40$ and $348-5549$, respectively, referring to a sentence-based compression rate of approximately $0.01-0.02\%$. As such, both word- and sentence- based compression rate are less than 1%. This is a key challenge in related work summarization, since the compression length rate is very high (less than 1%).

RWSData summaries also average 17.9 sentences, 522 words in length, citing an average of 10.9 articles. As such, the task of related work summarization needs to take multiple articles in the input. If the input has many articles, overlapping and novel infor-

[2] `http://l2r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS`



(a) Word-based length (b) Sentence-based length

Figure 2.1: Word- (left) and sentence- (right) based correlation between reference text length and related work section length, over the 20 articles in the **RWSData** dataset.

mation among articles will increase. This adds further difficulties for the summarization task in handling multiple input but also lends the opportunity to utilize more evidence to base our summarization processes on.

Details on the demographics of RWSData are shown in Table 2.3. The **RWSData** dataset is currently publicly available for research purposes[3].

Measure	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**
average	17.9	522.4	240.6	10.9	2386.0	217.0	51739.6	4785.8	4446.5	3.3	1.8
stdev	7.9	216.5	75.3	5.6	1306.7	53.9	26682.3	1212.3	1297.9	1.7	0.6
min	6	179	106	2	348	116	8580	2716	1564	1	1
max	40	922	359	26	5549	330	112267	7964	6464	7	3

Table 2.3: Statistics with average, stdev (STandard DEViation), min (MINimum), and max (MAXimum) of values of N1−N11 denoted in Table 2.2 in the **RWSData** dataset.

[3] `http://wwww.comp.nus.edu.sg/~hcdvu/RWSData/RWSData.htm`

2.2 Characteristics of Related Work Summaries

2.2.1 Definition

A related work (abbreviated to **RW**) summary is a text summary which describes briefly the main ideas of previous or recent works, indicating their relevant aspects in the context of the current paper's topics. Specifically, a RW summary should clearly identify the similarities and dissimilarities among articles, as well as discuss the previous works in an appropriate manner. Figure 2.2 gives a prototypical example of a RW summary.

Little work has been done on contradiction detection. The PASCAL Recognizing Textual Entailment (RTE) Challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007) focused on textual inference in any domain. Condoravdi et al. (2003) first recognized the importance of handling entailment and contradiction for text understanding, but they rely on a strict logical definition of these phenomena and do not report empirical results. To our knowledge, Harabagiu et al. (2006) provide the first empirical results for contradiction detection, but they focus on specific kinds of contradiction: those featuring negation and those formed by paraphrases. They constructed two corpora for evaluating their system. One was created by overtly negating each entailment in the RTE2 data, producing a balanced dataset (LCC_negation). To avoid overtraining, negative markers were also added to each non-entailment, ensuring that they did not create contradictions. The other was produced by paraphrasing the hypothesis sentences from LCC_negation, removing the negation (LCC_paraphrase): *A hunger strike was not attempted* → *A hunger strike was called off*. They achieved very good performance: accuracies of 75.63% on LCC_negation and 62.55% on LCC_paraphrase. Yet, contradictions are not limited to these constructions; to be practically useful, any system must provide broader coverage.

Figure 2.2: An actual example RW summary from a published conference paper (de Marneffe et al., 2008).

2.2.2 Position

In scientific writing, a RW summary (often occurring as an independent section) can be placed at two different positions depending the purpose of authors. At the position either within the introduction section or the section on its own at the beginning of the article immediately after the Introduction section, a RW summary should be give sufficient descriptions as well as possible stance about previous works. Meanwhile, at the position right before the Conclusion section, it should give a relatively short outline of previous

studies and adequate comparisons between the technical content of the paper and previous studies. A RW summary positioned at the end of the article may be more complicated to create automatically as it needs extensive semantic processing, which are beyond the current ability of NLP techniques, for example generating comparisons between current proposed method and previous methods. Thus, in this study, I target on generating RW summaries which target to be placed at the first, beginning position.

2.2.3 Topical Structure

I conducted a first preliminary analysis on human-written structures of existing RW summaries within the **RWSData** dataset. I carried out my analysis by reading all RW sections and then exploring the discourse strategies how RW summaries can be written. From my analysis, I propose a general structure for RW summaries, which I show in Figure 2.3.

The structure of a RW section follows a topic hierarchy tree in which the root node is the general topic of the RW summary. The content of the general topic usually starts with a topic sentence following by the general background or description on that topic. This content is optional and can be ignored depending on the authors' purposes. Further, this general topic may have a number of topics, each of which has the structure comprising of different sections: Background, Problem Description, Result, Comment, and Claim. Each of such a topic may have sub-topics which recursively use the same structure.

In addition, the optional section describing the individual proposed statement of authors should be included. Importantly, according to my understanding, the contents inside the dark rounded rectangle boxes are capable of being generated automatically. In contrast, those inside the dashed rounded rectangle boxes seem to be very difficult to generate. Figure 2.4 gives an example that narrates the structure of the RW summary.

In Figure 2.4, the topic hierarchy tree is comprised of the root node with the general topic "text classification" (lines 1–5) followed by the topic 1 "monolingual classifi-

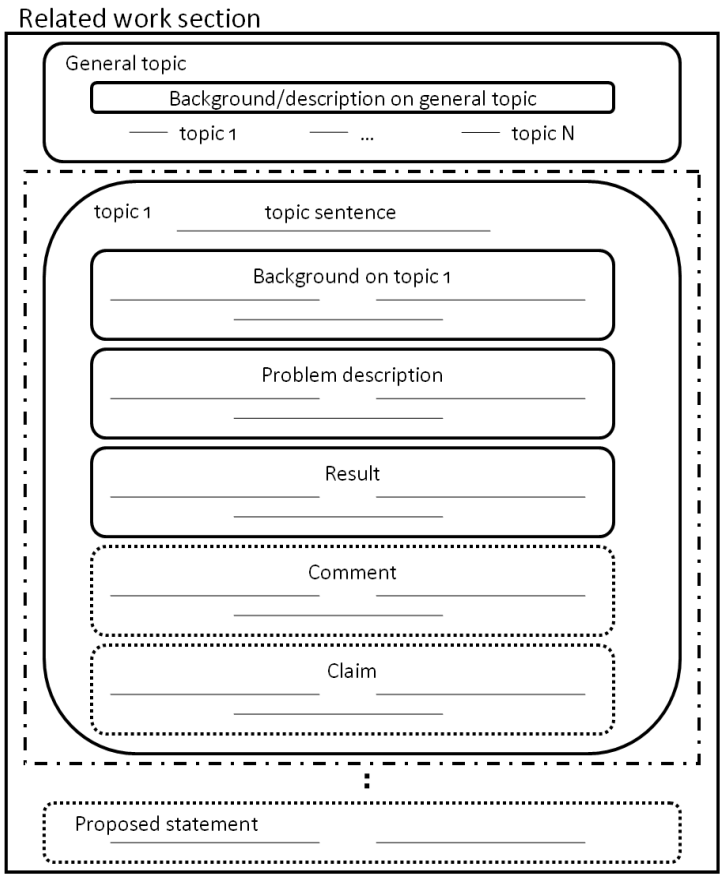


Figure 2.3: A general structure for RW summaries in scientific articles

cation" (lines 5–34) and topic 2 "cross-lingual classification" (lines 35–71). The topic 1 may contain two sub-topics "feature selection" (lines 6–19) and "probabilistic classifiers" (lines 20–33) whereas the topic 2 contains two other sub-topics "poly-lingual approach" (lines 45–58) and "cross-lingual approach" (lines 59–71). Each topic is usually presented with background knowledge. Various approaches of previous related works were then discussed to elaborate on each topic. Finally, the proposed statement is discussed (lines 73–78).

Since each RW summary can implicitly be associated with a topic hierarchy tree, the annotation of topical information in the **RWSData** dataset is required. I note that the

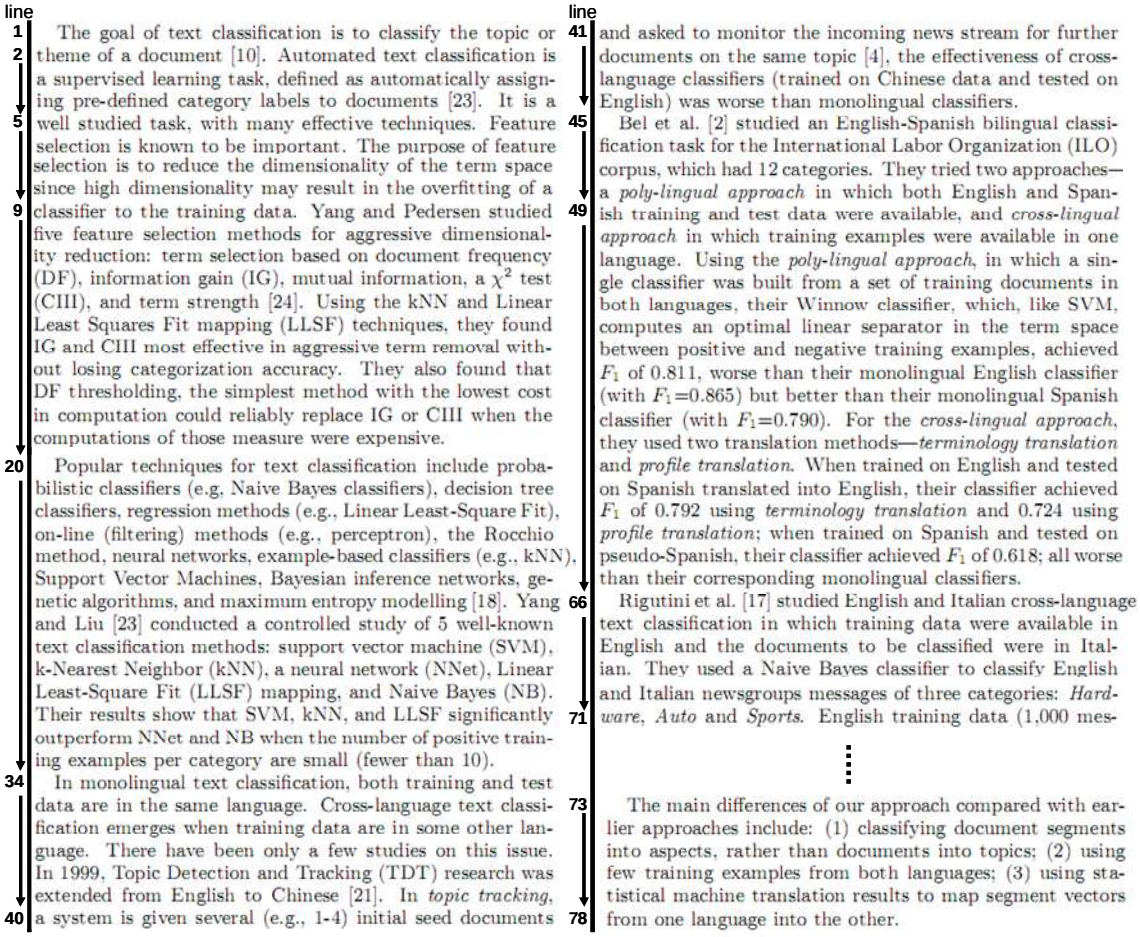


Figure 2.4: An example about structure of a RW summary in (Wu and Oard, 2008)

construction of topic hierarchy tree is subjective and that different annotators will end up with different topic hierarchy trees. I annotated this information for the **RWSData** dataset, following the general guidelines below.

- Carefully note the important topics for each related work.
- Identify the relationships (parent-child) among topics and construct the topic hierarchy tree.
- For each topic, provide a set of associated keywords. These keywords can appear

in RW sections. Note that it is unnecessary to read the original referenced articles to find keywords. Also, if a keyword already appears in the parent topic, it should not to appear in the children. Topics which have common parents may contain overlapping keywords.

After manually constructing the topic hierarchy trees, I compiled demographics on the dataset, as shown in Table 2.2 and Table 2.3 (columns N10, N11). As can be seen, the topic trees are simple, averaging 3.3 topic nodes in size and average depth of 1.8. Their simplicity furthers our claim that automated methods would be able to create such trees.

In addition to structure of RW summaries, I also explored the way the authors use citations within a RW summary. When describing related works (means referring to citations), authors have to choose some aspects of these works relevant to their current work to discuss. Some of aspects can be identical or complementary. Based on my observation on the **RWSData** dataset, I categorize how authors use citations in three ways:

- Citations that describe a unique aspect of a work. In this way, each recognized aspect is associated separately with an citation.

  For example:

  1) Zens and Ney (2007) remove constraints imposed by the size of main memory by using an external data structure. Johnson et al. (2007) substantially reduce model size with a filtering method.

- Citations that describe an aspect in common with other works. In this way, two or more citations are discussed in tandem.

  For example:

  1) Chan et al. (2007) and Carpuat and Wu (2007) improve translation accuracy using discriminatively trained models with contextual features of source phrases.

2) Training the transliteration model is typically done under supervised settings (Bergsma and Kondrak, 2007; Goldwasser and Roth, 2008b), or weakly supervised settings with additional temporal information (Sproat et al., 2006; Klementiev and Roth, 2006a).

- Citations that describe two or more complementary aspects, that differ from the authors' current work. This is usually to set up a contrast to show the advantages of the current work.

  For example:

  1) Unlike previous annotations of sentiment or subjective (Wiebe et al., 2005, Pang and Lee, 2004), which typically relied on binary 0/1 annotations, we decided to use a finer-grained scale, hence allowing the annotators to select different degrees of emontional load.

  2) Our chunk-based system takes the last word of the chunk as its head word for the purposes of predicting roles, but does not make use of the identities of the chunk's other words or the intervening words between a chunk and the predicate, unlike Hidden Markov Model-like systems such as Bikel et al. (1997), McCallum et al. (2000) and Lafferty et al. (2001).

Each of the above ways offers different levels of difficulty in exploring strategies for summarizing RW sentences. According to my understanding, the first way is the simplest for summarization. The second and third ones are harder because they require semantic processing to decide what is similar or dissimilar among relevant works. Automating such a step is beyond the current state-of-the-art NLP techniques.

2.3 Decomposition of Related Work Summaries

2.3.1 Related Studies

How do we ourselves (as humans) compose RW sections? A way to introspect on this human process is to decompose it. Solving the decomposition process may help figure out the feasible approach to RW summarization. Also, the approaches for decomposition vary, depending on the nature of the summaries. A useful distinction I find is to differentiate between single-document (Jing and McKeown, 1999; Jing, 2002; Ceylan and Mihalcea, 2009) and multi-document summaries (Banko and Vanderwende, 2004).

(Jing and McKeown, 1999; Jing, 2002) initiated the exploration of decomposing human-written summaries for news articles. They defined the decomposition as the process to infer the relations between the phrases in a summary composed by human summarizers and phrases in the original document. The studies hypothesized that such relations may come from the cut-and-paste operations which humans use to extract relevant texts from the original document to produce the summary. Specifically, the cut-and-paste operations comprise six main operations which are usually performed by humans such as: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, generalization/specification, and reordering. More descriptions of them can be found in (Jing, 2002).

Their decomposition shed light on the following three questions:

- Whether the summary is created by human cut-and-paste operations?
- Which components in the summary sentence come from the original documents and where in the original document do they come from? Note that the components may be of various granularity (e.g. words, phrases, clauses, or even sentences).
- How such components are constructed? Which human operations are used?

Their decomposition process for single-document summaries uses the Hidden Markov Model (HMM) which utilizes the underlying Viterbi algorithm. The algorithm starts by modeling each word in a summary sentence as a node in the HMM model. The transition among nodes is drawn based on the assumption that humans prefer to extract phrases than isolated words and are more likely to combine the adjacent sentences rather than combine sentences that are far apart. This assumption lead to some heuristic rules to assign the transition probabilities for HMM model. The decomposition was formulated as the problem of finding the most likely document position for each word in the input summary sentences. A case study in news domain was then carried out to examine the algorithm using both automatic and subjective human evaluation. The results showed that the proposed algorithm to decomposition using the HMM model worked very well on the selected corpus. It also suggested that approximately 78% of summary sentences in news articles was produced by humans using cut-and-paste operations on the original articles. Also, the technique of the decomposition of human-written summaries using HMM modeling was also applied successfully to the analysis of Japanese broadcast news domain in (Hideki Tanaka and Itoh, 2005). Recently, Ceylan and Mihalcea (2009) successfully adapted the above decomposition methodology capable of dealing with technical books. These promising results are interesting as I also want to examine the decomposition in the context of RW summaries.

2.3.2 The Alignment

Previous decomposition approaches which dealt with single-document summaries cannot be applied to my task of RW summarization, as this task takes input from multiple sources. It is also important to consider that scientific writing places firm limits on plagiarism; thus authors often limit their copying of set words or phrases from the original references. Due to this reason, they must use their own words to compose the RW summaries. This factor adds more difficulty to the decomposition of RW summaries.

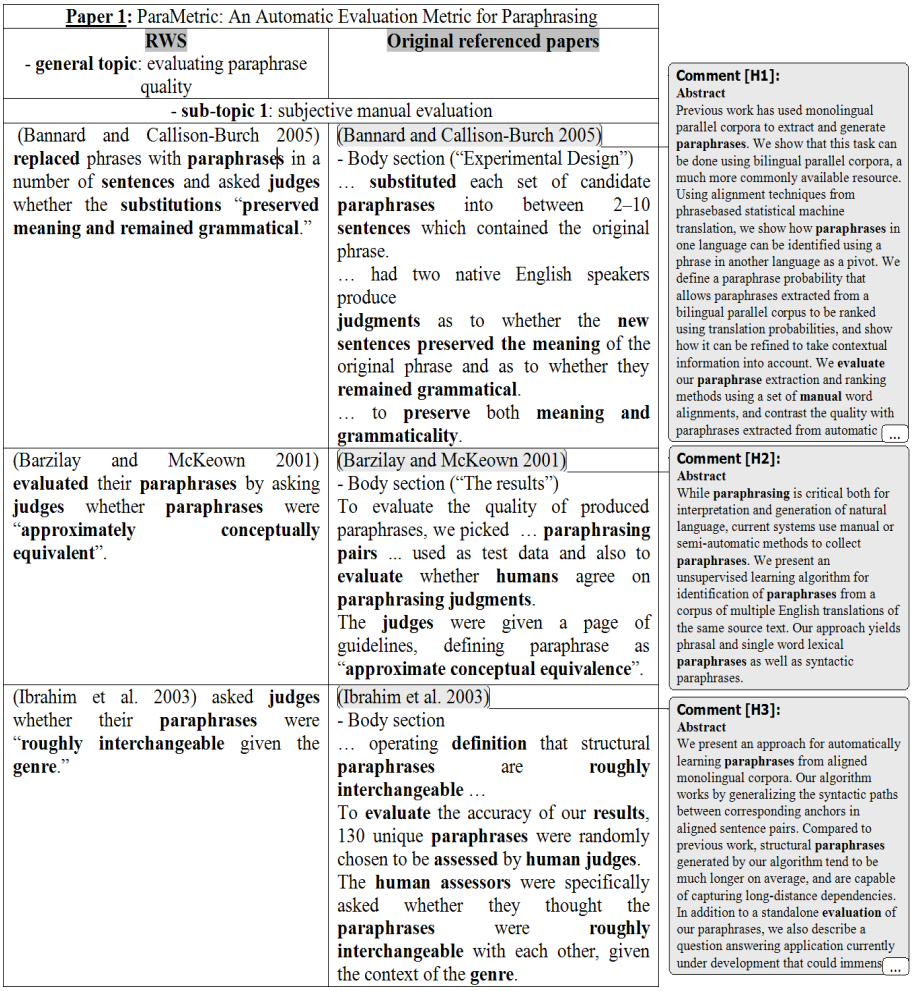


Figure 2.5: An illustrating example describing the analysis process

Thus, I conducted a manual analysis to examine whether the RW summaries contain words and phrases that originate from the referenced articles, as in the cut-and-paste technique. I randomly selected five RW summaries in the **RWSData** dataset and aligned them to the original referenced articles. The alignment was performed on components at various granularity such as: word, phrases, sentences. I also pinpointed which sections (e.g. abstract, introduction, body, discussion, conclusion, ...) these components come from.

Consider the first example in Figure 2.5 referring to the article **(Bannard and Callison-Burch, 2005)**. In this example, I observed that various words (e.g. "paraphrase") or phrases (e.g. "preserved the meaning and remained grammatical") are matched in both RW sentences and text fragments from original referenced articles. As observed, these words or phrases do not appear in the Abstract section of the referenced article.

After analyzing the five articles, I observed that a RW summary often refers to just some specific aspects (e.g. methods, results, evaluation processes ...) that relate to the topic of interest in the current paper. Thus, the RW sentences may be constructed from the text fragments that come from various sections in original referenced articles.

Further, based on my observation on the **RWSData** dataset, I categorize RW sentences into three categories:

- **RWS1**: (XX, 2000) ... - a summary of an aspect mentioned in referenced article with respect to a specific topic. For example: **(Barzilay and McKeown 2001)** evaluated their paraphrases by asking judges whether paraphrases were "approximately conceptually equivalent".

- **RWS2**: Topic (XX, 2000) ... - summary of a topic. For example: Supervised approaches such as **(Black et al. 1998)** have used clustering to group together different nominals ...

- **RWS3**: Fact or Opinion (XX, 2000) ... - evidence-based reference. For example: Co-training **(Riloff and Jones, 1999; Collins and Singer, 1999)** begins with ...

- **RWST**: template-based summary, focus mainly on something about survey paper, dataset, metric, tool, and so on. For example: Sebastiani's **survey** paper **[23]** provides an overview of techniques in text categorization, ...

Figure 2.6 shows the statistics (occurrence frequency) about possible positions of all RW categories in the original referenced articles.

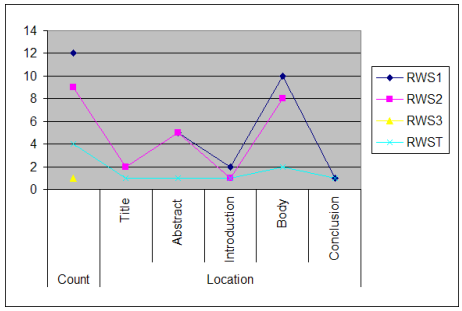


Figure 2.6: Statistics of possible positions of all RW categories

As can be seen in Figure 2.6, the most likely positions which RW summary sentences usually come from is the body section of the referenced articles decreasingly following by the Abstract, the Title, the Introduction and the Conclusion sections. Note the count in this figure means the number of instances to be analyzed.

2.3.3 Revisions by Human Writers

Another concern in the decomposition process is to find out which operations (also called revisions) human summarizers use to construct the RW summaries. Here I adapt five of the original operations as defined in (Jing and McKeown, 1999) that are used in creating RW summaries by humans observed in the **RWSData** dataset.

Sentence Reduction

This operation aims to remove less important components from a sentence and then use the reduced sentence in a summary.

Text fragment 1: ... substituted each set of candidate **paraphrases** into between 2-10 **sentences** which contained the original **phrase**.

RW sentence: (Bannard and Callison-Burch 2005) replaced **phrases** with **paraphrases** in a number of **sentences** ...

Sentence Combination

This operation combines several different fragments/sentences together to construct a new sentence. Sentence combination can be used in combination with sentence reduction.

Text fragment 1: ... **substituted** each set of candidate **paraphrases** into between 2-10 **sentences** which contained the original **phrase**.

Text fragment 2: ... had two native English speakers produce **judgments** as to whether the new sentences **preserved** the **meaning** of the original **phrase** and as to whether they remained **grammatical**.

RW sentence: (Bannard and Callison-Burch 2005) **replaced phrases** with **paraphrases** in a number of **sentences** and asked **judges** whether the **substitutions** "preserved **meaning** and remained **grammatical**".

Syntactic Transformation

This operation transforms some components into other syntactic forms. An example is the movement of a subject or a change in word ordering.

Text fragment 1: ... to **preserve** both meaning and **grammaticality**.

RW sentence: ... "**preserved** meaning and remained **grammatical**".

Lexical Paraphrasing

This operation replaces other phrases/words in a sentence. Consider the following example in which the word "substituted" is replaced by another word "replaced":

Text fragment 1: ... **substituted** each set of candidate paraphrases into between 2-10 sentences which contained the original phrase.

RW sentence: (Bannard and Callison-Burch 2005) **replaced** phrases with paraphrases in a number of sentences ...

Generalization/Specification

This operation replaces some certain phrases/words in a sentence with a higher- (generalization) or lower- (specification) level descriptions. In the following example, "large text corpora" in the original sentence is replaced by "the Web" in the summary sentence. This is the case of generalization.

Text fragment 1: We present an unsupervised learning algorithm that mines **large text corpora** for patterns that express implicit semantic relations.

RW sentence: (Turney 2006a) presents an unsupervised algorithm for mining **the Web** for patterns expressing implicit semantic relations.

Note that the overall meaning of a sentence after using the above revisions needs to be preserved. Also, all of the above revisions are not used alone but usually combined together. Intuitively, handling **all** of the above revisions for RW summarization is not feasible due to their complexity, especially in two revisions: lexical paraphrasing and generalization/specification. Thus, I assume that the RW summaries are supposed to be constructed from three revisions: sentence reduction, sentence combination and syntactic transformation.

2.4 Related Work Representation

The previous discussion has focused on describing the characteristics of RW summaries which can be beneficially used in ATS. The next step is to examine how to generate and represent a complete RW summary. My aim here is to investigate which important factors make such summaries easy-to-read and fluent in terms of cohesion and coherence. Cohesive[4] is a grammatical and lexical relationship within a text or sentence, indicating surface and textual units and their interconnectedness. In contrast to cohesion, coherence[5] normally refers to a discourse relation between larger units of text (*e.g.* clauses,

[4] http://en.wikipedia.org/wiki/Cohesion_(linguistics)
[5] http://en.wikipedia.org/wiki/Coherence_(linguistics)

sentences, paragraph) which represents structuring of the text at a macro level by text schemes and rhetorical structures. Text cohesion and coherence can greatly contribute to text readability. Classic frameworks that describe computational cohesion and coherence include (Grosz et al., 1995; Kibble and Power, 2004; Barzilay, 2005). In the context of RW summarization, there are two main factors which reflect summary representation. They are topic transition and local coherence.

This section will give a deep manual analysis on RW representation based on topic transition and local coherence and then figure out the appropriate representation which are developed in the proposed system discussed in Chapter 4.

The analysis was carried out over a set of published conference articles in Computational Linguistics. I randomly chose 30 articles in leading major conferences (e.g. ACL, NAACL) over years for my analysis. There are 5 articles from NAACL'09, 12 ones from ACL'07 and the rest from ACL'09. I refer to this portion of the original dataset as **RWSData-Sub**. Note that the **RWSData-Sub** dataset differs from the **RWSData** dataset because the **RWSData** dataset will be used to weakly supervise the summarization process in the system (discussed in Chapter 4) whereas the **RWSData-Sub** will be used as a post-processing step in the generation process. As such, the evaluation of generated RW summaries *versus* gold standard RW summaries will be fair.

Since a RW summary is a topic-biased summary in a hierarchical fashion, topic transition refers to the appropriate topic representation and ordering which ensures that the output summary is coherent. Given a topic hierarchy tree, nodes first are ordered in either a depth-first or breath-first traversal. According to my observation on real RW summaries, depth-first traversal is preferred. Then, each topic node together with associated summarized information is presented.

2.4.1 Topic Transition

My analysis reveals that there are two main types of topic representation within RW summaries. Type 1 uses transition sentences to connect ordered topic nodes. Type 2 is simpler, referring to the representation of topics nodes as topic titles. Figures 2.7 and 2.8 give examples of Type 1 and Type 2 topic representations, in which a RW section is associated with a topic hierarchy tree and topic descriptions. Each node in the figures is linked with a text fragment (surrounded by a rectangle with node notation at the upper left corner) which describes its content.

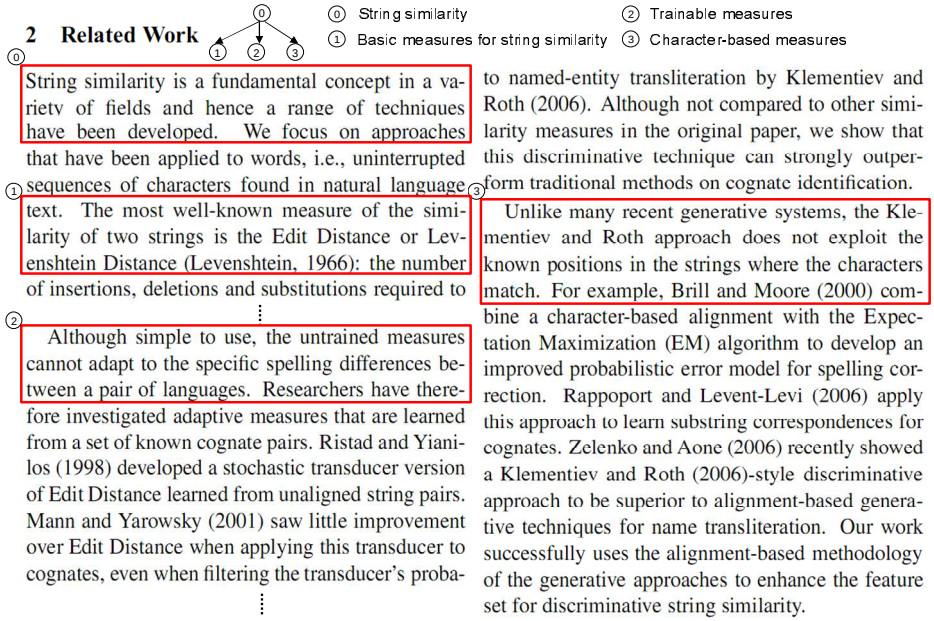


Figure 2.7: An example of Type 1 topic representation in the RW section of (Bergsma and Kondrak, 2007).

In the Figure 2.7, the authors first introduced the general topic (node 0) following by a sub topic (node 1). Moving from node 0 to node 1, they started the statement with "the most well-known measures ..." to introduce node 1. After finishing the discussion on node 1, they gave their ideas on node 1 (i.e., simple to use, recognized that measures mentioned in this topic are untrained ones) to move the discussion to node which refers to trainable measures contrary to node 1. Actually, this expression can be thought of as

a discourse relation (i.e., a *CONTRAST* relation). Similarly, the movement from node 1 to node 2 also uses the CONTRAST discourse relation. Thus, for Type 1, topic nodes are implicitly expressed using transition sentences. Meanwhile, in the Figure 2.8, the authors explicitly show topic nodes by using topic sections. Such topic sections is then discussed separately. If a topic has sub-topics, its topic section will be structured with sub-topic sections. As such, Type 1 and 2 show two different ways in representing the transitions between topics given the structure of a topic hierarchy tree. Each of two has its advantages and disadvantages. Generally, Type 1 seems to be more natural in terms of topic coherence and easier to read than ones using Type 2.

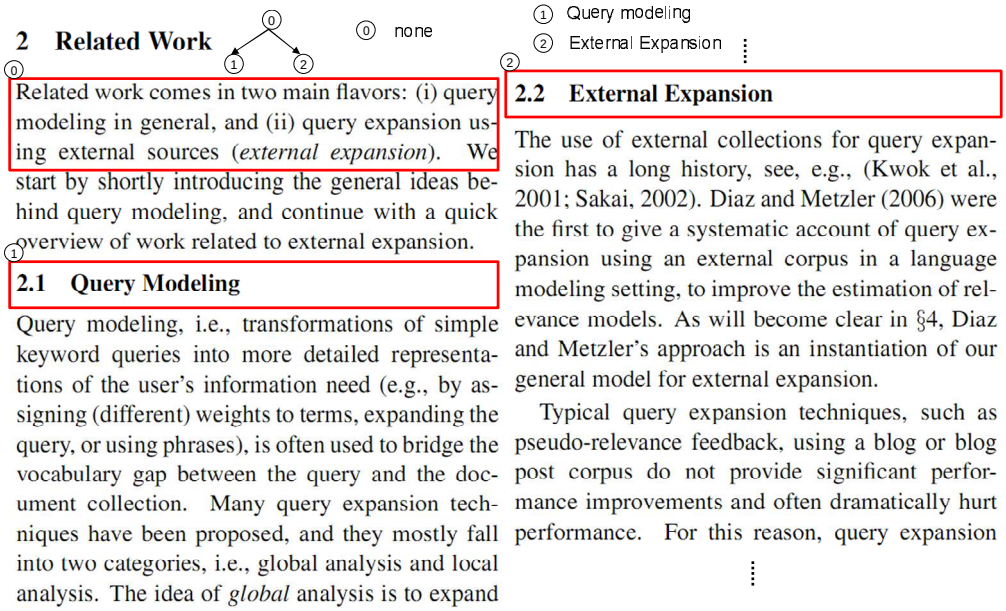


Figure 2.8: An example of Type 2 topic representation in the RW section of (Weerkamp et al., 2009).

To gain further insight, I also counted on how many articles used RW sections for each type of topic representation. This exercise showed that the majority – 23 of 30 – of the RW summaries used a Type 1 representation.

Further, which topic representation type should be used in the automatic system for RW summarization? In fact, given a topic hierarchy tree, Type 2 representation is

simple to process. Type 1 representation is non-trivial because it requires an external discourse processor to assign pre-defined discourse relations (e.g. *CONTRAST*, *ELABORATION*) for a given pair of topic nodes. To do this raises the difficult problem of discourse processing, which I feel is out of scope for my thesis. Hence, I need to prove that Type 1 is sufficient for topic representation. The following section turns to local coherence of both Type 1 and 2 to validate this.

2.4.2 Local Coherence

Local coherence refers to an instance of discourse processing which aims to reflect two main factors – the syntactic realization of discourse entities and transitions between focused entities (Nenkova and McKeown, 2003). In summaries of news articles, the focus is on mentions of people (Nenkova and McKeown, 2003). In the context of RW summaries, entities refers to citations which are referenced articles mentioned in the summaries. Nenkova and McKeown (2003) did a corpus study to derive a statistical model based on Markov Chains to resolve the syntactic realization of mentions to people in news summaries. The study investigated the differences between first and subsequent mentions corresponding to people, analyzing the realization of their components: premodifiers, names, and post-modifiers. These kinds of mentions then help to infer implicit features to automatically build natural co-reference chains, i.e., the chain of all mentions of an entity within a summary. The summary post-corrected with this automatic resolving step was proved to be more coherent than the original one.

I found that entities (*a.k.a.* mentions to people) in news summaries are usually repeated. Also, events of these entities are continuous. It helps to easily build co-reference chain of entities. This differs from RW summaries since entities (*a.k.a.* mentions to citations) only appear at certain places within each topic. Thus, the method that the earlier work suggested may not apply.

In this section, I will examine various relevant issues about how the mentions to

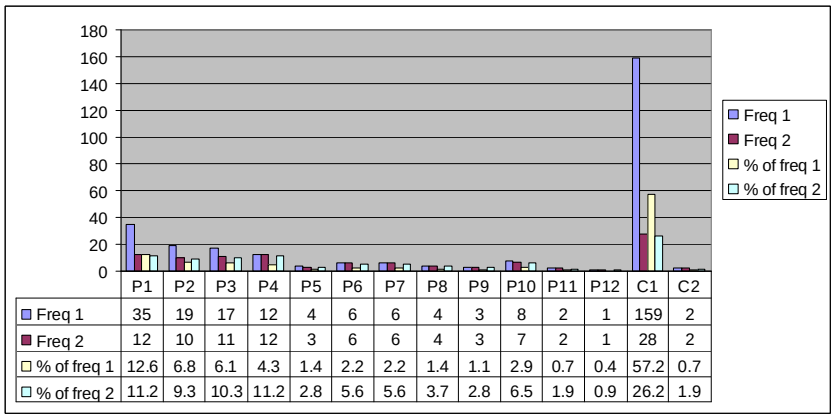
citations are presented within RW summaries, by analyzing 30 articles from **RWSData-Sub**. Given the focus on mentions to citations, I identified 14 patterns that are regularly used within realistic RW summaries. A pattern here is a first or subsequent mention to a citation. Descriptions and examples of these patterns are given in Table 2.4.

No.	Pattern	Notation	Mention	Example
1	<ref1>... They/he/she ...	P1	subsequent	**Hearst (1998)** presents a method to automate the discovery of WordNet relations, ... **She** explores several patterns for ...
2	<ref1>... (T)heir/his/its [model/approach/algorithm/...](s) ...	P2	subsequent	**Lauer (1995)** tackles the problem of semantically disambiguating noun phrases by ... **His method** involves searching a corpus ...
3	<ref1>...<ref2>... Such/these/the studies/approaches/algorithms ...	P3	subsequent	**(Hasegawaetal, 2004; Hassanetal, 2006)** proposed unsupervised clustering methods that ... **These studies**, however, focused on the classification of pairs that ...
4	<ref1>... [This/that work/approach/task/strategy/...] ...	P4	subsequent	**Pasca (Pasca, 2007b; Pasca, 2007a)** illustrated a set expansion approach that ... **This approach** is similar in flavor to ...
5	(T)he [work/use/...] of <ref1>...	P5	first	**The work of Och et al (2004)** is perhaps the best known study of new features and ...
6	<ref1>...<ref2>... (O)ther(s) (work)	P6	subsequent	Some approaches coarsely discriminate between biographical and non-biographical information **(Zhouetal., 2004; Biadsyetal., 2008)**, while **others** go beyond binary distinction by ...
7	More/some recent approaches <ref1>...<ref2>...	P7	first	**Some recent work (Li et al., 2006; Xu et al., 2006)** has attempted to introduce preference into a probabilistic framework ...
8	<ref1>'s [work/study/...] ...	P8	first	A third difficulty with **(Och et al., 2002)'s study** was that it used MERT, which ...
9	<ref1>... (<ref2>) ... This/the line of work/research ...	P9	subsequent	**Another line of research (Watanabe et al., 2007; Chiang et al., 2008)** tries to squeeze as many features as possible from ...
10	The/Another work/study <ref1>...	P10	first	**Another work (Koehn and Knight, 2003)** showed improvements by ...
11	<ref1>...<ref2>... [All these/All of the] [systems/works]	P11	subsequent	**Pasca (Pasca, 2004)** presented a method for acquiring named entities in ... **Etzioni et al. (Etzionietal., 2005)** presented the KnowItAll system that ... **All the systems** mentioned rely on ...
12	In <ref1>, (the authors) ...	P12	subsequent	**In (Harabagiu et al., 2001)**, the path patterns in WordNet are utilized to ...
13	<ref1>...	C1	first	**Ponzetto and Strube (2006)** suggest to mine semantic relatedness from Wikipedia, ...
14	<ref1>...<ref1>...	C2	subsequent	Another measure, suggested by **Church and Gale (1995a)** is burstiness which ... **Church and Gale** also noted that ...

Table 2.4: Details on 14 patterns explored in the analysis.

Such patterns show that people tend to use a variety of patterns to represent mentions to citations. Each pattern plays an important role in connecting sentences in the

summary. Note that patterns C1 and C2 are special; in that they represent the direct uses of the citation (see examples on C1 and C2 patterns in Table 2.4). In addition, the fourth column in Table 2.4 additionally gives two kinds of mentions which each pattern associates with. As a result, there are 5 “first” and 9 “subsequent” mentions recognized in this analysis.

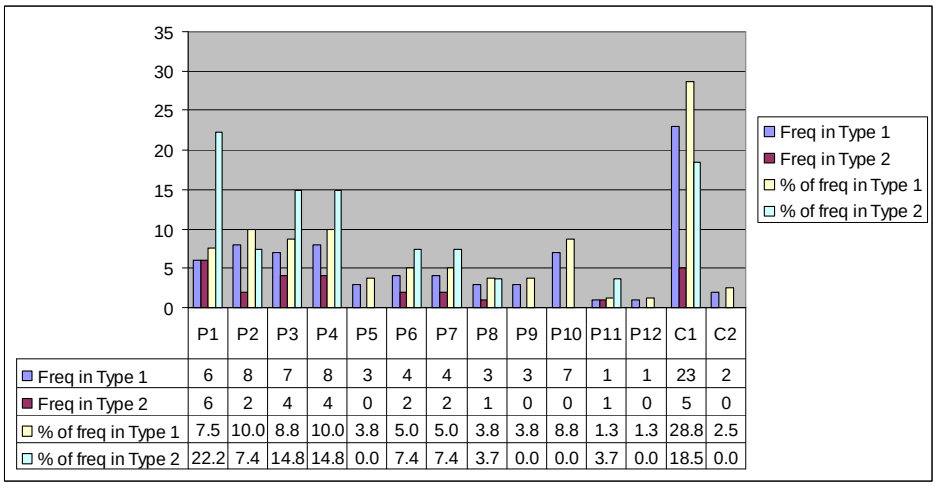


	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	C1	C2
Freq 1	35	19	17	12	4	6	6	4	3	8	2	1	159	2
Freq 2	12	10	11	12	3	6	6	4	3	7	2	1	28	2
% of freq 1	12.6	6.8	6.1	4.3	1.4	2.2	2.2	1.4	1.1	2.9	0.7	0.4	57.2	0.7
% of freq 2	11.2	9.3	10.3	11.2	2.8	5.6	5.6	3.7	2.8	6.5	1.9	0.9	26.2	1.9

Figure 2.9: **Statistics for 14 patterns over the RWSData-Sub dataset**. Note that each pattern is associated with four columns. The first column (“Freq 1”) means the number of instances which each pattern appears over the dataset. The second one (“Freq 2”) means the number of RW sections (over 30 in the dataset) in which each pattern appear. The third and fourth ones are the percentages of “Freq 1” and “Freq 2” over 14 patterns, respectively.

To explore how frequent such patterns are used in RW summaries, I conducted the calculation on frequencies of patterns over the dataset. The calculation is simply based on the number of instances of each pattern observed from sample RW sections. The detail of statistics is given in Figure 2.9.

Figure 2.9 shows that the pattern of direct citation representation (C1) is used most frequently (57.2%). This pattern is the simplest way to mention to a citation. Most observed RW summaries (28/30) use this pattern. Meanwhile, people rarely (2/30) use the pattern C2 (note that C2 means the use of C1 repeatedly). This justifies the statement about human preference of less informative subsequent mentions (Krahmer and Theune, 2002). Remarkably, patterns that are used frequently following the pattern C1 are P1,



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	C1	C2
Freq in Type 1	6	8	7	8	3	4	4	3	3	7	1	1	23	2
Freq in Type 2	6	2	4	4	0	2	2	1	0	0	1	0	5	0
% of freq in Type 1	7.5	10.0	8.8	10.0	3.8	5.0	5.0	3.8	3.8	8.8	1.3	1.3	28.8	2.5
% of freq in Type 2	22.2	7.4	14.8	14.8	0.0	7.4	7.4	3.7	0.0	0.0	3.7	0.0	18.5	0.0

Figure 2.10: **Statistics for 14 patterns that appear in each type of topic transition representation over the RWSData-Sub dataset**. Note that each pattern is associated with four columns. The two first columns are the number of RW sections (over a total of 30 in the dataset) in which each pattern appears referring to each type of topic representation. The two final columns are percentages of the first two over the 14 patterns.

P2, P3, and P4 with percentages 12.6%, 6.8%, 6.1%, and 4.3%, respectively. Note that all these patterns are subsequent patterns.

They also appear in more than 10 RW sections in **RWSData-Sub**. The observation is that people tend to prefer relatively simple patterns to represent mentions (e.g. P1, P2, P3, P4 and C1). Other patterns (P5 to P11) are more complex and used in specific cases. Especially P12 is quite simple but is not used frequently (only 1 time). Also, people usually use patterns that are combined together to flexibly represent citations (e.g. C1 combining with P1, P2). Such the flexible use of patterns makes the created RW sections easier-to-read and coherent. However, based on my observation over the **RWSData-Sub** dataset, patterns are combined together without specific combination rules. Thus, it makes the automatic generation for such patterns problematic.

Figure 2.10 shows the statistics of patterns associated with topic transitions. For simplicity, this figure only shows the number of RW summaries in which each pattern

appears, with respect to topic representation types. I observe that patterns C1, P1, P2, P3, and P4 appear most frequently in each type of topic representation as compared to other patterns. However, the pattern C1 in Type 1 no longer holds skew distribution as in Type 2 (23 in Type 1 *vs.* 5 in Type 2). In particular, three patterns (P1, P3, P4) are used in Type 2 more frequently than in Type 1, especially P1 (increased dramatically from 7.5% to 22.2%). Some other patterns (P5, P9, P10, and P12) are no longer used in Type 2. In sum, over 14 patterns, patterns P1, P2, P3, P4 and C1 are sufficient for both Type 1 (65.1%) & 2 (59.2%).

Table 2.5 counts the appearance of each pattern, and provides information on the sentence length of the summary, and topic representation type. The average and standard deviation of appearance of each pattern in the summary is 3.5 and 1.3, respectively. The average and standard deviation of sentence-based summary length is 17.1 and 5.2, respectively. As such, a RW summary which has the sentence-based length in range of 17.1 ± 5.2 may use 3.5 ± 1.3 transition patterns.

In sum, in order to decide the appropriate setting for representing RW summaries, one may depend on two factors: 1) choose between the two topic transition types and, 2) decide the appropriate patterns and their combinations for local coherence with respect to the chosen type of topic transition. Though an appropriate setting for RW representation can be chosen easily at a human level, however, this is still problematic for computer programs.

The detailed analysis above has explored discrete statistics in which humans use topic transition and local coherence for RW representation. From this analysis, I believe that creating topic transitions only using Type 2 transitions, along with patterns (*e.g.* P1, P2, C1) for representing local coherence, are sufficient for people to understand a RW summary. In my work, I will choose this setting for representing RW summaries during generation stage implemented in the proposed ReWoS system in Chapter 4.

2.4.3 Citation Representation

No.	Summary ID	Patterns	Freq1	Freq2	Length	Type
1	N09-1002	P2, C1(6)	7	2	21	Type 1
2	N09-1022	P1, P2, P8, C1(7)	10	4	12	Type 1
3	N09-1025	P5(3), P2, P6, P8, P9, C1(2)	9	6	9	Type 1
4	N09-1048	P1, P2, C1(8)	10	3	10	Type 1
5	N09-1060	C1(2), C2(1)	3	2	19	Type 1
6	P07-1016	P7, C1(11)	12	2	15	Type 1
7	P07-1017	P7, P9, P10, C1(6)	9	4	17	Type 1
8	P07-1030	P7, P3(2), P4, C1(14)	18	4	12	Type 1
9	P07-1036	P4(3), P10, C1(8)	4	3	19	Type 1
10	P07-1055	P2, P3, P4, P10, C1(8)	12	5	16	Type 1
11	P07-1067	P5, P1(6), P2(3), P12, P8, C1(3)	15	6	22	Type 1
12	P07-1069	P3, P10, C1(4)	6	3	15	Type 1
13	P07-1072	P1(2), P6, C1(5)	8	3	10	Type 1
14	P07-1083	P3, C1(7)	8	2	22	Type 1
15	P07-1124	P1(5), P3, P4, P6, P7, C1(1)	10	6	24	Type 2
16	P07-1125	P1(2), P2(2), P6, P7, C1(6)	12	5	14	Type 2
17	P07-3014	P1(5), P2(3), P4, C1(2)	11	4	31	Type 2
18	P09-1002	P4(2), P5, C1(5), C2(1)	8	4	24	Type 1
19	P09-1009	P4, P10(2), C1(4)	7	3	17	Type 1
20	P09-1010	P3(3)	3	1	13	Type 2
21	P09-1024	P6, P3(3), P4, C1(2)	7	4	15	Type 1
22	P09-1050	P4(2), P11, C1(5)	8	3	16	Type 1
23	P09-1055	P9, P3(2), P10, C1(3)	7	4	13	Type 1
24	P09-1062	P1, P3, P11	3	3	15	Type 2
25	P09-1077	P1(3), P2(2), P4, C1(7)	13	4	16	Type 2
26	P09-1083	P1(3), P3(2), C1(3)	8	3	20	Type 1
27	P09-1113	P2(2), P6, P7, C1(5)	9	4	10	Type 1
28	P09-1114	P1(5), C1(10)	15	2	22	Type 1
29	P09-1119	P1, P8, C1(4)	6	3	24	Type 2
30	P09-1120	P1, P2, P10, C1(11)	14	4	20	Type 1

Table 2.5: **Detailed counts of the 14 patterns in 30 RW sections in RWSData-Sub**. "Summary ID" is the ID of the RW summary; "Patterns" list the patterns that appear in the summary (the parenthetical numbers indicate the frequency of the corresponding pattern); "Freq1" and "Freq2" denote the total frequency and the distinct number of patterns that appear in the summary; "Length" gives the summary length in sentences; and "Type" refers to the type of topic representation.

The above analysis has stressed the importance of the use of direct citation representation (patterns C1, C2) in writing RW summaries. This section provides different ways to use them. The observation on the dataset shows that there are two categories of citation representation, being consistent with standard citation uses in scientific writing[6]:

- **Single Citation**. This category is divided into two sub categories as follows:

[6] `http://www.stat.psu.edu/~surajit/present/bib.htm`

– **Textual Cite (used under an LaTeX symbol: citet)** is usually used when starting new topic sentences. Citations usually appear as subjects of sentences.

For example: ***Cucerzan and Brill (2004)*** *pioneered the research of query spelling correction, with an excellent description of how a traditional dictionary based speller had to be ...*

No	Summary ID	Multiple Citation	Single Citation	
			Parenthetical Cite	Textual Cite
1	N09-1002	1	2	3
2	N09-1022	0	1	7
3	N09-1025	2	1	1
4	N09-1048	0	4	4
5	N09-1060	0	2	3
6	P07-1016	4	6	0
7	P07-1017	0	9	1
8	P07-1030	7	8	0
9	P07-1036	2	5	0
10	P07-1055	2	6	4
11	P07-1067	0	2	9
12	P07-1069	2	0	2
13	P07-1072	2	3	2
14	P07-1083	0	3	7
15	P07-1124	2	2	6
16	P07-1125	0	6	7
17	P07-3014	0	0	12
18	P09-1002	0	3	1
19	P09-1009	3	4	1
20	P09-1010	3	1	0
21	P09-1024	4	2	0
22	P09-1050	0	0	7
23	P09-1055	2	4	0
24	P09-1062	4	8	3
25	P09-1077	1	2	5
26	P09-1083	1	8	0
27	P09-1113	1	2	8
28	P09-1114	2	4	10
29	P09-1119	2	1	4
30	P09-1120	2	13	0

Table 2.6: Detailed statistics of categories for citation representation.

– **Parenthetical Cite (used under an LaTeX symbol: citep)** is used to mention specific topics/tools/data/papers/... that the authors want readers to refer to.

For example: On the other hand, there have been many semi-supervised approaches in numerous applications such as self-training in word sense dis-

*ambiguation **(Yarowsky, 2005)** and parsing **(McClosky et al., 2008)**.*

- **Multiple Citation**. This category aims to generally list multiple referenced articles to give support to topics mentioned.
  *For example: This was used, for example, by **(Thelen and Riloff, 2002; Collins and Singer, 1999)** in information extraction, and by **(Smith and Eisner, 2005)** in POS tagging.*

Depending on functionality of each category, one may choose the appropriate one that suits specific situations given.

Also, there are many realizations of the above representation categories. For example: people may use "Jones et al. (1990)" or "(Jones et al., 1990)" for **single citation** and "Jones et al. (1990); James et al. (1991)" or "(Jones et al., 1990; James et al. 1991)" for **multiple citation**.

Furthermore, it is helpful to observe how frequently each of the above citation representation category is used in realistic RW summaries. To do this, I also conducted a statistics over the same dataset (**RWSData-Sub**). Note that if a multiple citation is already counted, single citations within that multiple citation is not counted again. Table 2.6 provides such a detailed statistics.

As can been seen in this table, the observed RW summaries use all categories (11 times) or just a few (3 that use just one and 16 that use just two). This supports the observation that authors prefer using two or three categories for citation representation. The average and standard deviation of "multiple citation" is 1.6 and 1.7, and "single citation" category with "Parenthetical Cite" is 3.7 and 3.1, and with "Textual Cite" is 3.6 and 3.5, respectively. Together with the summary length information shown in Table 2.5, on the other hand, a RW summary with the length in range of 17.1 ± 5.2 uses 1.6 ± 1.7 time(s) for "multiple citation" category, 3.7 ± 3.1 time(s) for "single citation" category with "Parenthetical Cite" and 3.6 ± 3.1 time(s) with "Textual Cite".

Table 2.5 also shows that people may use both "Parenthetical Cite" and "Textual

Cite" categories for "single citation" (in most cases) instead of using standalone ones. In addition, people may not use "multiple citation" but **MUST** use at least one category of "single citation".

The manual analysis discussed so far on various aspects of RW summaries will be helpful in developing the summarization (Sections from 2.2 to 2.3) and generation methods (Section 2.4). It provides a detailed vision about the behaviour of people in writing complete RW summaries. Such a manual analysis will play a role as guideline towards automated summarization and generation of RW summaries, which leads to the implementation of the proposed ReWoS system (Chapter 4).

2.5 Evaluation Metrics

In order to assess the quality of output summaries, it is also worth considering evaluation methods. In this section, I first review evaluation measures used in summarization community and then assess whether they are sufficient for evaluation of RW summaries. In addition, I also present my thoughts about expected metrics for both automatic and manual means which are designed specific to the task of RW summarization.

2.5.1 Previous Metrics

There have been metrics developed expressively for the evaluation of automatic summarization. Such evaluation metrics are designed to be flexible and applicble to both single- and multi- document summarization. Here I consider three major metrics used regularly in the summarization literature: ROUGE (Lin, 2004), Pyramid (Nenkova et al., 2007), and DEPEVAL (Owczarzak, 2009).

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It was proposed by Lin in (Lin, 2004). ROUGE is based on the key idea which is to measure the content coverage at various granularity (e.g. n-grams, word sequences, word pairs)

between human-generated reference summaries and computer-generated summaries. Inspired from the calculation of content similarity, he suggested different variants of ROUGE including ROUGE-N (N-gram Co-Occurrence Statistics), ROUGE-L (Longest Common Subsequence), ROUGE-W (Weighted Longest Common Subsequence), and ROUGE-S (Skip-Bigram Co-Occurrence Statistics). These ROUGE scores were proven to correlate reasonably with human judges. Note that all ROUGE scores has been successfully implemented in the ROUGE package[7].

The Pyramid method (Nenkova et al., 2007) observes that the content of a summary is characterized by different "information nuggets" or Summary Content Units (SCUs). Each SCU can be assigned a weight to favor its importance. All possible SCUs are manually extracted from both human and automatic summaries. The assessors then determine how many SCUs are shared between them to score the summaries. This method is very expensive and time-consuming because it requires labour to create the requisite human judgments.

Recently, Owczarzak (2009) proposed a novel method (namely DEPEVAL) for automatic summarization evaluation based on lexical dependency relations in sentences. Each such relation is represented as a triplet: relation_name(governor, dependent) (*e.g.* subject(resign, John)), which is normally extracted from a statistical dependency parser (*e.g.* Stanford Parser (de Marneffe and Manning, 2008)). The basic idea behind automatic evaluation of DEPEVAL is that the correlation between human and automatic summaries is measured by the set of overlapping dependency relations both of them contain. The empirical evaluation on the TAC 2008 and the DUC 2007 data sets shows that DEPEVAL provides a comparable or better confidence than previous evaluation metrics like ROUGE scores.

[7] `http://berouge.com/default.aspx`

2.5.2 Observation and Suggested Metrics

I observe that existing methods may contain some problems applied to evaluation for RW summarization. For example, ROUGE may cause the inconsistent problem as shown in Figure 2.11. In this figure, assume that the reference summary and two candidate summaries have some text fragments referring to different referenced articles (e.g. articles [1] and [2]). Initially, if the reference information is not considered, the first candidate summary has four overlapping words with the reference summary whereas the second candidate summary only has three.

It turns out that the candidate summary 1 is preferred according to the way ROUGE is computed. Otherwise, the reference information is considered, C in [1] and A in [2] of the candidate summary 1 may not refer to C in [2] and A in [1] of the reference summary, respectively. Thus, the candidate summary 1 actually has only two overlapping words with the reference summary. In this way, the second candidate summary is preferred, in contrast to the previous case. This situation is also valid in using two other evaluation metrics, the Pyramid (Nenkova et al., 2007) and DEPEVAL (Owczarzak, 2009) because they only differ from ROUGE in the way the content similarity is evaluated (overlapping N-grams with ROUGE, content units extracted by humans with Pyramid, and dependency relations with DEPEVAL).

Thus, it is very important to adjust existing methods suitable for evaluation of automatic RW summaries. The main idea to adjust them is to select appropriate information in comparing between human and automatic summaries. Information within each referenced article in the automatically generated summary needs to be compared consistently with the appropriate correlate in the human summary.

Assume that ROUGE metric is given to compute lexical content similarity. In this case, any equivalent metric (*e.g.* DEPEVAL, Pyramid) can be used in replacement of ROUGE. I choose ROUGE as a typical example to represent my idea.

In light of these problems, I extend ROUGE to create two measures for the eval-

uation of RW summaries that address these issues. They are ROUGE-Ref (ROUGE **Ref**erence) and ROUGE-Ref-T (ROUGE **Ref**erence with **T**opic). ROUGE-Ref means that information referring to referenced articles is grouped together and the score is calculated within each group using original ROUGE scores. Meanwhile, ROUGE-Ref-T simply adds the topic information into ROUGE-Ref. ROUGE-Ref-T is based on the observation that two text fragments may be different according to their topics. Note that the **ROUGE-Ref-T** requires a topic assignment with respect to the topic hierarchy tree for each sentence in the final RW summaries. As such, topic hierarchy tree is an important prerequisite of calculating ROUGE-Ref-T. Intuitively, my two extended ROUGE measures are reasonable but they need to be examined the correlation with human judges to compare with the original ROUGE.

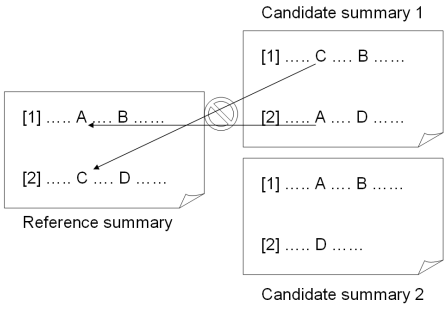


Figure 2.11: An illustrating example describing the inconsistent problem in evaluating the RW summaries using original ROUGE

Since automatic evaluation with metrics like ROUGE scores does not allow much introspection, I decide to turn to human evaluation. In this thesis, in addition to automatic evaluation with ROUGE metric, I will also propose different human evaluation metrics designed specific to the task of RW summarization which is given the details in Chapter 5.

Chapter 3

Literature Review

My proposed problem is a specific instance of the automatic text summarization (ATS) problem, which has been investigated within NLP community for nearly 50 years. While general ATS is outside the scope of my thesis, a general overview of summarization research is still instructive, and I refer the interested reader to a few excellent surveys (Ding, 2004; Das and Martins, 2007; Jones, 2007) and books (Mani, 2001; Hovy, 2003). Since the literature of summarization research is already well-documented by these sources, I focus mainly on reviewing the works on summarization in the domain of scientific texts that relate to the specific proposed problem of related work summarization.

Automated related work summarization is significantly different from traditional summarization (*e.g.* news) in several respects. First, it is limited to the domain of scientific discourse, which contains specific features that are have currently not been explored by others. Second, the related work summary should follow the structure of other example related work sections, which is more regular and formalized than in other domain-independent and general summarization tasks. Last, evaluation of this specific type of summary is non-trivial and requires special evaluation metrics. While there are no existing studies on this specific problem, there are closely related endeavors.

A line of research focuses mainly on exploring domain-specific features for sum-

marizing scientific texts. Such domain-specific features have been proven very effective in inferring suitable strategies across summarization problems.

Early works deal with the summarization problem of abstract creation for technical documents. Probably the well-known and oft-cited paper is that of (Luhn, 1958) done at IBM in 1959. In this work, the author presented a method of automatically creating the abstracts. The core idea in this method is to use the occurrence frequency and distribution of particular words in input documents to rank sentences. Similar work to Luhn is (Baxendale, 1958), which introduced the feature of sentence position. To examine the importance of sentence position feature, the author conducted a small study by manually checking 200 paragraphs, finding that the topic sentences come as the first in 85% of the cases, and occur last in 7% paragraphs. This study implies that a naïve summarization approach could just select the first few sentences of each paragraph. Contrary to this, my manual analysis on Chapter 2/Section 2.3 revealed that the selection of a few first sentences to construct related work summaries is not effective, requiring other strategies to locate appropriate information for summarization.

Based on these works, Edmundson (1969) presents an automatic system to produce the extracts for technical documents. He built an extractive summarizer which uses a linear function to rank the importance of sentences. This weighted linear function combines different kinds of features. Adding to the two previous features from Luhn and Baxendale's works (Luhn, 1958; Baxendale, 1958), Edmundson introduced new features specific to technical documents. These include extracting clues that come from two structural sources – the body and the skeleton (*e.g.* title, headings and format) of the documents – and two linguistic sources – the presence of cue words using cue dictionary and a key glossary which include all keyword candidates whose total frequency exceeds a pre-defined threshold. This work explored the importance and effectiveness of structural information and heuristics-based evidences in summarizing abstracts of technical documents. Such information is helpful but need to be adapted for use in related work

summarization.

According to my understanding, the approaches proposed in three above works are extractive approaches which lack inter-sentential or structural discourse analysis, and would not be reliable in producing coherent abstracts. Ono et al. (1994) addressed this problem by presenting an advanced approach which leverages the implicit discourse structure to generate abstracts automatically. In that work, discourse structure is defined as the rhetorical structure which can be represented as the compound of rhetorical relations between sentences or paragraphs in texts. These rhetorical relations operate at two layers: intra-paragraph (represented based on units as sentences) and inter-paragraph (represented based on units as paragraphs). In sentences, the rhetorical relations can be extracted in accordance with the respective connective expression. For example, consider the sentence “This approach works well because it operates on the news domain”. In this sentence, there exists a connective expression “because” which means the relation “reason”. Thus, the clause “it operates on the news domain” will be a reason for the previous clause “this approach works well”. Overall, there are totally 34 rhetorical relations manually defined in (Ono et al., 1994). Their approach, which uses a subsequent generation process, results in a high coverage rate of 74% of manually-judged key sentences, and demonstrates the effective use of rhetorical relations in identifying key sentences. However, this work did not provide guidance in detail how the rhetorical relations are defined. It is actually helpful for successors to adapt such rhetorical information to novel problems. Also, the evaluation of discourse-based approach should be compared with other non-discourse approaches to examine its effectiveness. Inspired from this idea, a possible strategy for related work summarization is to explore implicit rhetorical features specific to scientific articles in locating summary information.

Another study relies on specific domain on educational science to build summarizer. de la Chica et al. (2008) presented an extractive summarizer to construct contents for concepts within knowledge maps used for educational science. The summarizer uti-

lizes the explicit knowledge in educational science texts to infer features for summarization. In particular, it proposes new domain-specific features, including: the educational standards feature (measures the content relevance of a given sentence according to standards of educational science texts), the additional domain knowledge of human experts, and the gazetteer features (reflects the appearance of the geographical names in each sentence). These proposed features were proved effective in compared to baseline features (*e.g.* centroid, length, and sentence position), resulting better summarization performance.

Technical terms and their definitions now appear frequently in Wikipedia. Summarization of Wikipedia has become a research topic in its own right. For example, Ye et al. (2009) investigated an approach for summarizing definitions for Wikipedia articles. Unlike normal texts, Wikipedia texts usually contain some specific features: wiki concept links which are multi-word terms indicating important content units in sentences, or two structural features with outlines which refer to a hierarchical clustering of sub-topics assigned by authors, and infoboxes which tabulate the key properties about topics of wiki articles. Such specific features are then integrated into unique summarization framework to produce Wikipedia definitions.

In the context of related work summarization problem, I believe that there exists implicit features in scientific articles which are likely to infer effective strategies for summarization processes. My work in this thesis will work on how to explore such features.

Unlike the above works which focus mainly on surface features (*e.g.* sentence position, cue phrases, . . .) or rhetorical structure in summarizing scientific texts, another line of work utilizes citation texts.

A citation is one method by which authors tell readers that a certain material should be credited to another source. A dereferenced citation may lead to a bibliographic reference providing the necessary details to unambiguously locate its source (e.g. au-

thors, the title of the work, published date and conference details). A “citation text” is text that discusses the cited work. Different citation texts that cite the same paper may highlight different aspects of the cited work. An interesting application is to use such citation texts to construct a “citation summary”.

Nanba and Okumura (1999) report work on a system to support the creation of technical surveys. Given a database of multiple papers, the system firstly identifies the reference relationships between papers and the additional information derived from the description around the references. This reference information is classified into three reference types including: type B (the references mention other researchers’ theories or methods), type C (the references compare with relevant works and point out the proposed problems) and type O (other than types B and C). The classification is based on 160 manually-created heuristics rules built from cue phrases. Similarities and dissimilarities are detected among papers based on these reference fragments, and finally presented in an interactive tool.

Another study that utilizes and stresses the important roles of citation texts is (Elkiss et al., 2008). In this study, the authors argue that the summaries using citation texts can serve as a surrogate for the actual article in various circumstances. They also explored the issue of little overlap between citation summaries and abstracts. The citation summaries may provide more details on different aspects of the actual article than the abstracts do. This claim was evaluated using a proposed lexical similarity metric called cohesion between abstract and citing sentences or among citing sentences to quantify their correlation. Even though the data domain used in this study is limited on biomedical domain only, however, the result is very valuable for further research. However, the study did not explore the role of full text of articles and its relationship with citation texts or abstracts. In fact, this issue has not been explored in the literature so far. This thesis aims to examine the roles of full text of articles and abstracts in the context of related work summarization.

Recent studies have directly utilized citation texts to explore how they impact scientific document summarization. The first study approached **single** article summarization (Qazvinian and Radev, 2008). Given a target scientific article and its citation summary, a graph-based approach was proposed to produce the final summary of that article. Here, a citation summary is represented as a complete undirected weighted graph with nodes (sentences). Edges between nodes are weighted by tf-idf based cosine similarity of two corresponding nodes. A graph clustering method is performed to cluster the nodes of graph. Different sentence extraction strategies were applied to the clusters in the evaluation.

Further, Mei and Zhai (2008) also demonstrated the usefulness of the citation summary in single article summarization. Given a scientific article and its citation summary, this study focused on generating a summary to best reflect the article's most influential aspect. They termed such a summary an impact-based summary, where the task is to extract the salient sentences from the input which best reflects the citation summary.

Even though two above studies obtained promising results in improving summarization performance, there are still unexplored challenges. First, both of studies did not deal with redundancy – how to extract unique information, how to fuse overlapping information across sentences. This issue needs to be solved in order to reduce redundancy and succinctly capture the novelty of the input in the output summary. Also, given a target paper, an abstract gives perspectives of authors about that paper, whereas a citation summary gives perspective of other works to that paper. Both sources are useful for related work summarization. This perspective has yet to be explored in the literature. Finally, rhetorical features specific to scientific domain have not been explored. Only surface features in those studies were examined. I believe that scientific texts may contain more rhetorical features that are helpful for summarization and generation processes.

The above studies are the initial efforts on single article summarization towards the future research of "topic summarization", where a system takes an input specifying

a research topic and automatically generates a summary of prior, relevant works. This research problem is very challenging due to the complexity of the task. Along these lines the iOPENER project has been initiated by leading researchers at the University of Michigan and the University of Maryland since 2008. This project initially investigates robust methods towards automatic generation of technical surveys given a set of articles. The ultimate goal of such generated technical surveys is to help readers understand large amounts of technical materials in the research literature as quickly as possible.

The first results from this project is (Mohammad et al., 2009). The authors re-examined some state-of-the-art generic multi-document summarization algorithms applied to the creation of the technical surveys. The key contribution here was in exploring the various methods – citation summaries, abstracts and full text – that could be employed to create technical surveys. To explore the structure of a technical survey, they conducted a manual analysis of chapter notes in technical books, which are prototypical examples of an actual technical survey. The analysis revealed that this structure is created from a set of rhetorical patterns: introductory statement, definitional follow up, elaboration of definition, deeper elaboration, contrasting definition, historical background and references to other prior works (Mohammad et al., 2009). The last pattern, on the other hand, accounts for the citation texts. Initially, this work took a first step on using this pattern towards the complete use of all patterns in generating a technical survey. In fact, the structure of an actual technical survey is much more complicated. Future investigation on this issue poses an interesting research problem.

Further, unlike (Qazvinian and Radev, 2008; Mei and Zhai, 2008) which target the problem of single article summarization, Mohammad et al. (2009) examined the problem of **multiple** article summarization. Various experiments show that the use of citation texts and abstracts in such context are very effective as compared to the articles' full text. Citation texts and abstracts may contain useful information that is not available in the full text of articles. However, the use of combination of both citation texts and

abstracts for summarization has not been explored. They may have some overlap in their content, and each of them may contain additional information that is not included in the remaining one. Also, I note that the evaluation in this study was limited to computational linguistics, so an extended evaluation over a wider set of domains is warranted.

In the work of (Mohammad et al., 2009), the output summary of a paper includes single sentences which does not express their full meanings. In this case, the contextual sentences (called background information) can help provide additional useful evidence which help readers quickly understand major contributions of that paper. Qazvinian and Radev (2010) examine the problem of automatically identifying such background information. To extend the work of (Mohammad et al., 2009), this work tries to use this background information in creating technical surveys and showed that such summaries have higher quality, compared to using citation summaries alone, in both automatic and human evaluation.

Such citation information may have great potential in other research domains, for example in mining the bioscience literature. Schwartz and Hearst (2006) utilized citation summaries to summarize key concepts and entities in bioscience, arguing that citation sentences may contain more informative and important contributions of a paper than its original abstract.

These works all center on the role of citations and their contexts in creating a summary, using citation information to rank content for extraction. However, they did not study the rhetorical structure of the intended summaries, targeting more on deriving useful content. Moreover, in the case that the citation summaries are unavailable, these approaches cannot work. My work takes advantage of full text of articles and explore their rhetorical structure, making the summarization problem solvable.

For work along this vein, I turn to studies on the rhetorical structure of scientific articles. Perhaps the most relevant is work by Teufel (1999); Teufel and Moens (2002); Merity et al. (2009) who defined and studied the argumentative zoning of texts, especially

ones in computational linguistics.

Notation	Category	Description
AIM	AIM	Statement of research goal.
BKG	BACKGROUND	Description of generally accepted background knowledge.
BAS	BASIS	Existing knowledge claim provides basis for new knowledge claim.
CTR	CONTRAST	An existing knowledge claim is contrasted, compared, or presented as weak.
OTH	OTHER	Description of existing knowledge claim.
OWN	OWN	Description of any other aspect of new knowledge claim.
TXT	TEXTUAL	Indication of papers textual structure.

Table 3.1: AZ-I rhetorical annotation scheme defined in (Teufel, 1999; Teufel and Moens, 2002).

First, they did an annotation analysis on a set of computational linguistics articles to assign what they term as "rhetorical status" for each sentence in the texts. They defined the task of argumentative zoning (AZ), which is the text classification of rhetorical status per sentence. The different types of rhetorical status express different communicative functions of each sentence with respect to the context of the whole article. Table 3.1 shows their rhetorical annotation scheme (called AZ-1) which is comprised of rhetorical labels and their descriptions. Consider the following example sentences:

- Paraphrases are alternative ways of conveying the same information. *(rhetorical status: BKG)*
- The remainder of this paper is as follows: Section 2 contrasts our method for extracting paraphrases with the monolingual case, and describes how we rank the extracted paraphrases with a probability assignment. *(rhetorical status: TXT)*
- In this paper we introduce a novel method for extracting paraphrases that uses bilingual parallel corpora. *(rhetorical status: AIM)*

Scientific research articles are main sources of information for researchers to learn about current cutting-edge technologies. Different from news articles of which structure

usually happens in time-linear manner, the structure of scientific research articles expresses the intellectual work conducted within a certain time period, focusing on problem bias and scientific argumentation. Some scientific articles are problem-biased because they describe the author's work from their own viewpoint and try to convince the reader the validity of a given work. Other articles are argumentative, discussing others' works in an objective manner, revealing advantages and disadvantages of a given approach. Thus, the structure designed for scientific research articles requires a specific rhetorical and argumentative analysis. Previous works presented in (Teufel, 1999; Teufel and Moens, 2002) took on the first effort in the construction of important concepts for the rhetorical analysis at the sentence level towards a complete meta-discourse analysis at document level for analyzing scientific research articles.

Recent work (Teufel et al., 2009) has extended these previous analyses for the domain in chemistry, expanding the original seven classes, as shown in Table 3.2. As can be seen, rhetorical status label *OWN* in AZ-I is extended to three different rhetorical status labels *OWN_METHOD, OWN_FAIL, and OWN_RES* to elaborate aspects about own work (*OWN* label) in more detailed manner, better suiting the styles demonstrated in chemistry publications. Even though the above argumentative zoning schemes are still underway, such efforts are promising to take further steps towards independent discipline for argumentative zoning in analyzing scientific texts.

While these studies studied the structure of an entire article, it is clear from their study that a related work section would contain general background knowledge (BACKGROUND zone) as well as specific information credited to others (OTHER and BASIS zones). This vein of work has been followed by Angrosh et al. (2010) which proposed the rhetorical classification scheme for the roles of each sentence in related work sections.

Recently, Jaidka et al. (2010) also present the beginnings of a corpus study of literature reviews, where they differentiate integrative and descriptive strategies in pre-

Category	Description	Category	Description
AIM	Statement of research goal or hypothesis of current paper	OWN_CONC	Findings, conclusions (non-measurable) of own work
NOV_ADV	Novelty or advantage of own approach	CODI	Comparison, contrast, difference to other solution (neutral)
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)	GAP_WEAK	Lack of solution in field, problem with other solutions
OTHR	Knowledge claim (significant for paper) held by somebody else. Neutral description	ANTISUPP	Clash with somebody else's results or theory; superiority of own work
PREV_OWN	Knowledge claim (significant) held by authors in a previous paper. Neutral description.	SUPPORT	Other work supports current work or is supported by current work
OWN_METHD	New knowledge claim, own work: methods	USE	Other work is used in own work
OWN_FAIL	A solution/method/experiment/ in the paper that did not work	FUT	Statements/suggestions about future work (own or general)
OWN_RES	Measurable/objective outcome of own work		

Table 3.2: AZ-II rhetorical annotation scheme defined in (Teufel et al., 2009).

senting discourse work. I see my differentiation between general and detailed topics in a topic tree (as discussed in Chapter 4/Section 4.2) as a natural parallel to their notion of integrative and descriptive strategies.

Further, the task of related work summarization is topic-biased, multi-document summarization problem that takes in a set of keywords arranged in a hierarchical fashion that describes topics of interest. Despite the bulk of previous works that addressed the topic-biased summarization problem for news texts, there exists no work for scientific texts.

In my task, a topic hierarchy tree is a bit similar to two previous studies (Branavan et al., 2007; Sauper and Barzilay, 2009). Sauper and Barzilay (2009) addressed the problem of automatically generating the summaries according to structural topic information given. The structural topic information differs from a topic hierarchy tree in terms of the depth of topic tree. Their tree is non-hierarchical. Meanwhile, Branavan et al. (2007) presented a problem that given the hierarchical segmentation of a text, the task is to automatically generate a table-of-contents for that tree with the desired length. Contrary to my proposed problem, given a topic hierarchy tree, I want to generate a text summary of

related works driven by that tree. Another concern is the position of topic nodes in tree. In particular, related work summarizer may treat leaf and intermediate nodes of topic tree in different ways in selecting appropriate information for summarization.

Chapter 4

Proposed System

The goal of this chapter is to develop a fully automatic system for RW summarization. Such a fully automatic system requires an input of multiple articles (*e.g.*, conference/journal papers) and a desired summary length. The system I develop here implicitly tries to organize the summary information following a hierarchy of topics. As discussed in Chapter 2, automatic generation of such hierarchy of topics is non-trivial, beyond the scope of this thesis. Thus, I alleviate the problem by providing a topic hierarchy tree as an additional input for the system. As a result, the semi-automatic system takes the input of multiple articles (*e.g.*, conference/journal papers), summary length and additionally a topic hierarchy tree.

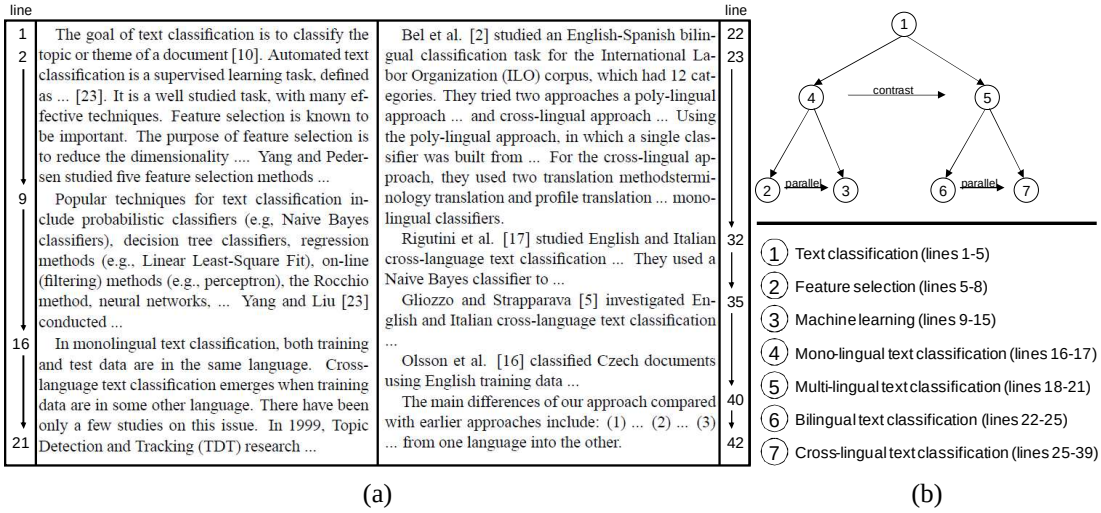
The proposed system is comprised of two main modules: 1) **Content Selection** and 2) **Generation**. The Content Selection module aims to extract all possible information at various granularity levels (e.g. words, phrases, sentences) relevant to a given topic hierarchy tree (THT). The Generation module then organizes the extracted information from Content Selection into a final comprehensive summary. Sections 4.2 and 4.3 discusses the Content Selection module whereas Section 4.4 describes the Generation module.

4.1 Problem Formulation

My problem formulation is based on the characteristics of RW summaries as well as problems and challenges discussed in Chapter 2. Given multiple articles (*e.g.*, conference or journal papers) as input, and a set of keywords in a hierarchical fashion that describes a target paper's topics of interest, a RW summarization system is expected to create a topic-biased summary of RW specific to the target paper. I assume that all input articles may share relevant topics which help to summarize the RW summaries. Note that I do not consider any structural information of the input articles (e.g. Title, Abstract, Introduction, Body, Conclusion) because such information makes data preprocessing step complicated. Moreover, the earlier discussion (Chapter 2/Section 2.3.2) hypothesizes that information to be extracted may not appear in any fixed section. Again, a topic hierarchy tree is very important and compulsory for RW summarization because it guides the summarizer to which relevant information is required to be summarized. Each node in the tree provides an associated set of keywords (e.g. words, phrases). The depth of the topic hierarchy tree may be varied depending on users' needs. According to my observation on **RWSData**, the maximum value for the tree depth is around 2 or 3. In fact, the content of a RW summary is strongly affected by the information provided in topic hierarchy tree. Basically, the topic hierarchy tree can be generated by employing hierarchical topic modeling algorithms like (Blei et al., 2004). However, the scientific domain may cause some unexpected problems which make using topic modeling may be non-trivial and complicated. Thus, I alleviate this problem by making a reasonable assumption that a topic hierarchy tree is provided in the input.

It turns out that my proposed problem has some novel specific characteristics that are not explored before. To start approaching it, some motivated questions should be considered as follows:

- How the structure of RW summary can be used to deduce the future approach for RW summarization?



(a) (b)

Figure 4.1: a) A RW summary extracted from (Wu and Oard, 2008); b) An associated topic hierarchy tree of a).

- How are the generated RW summaries ensured to maximize the text coverage and coherence with respect to the input topic hierarchy tree?
- How to generate the RW summaries that look like human-written ones?

4.2 Rhetorical Analysis on RW Summaries

I first extend the work on rhetorical analysis, concentrating on RW summaries. By studying examples in detail, I gain insight on how to approach RW summarization. I focus on a concrete RW summary example for illustration, an excerpt of which is shown in Figure 4.1a. Focusing on the argumentative progression of the text, I note the flow through different topics is hierarchical and can be represented as a topic tree as in Figure 4.1b.

This summary provides background knowledge for a paper on text classification, which is the root of the topic tree (node 1; lines 1–5). Two topics (“feature selection” and “machine learning”) are then presented in parallel (nodes 2 & 3; lines 5–8 & 9–15),

where specific details on relevant works are selected to describe two topics. These two topics are implicitly understood as subtopics of a more general topic, namely “monolingual text classification” (node 4; lines 16–17). The authors use the monolingual topic to contrast it with the subsequent subtopic “multi-lingual text classification” (node 5; lines 18–21). This topic is described by elaborating its details through two sub-topics: “bilingual text classification” and “cross-lingual text classification” (nodes 6 & 7; lines 22–25 & 25–39) where again, various example works are described and cited. The authors then conclude by contrasting their proposed approach with the introduced relevant approaches (lines 40–42).

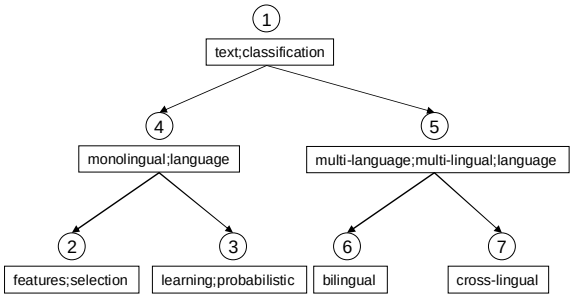


Figure 4.2: An associated topic tree of RW summary in Figure 4.1a, annotated with key words/phrases.

This summary illustrates three important points. First, the topic tree is an essential input to the summarization process. The topic tree can be thought of as a high-level rhetorical structure for which a process then attaches content. While it is certainly non-trivial to build such a tree, modifications to hierarchical topic modeling (Blei et al., 2004) or keyphrase extraction algorithms (Witten et al., 1999) I believe can be used to induce a suitable form. A resulting topic hierarchy from such a process would provide an associated set of key words or phrases that would describe the node, as shown in Figure 4.2.

Second, while summaries can be structured in many ways, they can be viewed as moves along the topic hierarchy tree. In the example, nodes 2 and 3 are discussed before their parent, as the parent node (node 4) serves as a useful contrast to introduce its sibling

(node 5). I find variants of depth-first traversal common, but breadth-first traversals of nodes with multiple descendants are more rare. They may be structured this way to ease the reader's burden on memory and attention. This is in line with other summary genres where information is ordered by high-level logical considerations that place macro level constraints (Barzilay et al., 2002).

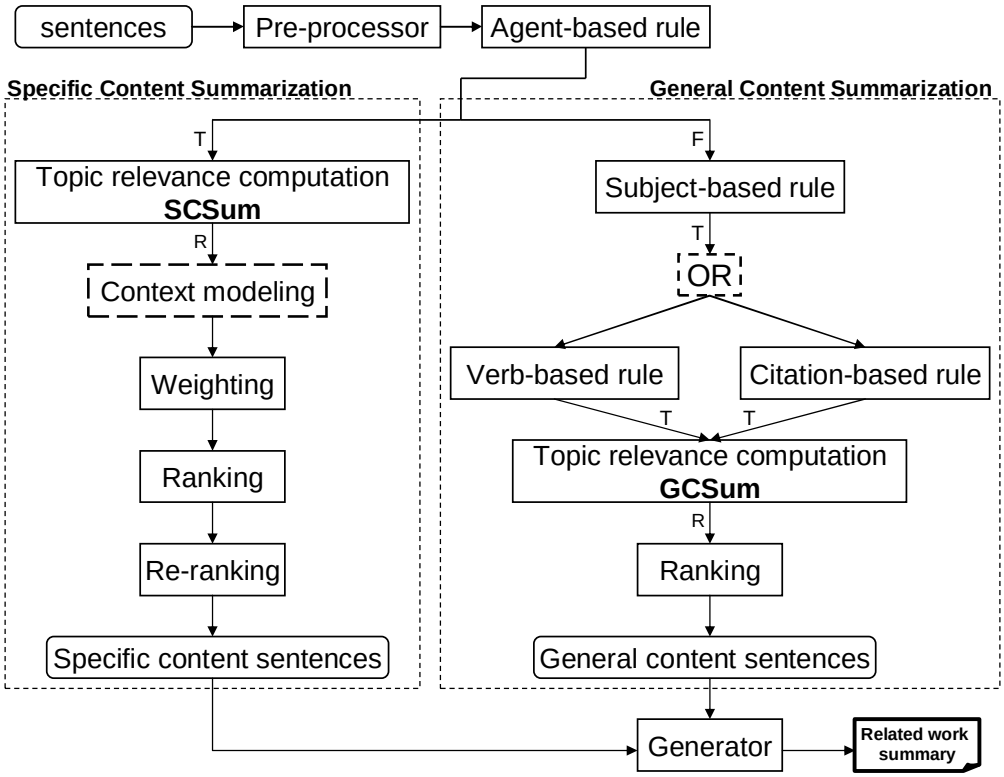


Figure 4.3: **The ReWoS architecture**. Decision edges are labeled as T (**T**rue), F (**F**alse) or R (**R**elevant).

Third, there is a clear distinction between sentences that describe a general topic and those that describe work in detail. Generic topics are often represented by background information, which is not tied to a particular prior work. These include definitions or descriptions of a topic's purpose. In contrast, detailed information forms the bulk of the summary and often describes key related work that is attributable to specific authors.

4.3 ReWoS: paired general and specific summarization

Motivated by the above observations, I propose a novel strategy for RW summarization with respect to a given topic tree.

I posit that sentences within a RW section come about by means of two separate processes – a process that gives general background information and a process that describes specific author contributions. A key realization in my work is that these two processes are easily mapped to the topic tree topologically: general content is described in internal topic nodes of the tree, whereas leaf nodes contribute detailed specifics. In my approach, these two processes are independent, and combined to construct the final summary.

I have implemented my idea in **ReWoS** (**Re**lated **Wo**rk **S**ummarizer), whose general architecture is shown in Figure 4.3. **ReWoS** is a pipeline system that features three modules: a General Content Summarization (GCSum), a Specific Content Summarization (SCSum), and a Generation.

Before discussing the modules, note that in the top of Figure 4.3, the input sentences (*i.e.*, the set of sentences from each related/cited article) are first preprocessed and subjected to an agent-based rule. The preprocessing removes redundant sentences, based on heuristic rules of sentence length and lexical clues. For example, sentences of which token-based length is too short (< 7) or too long (> 80), sentences referring to future tenses, or sentences containing obviously redundant clues such as: “in the section ...”, “figure XXX shows ...”, “for instance”. Lowercase and stemming for sentences are also performed.

The agent-based rule attempts to distinguish whether the sentence describes an author’s own work or not. **ReWoS** looks for the presence of tokens that signals work done by the author, such as “we”, “our”, “us”, “this approach”, and “this method”. I compiled a list of such 30 tokens (see details in Appendix A.1). For example, the following sentences contain tokens which are identified by the agent-based rule:

- Sentence 1: *the goal of customer satisfaction studies in business intelligence is to discover opinions about a company ' s products , features , services , and businesses .*

- Sentence 2: ***we** present a prototype system , code-named pulse , for mining topics and sentiment orientation jointly from free text customer feedback .*

Sentences that are marked with such tokens are routed for Specific Content Summarization (such as Sentence 2); sentences without such tokens are routed for General Content Summarization (such as Sentence 1).

4.3.1 General Content Summarization

The objective of general content summarization (GCSum) is to extract sentences containing useful background information on the topics of the internal node in focus. Note that since general content sentences do not specifically describe work done by the authors, I only take sentences that do not have the author-as-agent as input.

I divide such general content sentences into two groups: indicative and informative. Informative sentences give detail on a specific aspect of the problem. They often give definitions, purpose or application of the topic, for examples:

- *Text classification is a task that assigns a certain number of pre-defined labels for a given text.*

- *Statistical machine translation (SMT) seeks to develop mathematical models of the translation process whose parameters can be automatically estimated from a parallel corpus.*

- *the goal of answer selection is to choose from a pool of answer candidates the most likely answer for a question.*

In contrast, indicative sentences are simpler, inserted to make the topic transition explicit and rhetorically sound, for examples:

- *Many previous studies have approached monolingual text classification.*
- *This section reviews past methods for paraphrase evaluation.*
- *Sentiment analysis has been studied by many researchers recently.*

Indicative sentences can be easily generated by templates, as the primary information that is transmitted is the identity of the topic itself. Informative sentences, on the other hand, are better extracted from the source articles themselves, requiring a specific strategy. As informative sentences contain more content, my strategy with GCSum is to attempt to locate informative sentences to describe the internal nodes, failing which GCSum falls back to using predefined templates to generate an indicative placeholder.

To implement GCSum's informative extractor, I use a set of heuristics in a decision tree to first filter inappropriate sentences (as shown on the RHS of Figure 4.3). Remaining candidates (if any) are then ranked by a topic relevance computation, of which the top n high-score sentences are selected for the topic.

This heuristic cascade's purpose is to remove sentences that do not suit the syntactic structure of commonly-observed informative sentences. A useful informative sentence should discuss the topic directly; so GCSum first checks the subject of each candidate sentence, filtering sentences whose subject do not contain at least one topic key word/phrase. I also observe that informative background sentences often feature specific verbs or citations. GCSum thus also checks whether stock verb phrases (*i.e.*, "based on", "make use of" and 23 other patterns, listed comprehensively in Appendix A.2) are used as the main verb. Otherwise, GCSum checks for the presence of at least one citation – general sentences may list a set of citations as examples. In this case, the regular expression based citation recognition in texts is performed (see details in Appendix A.3). If both the cue verb and citation checks fail, the sentence is filtered out. Sentences that

remain are plausible candidates for extraction in GCSum and need to be ranked for their fitness for the summary.

GCSum's topic relevance computation ranks sentences based on keyword content. Specifically, I state that the topic of an internal node is affected by its surrounding nodes – ancestor, descendants and siblings. Based on this idea, the score of a sentence is computed in a discriminative way using the following linear combination:

$$score_S \rightarrow score_S^{QA} + score_S^{Q} - score_S^{QR} \tag{4.1}$$

where $score_S$ is the final relevance score, and $score_S^{QA}$, $score_S^{Q}$, and $score_S^{QR}$ mean the component relevance score of the sentence S with respect to the ancestor, current or other remaining nodes, respectively. I give positive credit to a sentence that contains keywords from an ancestor node, but penalize sentences with keywords from other topics (as such sentences would be better descriptors for those other topics).

To obtain each component relevance score, I employ TF×ISF relevance computation (Otterbacher et al., 2005). Term Frequency × Inverse Sentence Frequency (TF×ISF) is simply a sentence-level variation of TF×IDF:

$$\begin{aligned} score_S^{Q} &= \frac{rel(S,Q)}{\sum_{Q'} rel(S,Q'))} \\ &= \frac{\sum_{w \in Q} \log(tf_w^S + 1) \times \log(tf_w^Q + 1) \times isf_w}{Norm} \end{aligned} \tag{4.2}$$

where $rel(S,Q)$ is the relevance of S with respect to topic Q, $Norm$ is a normalization factor of $rel(S,Q)$ over all input sentences, tf_w^S and tf_w^Q are the term frequencies of token w within the sentence S or sentences that discuss topic Q, respectively. isf_w is the inverse sentence frequency of token w computed by $\log\left(\frac{1+N}{0.5+sf_w}\right)$, where sf_w is the sentence frequency of token w over all input sentences.

4.3.2 Specific Content Summarization

Sentences that are marked with author-as-agent are input to the Specific Content Summarization (SCSum) module. SCSum aims to extract sentences that contain detailed information about a specific author's work that is relevant to the input leaf nodes' topic.

SCSum starts by computing the topic relevance of each candidate sentence as shown in Equation (4.3). This process is identical to the Topic Relevance Computation step in the GCSum module, except that the term $score_S^{QR}$ in Equation (4.1) is replaced by $score_S^{QS}$, which is the relevance of the input sentence S with respect to its sibling nodes. I hypothesize that given a leaf node, sibling node topics may have an even more pronounced negative effect than other remaining nodes in the topic tree.

$$score_S \rightarrow score_S^{QA} + score_S^{Q} - score_S^{QS} \tag{4.3}$$

4.3.2.1 Context Modeling

I note that single sentences occasionally do not contain enough context to clearly express the idea mentioned in original articles. While agent-based sentences often introduce concepts, the pertinent details often are described later. Extracting just the agent-based sentence may incompletely describe a concept and lead to false inferences. Consider the example in Figure 4.4. In this figure, Sentences 0-5 are an contiguous extract of a source article being summarized, where Sentence 0 is an identified agent-based sentence. Sentence 6 shows a RW section sentence from a citing article that describes the original article. It is clear that the citing description is composed of information taken not only from the agent-based sentence but its context in the following sentences as well.

From this observation, I also choose nearby sentences within a contextual window after the agent-based sentence to represent the topic. I set the contextual window to 5 and extract a maximum of 2 additional sentences. These additional sentences are chosen based on their relevance scores to that topic using Equation (4.3). Sentences with non-zero scores are then added as contexts of the anchor agent-based sentence. As a

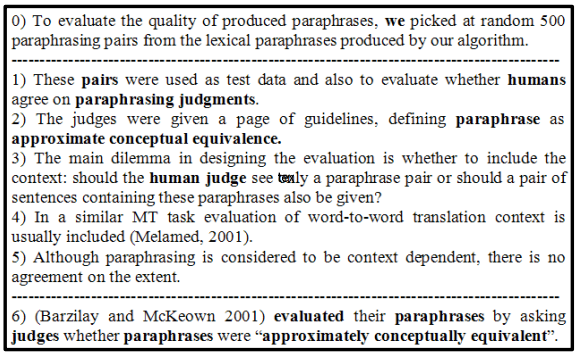


Figure 4.4: An example of agent-based sentence and its contexts.

result, some topics may contain only a single sentence, but others may be described by additional contextual sentences. Figure 4.5 shows an example of extracted RW summary using additional contextual sentences. As can be seen in the figure, some agent-based sentences can have two or one or none additional contextual sentences. For example, Sentences 1, 2, and 10 have two; Sentences 3, 5, and 6 have only one; and sentence 4 has none.

4.3.2.2 Weighting

The score of a candidate content sentence is computed from topic relevance computation (SCSum) that includes contributions for keywords present in the current, ancestor and sibling nodes. I observe that the presence of one or more of current, ancestor and sibling nodes may affect the final score from the computation. Thus, to partially address this, I add a new weighting coefficient for the score computed from the topic relevance computation (SCSum) (Equation (4.3)) as follows:

$$score^*_S = w_S^{QA,Q,QS} \times score_S \tag{4.4}$$

where: $w_S^{QA,Q,QS}$ is a weighting coefficient that takes on different values based on the presence of keywords in the sentence. Q, QA, and QS denote keywords from current,

Topic node: subjective manual evaluation
Ancestor keywords: paraphrase;evaluation;quality;
Node keywords: judge;judgement;assessment;human;subjective;manual;
1) [Bannard and Callison Burch 2005] had two native english speakers produce judgments as to whether the new sentences preserved the meaning of the original phrase and as to whether they remained grammatical . *paraphrases that were judged to preserve both meaning and grammaticality were considered to be correct , and examples which failed on either judgment were considered to be incorrect . the inter-annotator agreement for these judgements was measured at r . = 0.605 , which is conventionally interpreted as ? good ? agreement .*
2) because [Bannard and Callison Burch 2005] wanted to test their method independently of the quality of word alignment algorithms , they also developed a gold standard of word alignments for the set of phrases that they wanted to paraphrase . *they had two native english speakers produce judgments as to whether the new sentences preserved the meaning of the original phrase and as to whether they remained grammatical . paraphrases that were judged to preserve both meaning and grammaticality were considered to be correct , and examples which failed on either judgment were considered to be incorrect .*
3) the results of [Bannard and Callison Burch 2005] ' systems are not directly comparable , since barzilay and mckeown (2001) evaluated their paraphrases with a different set of criteria (they asked judges whether to judge paraphrases based on ? approximate conceptual equivalence ?) . *they evaluated their system with human judges who were asked whether the paraphrases were ? roughly interchangeable given the genre ? , scored an average of 41 % on a set of 130 paraphrases , with the judges all agreeing 75 % of the time , and a correlation of 0.66 .*
4) [Bannard and Callison Burch 2005] evaluate their paraphrase extraction and ranking methods using a set of manual word alignments , and contrast the quality with paraphrases extracted from automatic alignments .
5) [Barzilay and McKeown 2001] ' algorithm produced 9483 pairs of lexical paraphrases and 25 morpho-syntactic rules . *these pairs were used as test data and also to evaluate whether humans agree on paraphrasing judgments . the main dilemma in designing the evaluation is whether to include the context : should the human judge see only a paraphrase pair or should a pair of sentences containing these paraphrases also be given ?*
6) [Cohn et al. to appear] discuss how the corpus can be usefully employed in evaluating paraphrase systems automatically (e.g. , by measuring precision , recall , and f1) and also in developing linguistically rich paraphrase models based on syntactic structure . *the obtained paraphrases are typically evaluated via human judgments .*
7) [Ibrahim et al. 2003] ' results also show that judging structural paraphrases is a difficult task and inter-assessor agreement is rather low . *all of the evaluators agreed on the judgments (either positive or negative) only 75.4 % of the time . the average correlation constant of the judgments is only 0.66 .*
8) to evaluate the accuracy of [Ibrahim et al. 2003] ' results , 130 paraphrases were roughly interchangeable with each other , given the context of the genre . *their results also show that judging structural paraphrases is a difficult task and inter-assessor agreement is rather low . all of the evaluators agreed on the judgments (either positive or negative) only 75.4 % of the time .*
9) however , [Nenkova et al., 2007] explicitly aim at developing a metric for evaluating content selection , under the assumption that a separate linguistic quality evaluation of the summaries will be done as well . *the proposed characterization of optimal content is predictive : among summaries produced by humans , many seem equally good without having identical content .*
10) [Papineni et al., 2002] see that s2 is quite a bit better than s1 (by a mean opinion score difference of 0.326 on the 5-point scale) , while s3 is judged a little better (by 0.114) . *the high correlation coefficient of 0.99 indicates that bleu tracks human judgment well . particularly interesting is how well bleu distinguishes between s2 and s3 which they now take the worst system as a reference point and compare the bleu scores with the human judgment scores of the remaining systems relative to the worst system .*

Figure 4.5: **An example of extracted sentences with their contextual sentences according to a topic node**. Red-color marked and italic sentences are additional contextual ones.

ancestor and sibling nodes. If the sentence contains keywords from other sibling nodes, I assign a penalty of 0.1. Otherwise, I assign a weight of 1.0, 0.5, or 0.25, based on whether keywords are present from both the ancestor node(s) and current node, just the current node or just the ancestor nodes.

Given the above weighting, **ReWoS** ranks the sentences selected from the previous components for an input node. I select the top n sentences to represent the input leaf topic node. However, as the extracted sentences may contain redundant information, I employ the notion of Maximum Marginal Relevance – MMR (Goldstein and Carbonell,

1996) in the simplified form of SimRank (Li et al., 2008). SimRank only checks the similarity between extracted sentences without checking the topic relevance of sentences. A sentence X is removed if it has the maximum cosine similarity value exceeding a predefined threshold (0.75) with any sentence Y which is already chosen at previous steps of SimRank.

4.4 Generation

The extracted information from the two above summarization processes (general and specific content summarization) are inputted to the generation process. In fact, a full-fledged generation of natural texts for our task would be complex. In my **ReWoS** system, I generate the RW summaries by using depth-first traversals to form the ordering of topic nodes in a topic tree. For example, given a topic tree as shown in Figure 4.1b, the ordering of topic nodes in generating the summary is $1-4-2-3-5-6-7$.

As I discussed in Section 2.4, my manual analysis revealed that the Type 2 topic transitions along with citation realization patterns (*e.g.* P1, P2, C1) are sufficient for people to understand a RW summary. As such, each topic in topic tree is then represented by topic title which is provided in the input.

Furthermore, for each topic node, sentences within an input article are put together. Sentences with higher relevance scores are presented first. The order of referenced articles are sorted alphabetically. The summary length for each topic node is assigned equivalently in my experiment. Sample outputs to demonstrate our RW summary is shown in Appendix A.4.2 and A.4.3. Readers can refer to Appendix A.4.1 to compare automatically ReWoS-generated RW summaries with the ones generated by humans.

The final generation component post-processes the chosen sentences to improve fluency, by resolving abbreviations found in the sentences. This step first builds a look-up table, which has two entries corresponding to abbreviations and their descriptions.

The table is built by utilizing dependency relations from the Stanford statistical parser (de Marneffe and Manning, 2008).

Consider an example, a text fragment *Statistical Machine Translation (SMT)* has dependency relations such as: *abbrev(Translation, SMT)*, *nn(Translation, Machine)*, and *amod(Translation, Statistical)*. *SMT* is then recognized as an abbreviation of *Statistical Machine Translation*.

In summary, this chapter provides a detailed description on my initial prototype system (namely **ReWoS**) for the proposed task of **RW** Summarization. The analysis in Chapter 2 reveals that a related work summary is implicitly structured by a topic tree. Based on this, I formulated the **ReWoS** system which takes in a set of referenced articles, a summary length, and a manually-built topic tree as well. Also, inspired from the idea of the rhetorical analysis on human-written **RW** summaries, which differentiates between *internal* and *leaf* nodes of a topic tree in structuring *general* and *specific* summary content, I developed my **ReWoS** system including two separate processes: **General Content Summarization - GCSum** and **Specific Content Summarization - SCSum**. Each of them itself employs various heuristics-based strategies and computations to extract appropriate information. In addition to **GCSum** and **SCSum**, the **ReWoS** system also implements a **Generator** which in turn combines the outputs from **GCSum** and **SCSum**, arranges the summary content in a suitable fashion according to the topology of a input topic tree. The effectiveness of the **ReWoS** system will be assessed both in automatic and human evaluation, discussed in next chapter (Chapter 5).

Chapter 5

Evaluation

Previous chapter has discussed the details of the proposed ReWoS system developed for the task of RW summarization. This chapter aims to examine suitable methods for evaluation of generated RW summaries. At the first part of this chapter, I will present set-ups for the experiments and evaluation including selection of state-of-the-art baseline systems, automatic and human evaluation metric. The results and detailed analysis will conclude this chapter.

5.1 Evaluation & Experiment Set-up

I wish to assess the quality of the resulting ReWoS system as compared to state-of-the-art generic summarization systems. The assessment will follow up three following important criteria to gain the confidence:

- How to measure the quality and diversity of the generated summary content?
- How well the proposed ReWoS system benefits from topic hierarchy tree?
- Whether internal components of the proposed ReWoS system work well (*e.g.* context modeling)?

I first detail my baseline systems used for performance comparison, and defined evaluation measures specific to RW summary evaluation. In my evaluation, I use my manually compiled corpus – **RWSData** – as discussed earlier in Chapter 2/Section 2.1. I benchmark ReWoS against two baseline systems: LEAD and MEAD.

The LEAD baseline system represents each of the cited article with an equal number of sentences. The first n sentences are drawn from the article, meaning that the title and abstract are usually extracted. Simply, LEAD system constructs RW summaries by taking all those first sentences of each cited article with respect to the input summary length. The order of the article LEAD used in the resulting summary was determined by the order of articles to be processed. Basically, the LEAD system is said to be quite effective for newspaper summarization but is not sure to be still good for RW summarization. The results presented in next sections will validate this.

MEAD is a well-documented baseline extractive multi-document summarizer, developed in (Timothy et al., 2004; Radev et al., 2004). MEAD offers a set of different features that can be parameterized to create resulting summaries. I conducted an internal tuning of MEAD to maximize its performance on the RWSData dataset. The optimal configuration uses just two tuned features of *centroid* and *cosine similarity*. Note that neither baseline system utilizes the structure of topic hierarchy tree, which is central to my approach. In my experiments, I used the MEAD toolkit [1] to produce the summaries for LEAD and MEAD baseline systems.

Automatic evaluation was performed with ROUGE (Lin, 2004), a widely used and recognized automated summarization evaluation method. I employed a number of ROUGE variants, which have been proven to correlate with human judgments in multi-document summarization (Lin, 2004).

As discussed in Chapter 2/Section 2.5, automatic evaluation with ROUGE score suffers some unexpected problems that lead to inaccurate scoring of automatically-generated

[1] `http://www.summarization.com/mead/`

RW summaries in compared to golden RW summaries.

Since automatic evaluation ROUGE scores may not allow much introspection, I decide to investigate more on human evaluation. I conducted a human evaluation to assess more fine-grained qualities of my system. I asked 11 human judges to follow an evaluation guideline that I prepared, to evaluate the summary quality, consisting of the following evaluation measures:

Correctness: Is the summary content actually relevant to the hierarchical topics given?

Novelty: Does the summary introduce novel information that is significant in comparison with the human created summary?

Fluency: Does the summary's exposition flow well, in terms of syntax as well as discourse?

Usefulness: Is the summary acceptable in terms of its usefulness in supporting the researchers to quickly grasp the related works relevant to hierarchical topics given?

Each judge was asked to grade the four summaries according to the measures on a 5-point scale of 1 (very poor) to 5 (very good). Summaries 1 and 2 come from LEAD- and MEAD-based systems, respectively. Summaries 3 and 4 come from my proposed ReWoS systems, without (ReWoS−WCM) and with context modeling in SCSum (ReWoS−CM). All summarizers were set to yield a summary of the same length (1% of the original relevant articles, measured in sentences). Due to limited time, only 10 out of 20 evaluation sets were assessed by the evaluators. Each set was graded at least 3 times by 3 different evaluators; evaluators did not know the identities of the systems, which were randomized for each set examined.

System	ROUGE Recall Scores			
	ROUGE-1	ROUGE-2	ROUGE-S4	ROUGE-SU4
LEAD	0.501	0.096	0.116	0.181
MEAD	0.663	0.178	0.211	0.287
ReWoS−WCM	0.584	0.127	0.154	0.227
ReWoS−CM	0.698	0.183	0.218	0.298

Table 5.1: ROUGE-based automatic evaluation results for ReWoS variants and baselines.

5.2 Results

ROUGE results are summarized in Table 5.1. Surprisingly, the MEAD baseline system outperforms both LEAD baseline and ReWoS–WCM (without context modeling). Only ReWoS–CM (with context modeling) is significantly better than others, in terms of all ROUGE variants. I have some possible reasons to explain this phenomenon. First, ROUGE evaluation seems to work unreasonably when dealing with verbose summaries, often produced by MEAD. Second, RW summaries are multi-topic summaries of multi-article references. This may cause miscalculation from overlapping n-grams that occur across multiple topics or references. Chapter 2/Section 2.5.2 shows a typical example to validate this statement. Third, some RW summaries contain novel but correct information in comparing with gold summaries. This is not handled by ROUGE evaluation, which is just based on n-gram overlap. Moreover, gold summaries written by humans are not optimal summaries. Given a topic, people can compose different but still correct RW summaries.

Since automatic evaluation with ROUGE does not allow much introspection, I turn to my human evaluation. Results are summarized in Table 5.2. They show that both ReWoS–WCM and ReWoS–CM perform significantly better than baselines in terms of correctness, novelty, and usefulness. This is because my system utilized features developed specifically for related work summarization. Also, my proposed systems compare

System	Evaluation Measure			
	Correctness	Novelty	Fluency	Usefulness
LEAD	3.027	2.764	3.082	2.745
MEAD	3.009	3.109	2.591	2.700
ReWoS−WCM	3.618	3.391	3.391	3.609
ReWoS−CM	3.691	3.618	2.955	3.573

Table 5.2: Human evaluation results for ReWoS variants and baselines.

favorably with LEAD, showing that necessary information is not only located in titles or abstracts, but also in relevant portions of the research article body.

ReWoS–CM (with context modeling) performed equivalently to ReWoS–WCM (without it) in terms of correctness and usefulness. For novelty, ReWoS–CM is better than ReWoS–WCM. It showed that the proposed component of context modeling is useful in providing new information that is necessary for the RW summaries. For fluency, only ReWoS–CM is better than baseline systems. This is a negative result, but is not surprising because the summaries from the ReWoS–CM which uses context modeling seems to be longer than others. It makes the summaries quite hard to digest; some evaluators stated that they preferred the shorter summaries. An interesting extension in my future plan is that using information fusion techniques to fuse the contextual sentences with its anchor agentive sentence.

Note that both automatic and manual evaluation are not statistically significant due to the size of evaluation data (only tested on 10 evaluation sets). Thus, in the future, I would like to do my evaluation on a larger-scale basis.

A detailed error analysis of the results revealed that there are three main types of errors produced by my proposed systems. The first issue is in calculating topic relevance. In the context of related work summarization, my heuristics-based strategies for sentence extraction cannot capture fully this issue. Some sentences that have high rel-

evant scores to topics are not actually semantically relevant to the topics. The second problem of anaphoric expression is more addressable. Some extracted sentences still contain anaphoric expression (*e.g.*, "they", "these", "such", . . .), making final generated summaries incoherent. For example, a sentence (*[Papineni et al., 2002] present their method as an automated understudy to skilled* ***human judges*** *which substitutes for* ***them*** *when there is need for quick or frequent* ***evaluations***.) is relevant to the topic "human paraphrase evaluation" (keywords: human judges, evaluations) but not semantically relevant to it (first issue). Also, the word "them" referring to any entity presented earlier makes current sentence incoherent (second issue). The third issue is paraphrasing, where substituted paraphrases replace the original words and phrases in the source articles. For example, substituted paraphrase *judges* is used instead of the phrase *human assessors*.

In this chapter, I have tried both automatic and human evaluation methods for the task of RW summarization. Automatic evaluation with ROUGE scores has been proven to ineffective in assessing RW summaries, whereas human evaluation with four proposed measures is more accurate, but is an exhausted task, requiring much time and labour.

Chapter 6

Future Work

I envision that an expected fully automated related work summarization system should follow a pipeline framework as shown in Figure 6.1.

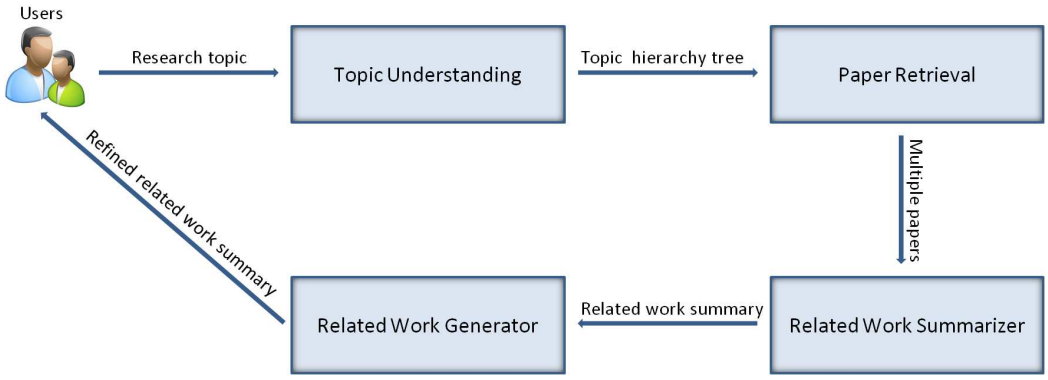


Figure 6.1: Expected framework for a fully automated related work summarization system

This system would work as follows. Given a research topic provided by users, a **Topic Understanding** module is responsible for exploring topic themes that implicitly reflect that topic. For example, given a research topic “text summarization”, two possible topics “multi-document summarization” and “single document summarization” should be recognized as sub-topics of the topic “text summarization”. The ultimate goal of this module is to provide topic themes under a hierarchical fashion, or also called a topic

hierarchy tree, for a **Paper Retrieval** module. Such a **Paper Retrieval** module would retrieve relevant papers that contain materials referring to a topic hierarchy tree provided by the **Topic Understanding** module. Both of the above modules may use the same resources for processing information. As a result, the outputs of two modules are a topic hierarchy tree and a set of relevant papers which are in turn provided to the **Related Work Summarizer and Generator** modules.

The **Related Work Summarizer** module aims to produce initial related work summaries which only contain raw information extracted from the input. The **Related Work Generator** then refines these initial summaries to produce the actual summaries which look like human-generated ones. To do this, a related work representation process is performed. Chapter 2/Section 2.4 shows in details what a representation process should do. Finally, the output is given back to users.

My initial prototype related work summarization system (namely **ReWoS**) developed in this thesis (as discussed in Chapter 4) has solved partially the pipeline framework of the expected system. The preliminary results show that the related work summaries produced by my system have better quality in terms of both automatic and human evaluation. However, my work shows that there is much room for additional improvement, for which I have outlined a few challenges that future research should pursue.

First, a shortcoming of my current system is that I assume that a topic hierarchy tree is given as input. It means that I ignore the **Topic Understanding** module in the development of my current system. Users are expected to provide such a topic hierarchy tree. I feel that this is an acceptable limitation because I feel existing techniques in topic modeling research will be able to create such input, and that the topic trees used in this study were quite simple. I plan to validate this by generating these topic trees automatically in my future work. Specifically, topic modeling research (Blei et al., 2010) is a good point to start.

Another shortcoming is that my prototype system takes the input with a set of re-

lated papers which is assumed to be provided by users. In this case, the **Paper Retrieval** module in the expected system is also ignored. In the future, I plan to automate this **Paper Retrieval** module.

The main focus of my initial system is on two modules **Related Work Summarizer and Generator**. The **Related Work Summarizer** module has been developed based on the idea using two different strategies (General and Specific Content Summarization) in locating the appropriate information for summarization process. The **Related Work Generator** module aims to refine the extracted information from the summarizer and produce the actual related work summaries. Though current system has obtained some promising result, there are still some open research problems which need more investigation.

First, I would like to develop a robust algorithm for automatic decomposition of related work summaries which current work in this thesis has not explored yet. Such an automatic decomposition will help create a golden corpus for related work summarization automatically.

As discussed earlier, the *context modeling* scheme included in the Related Work Summarizer module has been developed using a very simple strategy. Given an agent-based sentence, it just computes the topic relevancy of contextual sentences in a window size of 5 and then attach at most two additional sentences to that sentence. In the future, I plan to investigate a strategy that fuses contextual sentences with agent-based sentence to construct a new sentence. Such a process will condense the final summary but add more useful content into it. The research of sentence fusion in this case will have to handle the scientific domain which differs from news domain that most of previous works (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005) focused on.

In the **Related Work Generator** module as discussed in Chapter 4/Section 4.4, the related work representation I use is still simple. Only most popular simple patterns have been implemented in this module. I aim to investigate on more complex patterns to

better produce human-like final related work summaries.

Further, since human evaluation is an exhausted task, another interesting future work is to develop robust an automatic evaluation method specific to the task of RW summarization. Such a method will be expected to overcome problems of existing methods like ROUGE to better evaluate RW summaries. Chapter 2/Section 2.5 suggested two possible evaluation strategies that future work may work on.

Finally, I want to go towards practical applications that benefit from automated related work summarization research. For example, fully automated topic-biased related work summarization system integrated into scientific literature search (*e.g.* ACL Anthology search, DBLP search) is an extremely useful application for scholars who want to quickly understand an unfamiliar research topic.

Chapter 7

Conclusions

According to the best of my knowledge, the research of automated related work summarization has not been studied before. In this thesis, I have taken the initial steps towards solving the problem.

There are three main contributions in this thesis.

First, I constructed a new dataset (namely **RWSData**) specific to the task of RW summarization. This dataset is now publicly available for community use.

Second, I conducted a deep manual analysis on various aspects of related work summaries to identify their important characteristics in locating appropriate information for summarization and generation processes. Characteristics of RW summaries covered include definition, position, and topical structure. I also present some interesting problems in my analysis such as: the decomposition and alignment of RW summaries, RW representation, and observations on evaluation metrics. Such a manual analysis is very important and helpful for people who are interested in approaching the RW summarization problem.

Finally, I developed my initial prototype **Re**lated **Wo**rk **S**ummarization system, namely **ReWoS**, which creates its extractive summaries by dividing the task into general and specific content summarization processes for locating appropriate sentences for gen-

eral topics as well as detailed ones in a hierarchical fashion of a topic given. The proposed **ReWoS** system with two variants, with **ReWoS-CM** and without **ReWoS-WCM** context modeling worked well in compared to generic multi-document summarization baseline systems in human evaluation. Since the task of RW summarization is non-trivial, these results obtained in this thesis are very encouraging, pioneering an interesting research problem.

Exploring related work summarization comes at a timely moment, as scholars now have access to a preponderous amount of scholarly literature. Automated assistance in interpreting and organizing scholarly work will help build future applications for intelligent literature searching or integration with advanced digital libraries and reference management tools.

Bibliography

M. A. Angrosh, Stephen Cranefield, and Nigel Stanger. Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 293–302. ACM, 2010. ISBN 978-1-4503-0085-8. doi: http://doi.acm.org/10.1145/1816123.1816168.

Michele Banko and Lucy Vanderwende. Using n-grams to understand the nature of summaries. In *HLT-NAACL '04: Proceedings of HLT-NAACL 2004: Short Papers on XX*, pages 1–4, Morristown, NJ, USA, 2004. Association for Computational Linguistics. ISBN 1-932432-24-8.

Regina Barzilay. Modeling local coherence: An entity-based approach. In *In Proceedings of ACL 2005*, pages 141–148, 2005.

Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. volume 31, pages 297–328, Cambridge, MA, USA, 2005. MIT Press. doi: http://dx.doi.org/10.1162/089120105774321091.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. volume 17, pages 35–55, 2002.

P. B. Baxendale. Machine-made index for technical literature - an exper-

iment. *IBM Journal of Research Development*, 2(4):354–361, 1958. URL http://www.research.ibm.com/journal/rd/024/ibmrd0204L.pdf.

Shane Bergsma and Grzegorz Kondrak. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1083.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, page 2003, 2004.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. volume 57, pages 1–30, New York, NY, USA, 2010. ACM. doi: http://doi.acm.org/10.1145/1667053.1667056.

S. R. K. Branavan, Pawan Deshpande, and Regina Barzilay. Generating a table-of-contents. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 544–551, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1069.

Hakan Ceylan and Rada Mihalcea. The decomposition of human-written book summaries. In *CICLing '09: Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 582–593, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-00381-3. doi: http://dx.doi.org/10.1007/978-3-642-00382-0_47.

Dipanjan Das and Andr F.T. Martins. A survey on automatic text summarization. Technical report, Language Technologies Institute, Carnegie Mellon University., 2007.

Sebastian de la Chica, Faisal Ahmad, James H. Martin, and Tamara Sumner. Pedagogically useful extractive summaries for science education. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 177–184, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL http://www.aclweb.org/anthology/C08-1023.

Marie-Catherine de Marneffe and Christopher D. Manning. The stanford typed dependencies representation. In *In COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation.*, 2008.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P08/P08-1118.

Yuan Ding. A survey on multi-document summarization. Technical report, In fulfillment of the Written Preliminary Exam II Requirement, Department of Computer and Information Science, University of Pennsylvania, 2004. URL http://ydsite.googlepages.com/yding.wpe2.revised.pdf.

H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, 1969. URL http://eprints.kfupm.edu.sa/53107/1/53107.pdf.

Aaron Elkiss, Siwei Shen, Anthony Fader, Gunecs Erkan, David States, and Dragomir Radev. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of American Society of Information Science and Technology (JASIST)*, 59(1):51–62, 2008. ISSN 1532-2882. doi: http://dx.doi.org/10.1002/asi.v59:1.

Jade Goldstein and Jaime Carbonell. Summarization: (1) using mmr for diversity - based reranking and (2) evaluating summaries. In *Proceedings of a workshop on held at Baltimore, Maryland*, pages 181–195. ACL, 1996. doi: http://dx.doi.org/10.3115/1119089.1119120.

Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.

Masamichi Nishiwaki Hideki Tanaka, Tadashi Kumano and Takayuki Itoh. Analysis and modeling of manual summarization of japanese broadcast news. In *In Proceedings of Second International Joint Conference on Natural Language Processing (IJCNLP05)*, pages 49–54, 2005.

Eduard Hovy. Text summarisation. In Ruslan Mitkov, editor, *The Oxford Handbook of computational linguistics*, pages 583 – 598. Oxford University Press, 2003.

Hongyan Jing. Using hidden markov modeling to decompose human-written summaries. *Comput. Linguist.*, 28(4):527–543, 2002. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120102762671972.

Hongyan Jing and Kathleen R. McKeown. The decomposition of human-written summary sentences. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: http://doi.acm.org/10.1145/312624.312666.

Karen Sprck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449 – 1481, 2007. ISSN 0306-4573. doi: DOI:10.1016/j.ipm.2007.03.009.

Rodger Kibble and R. Power. Optimizing referential coherence in text generation. *Computational Linguistics*, 30 (4):pp. 401–416, 2004.

E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In *In Kees van Deemter and Rodger Kibble, editors, Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages pages 223–264, 2002.

Wenjie Li, Furu Wei, Qin Lu, and Yanxiang He. PNR2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In *Proceedings of (Coling 2008)*, pages 489–496, Manchester, UK, August 2008. URL `http://www.aclweb.org/anthology/C08-1062`.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. ACL. URL `www.law.kuleuven.ac.be/icri/conferences/Lin.pdf`.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958.

Inderjeet Mani. *Automatic Summarization*. John Benjamins, Amsterdam, 2001.

Erwin Marsi and Emiel Krahmer. Explorations in sentence fusion. In *In Proceedings of the 10th European Workshop on Natural Language Generation*, pages 109–117, 2005.

Qiaozhu Mei and ChengXiang Zhai. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio, June 2008. ACL. URL `http://www.aclweb.org/anthology/P/P08/P08-1093`.

Stephen Merity, Tara Murphy, and James R. Curran. Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26, Suntec City, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W09/W09-3603.

Rada Mihalcea and Hakan Ceylan. Explorations in automatic book summarization. In *Proceedings of EMNLP-CoNLL*, pages 380–389, Prague, Czech Republic, June 2007. ACL. URL http://www.aclweb.org/anthology/D/D07/D07-1040.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. Using citations to generate surveys of scientific paradigms. In *Proceedings of HLT-NAACL*, pages 584–592, Boulder, Colorado, June 2009. ACL. URL http://www.aclweb.org/anthology/N/N09/N09-1066.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *SIGKDD'08: Proceeding of the 14th SIGKDD*, pages 542–550. ACM, 2008. ISBN 978-1-60558-193-4. doi: http://doi.acm.org/10.1145/1401890.1401957.

Hidetsugu Nanba and Manabu Okumura. Towards multi-paper summarization using reference information. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 926–931, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-613-0. URL http://portal.acm.org/citation.cfm?id=687586.

Ani Nenkova and Kathleen McKeown. References to named entities: a corpus study. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages

70–72, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073483.1073507.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. volume 4, page 4, New York, NY, USA, 2007. ACM. doi: http://doi.acm.org/10.1145/1233912.1233913.

Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. In *Proceedings of the 15th conference on Computational linguistics*, pages 344–348, Morristown, NJ, USA, 1994. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/991886.991946.

Jahna Otterbacher, Güneş Erkan, and Dragomir R. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of HLT-EMNLP '05*, pages 915–922. ACL, 2005. doi: http://dx.doi.org/10.3115/1220575.1220690.

Karolina Owczarzak. Depeval(summ): Dependency-based evaluation for automatic summaries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 190–198, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P09/P09-1022`.

Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of Coling 2008*, pages 689–696, Manchester, UK, August 2008. URL `http://www.aclweb.org/anthology/C08-1087`.

Vahed Qazvinian and Dragomir R. Radev. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual*

Meeting of the Association for Computational Linguistics, pages 727–736, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1088.

Dragomir R. Radev, Hongyan Jing, Malgorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *IPM*, 40(6):919–938, 2004. ISSN 0306-4573. doi: 10.1016/j.ipm.2003.10.006. URL http://dx.doi.org/10.1016/j.ipm.2003.10.006.

Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P09/P09-1024.

Ariel S. Schwartz and Marti Hearst. Summarizing key concepts using citation sentences. In *Proceedings of BioNLP '06*, pages 134–135. ACL, 2006.

Simone Teufel. *Argumentative Zoning: Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh, 1999. URL http://www.cl.cam.ac.uk/users/sht25/az.html.

Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Journal of Computational Linguistics*, 28(4):409–445, 2002. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120102762671936.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Sin-

gapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D/D09/D09-1155.

Dragomir Radev Timothy, Timothy Allison, Sasha Blair-goldensohn, John Blitzer, Arda elebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Adam Winkel, and Zhu Zhang. Mead - a platform for multi-document multilingual text summarization. In *LREC 2004*, 2004. URL http://tangra.si.umich.edu/~radev/papers/lrec-mead04.pdf.

Wouter Weerkamp, Krisztian Balog, and Maarten de Rijke. A generative blog post retrieval model that uses query expansion based on external collections. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1057–1065, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P09/P09-1119.

Ian H. Witten, Gordon Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of Digital Libraries 99 (DL'99)*, pages 254–255. ACM Press, 1999.

Yejun Wu and Douglas W. Oard. Bilingual topic aspect classification with a few training examples. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 203–210, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: http://doi.acm.org/10.1145/1390334.1390371.

Shiren Ye, Tat-Seng Chua, and Jie LU. Summarizing definition from wikipedia. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages

199–207, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P09/P09-1023.

Appendix A

Appendix

A.1 Tokens used for the Agent-based Rules

- “this approach”, “this work”, “this article”, “this paper”, “this journal”, “this method”, “this survey”, “this model”, “this framework”, “this algorithm”
- “we”, “our”, “us”, “ours”, “ourselves”, “i”, “my”, “me”, “mine”, “myself”, “they”, “their”, “theirs”, “themselves”, “he”, “his”, “him”, “himself”, “she”, “her”, “hers”, “these”

A.2 Patterns for Stock Verb Phrases

The list of stock verb phrases is as follows: “based on”, “require”, “is to”, “make use of”, “applied in”, “used to”, “used in”, “aim to”, “aim at”, “suffer from”, “divided into”, “focused on”, “differ from”, “differ on”, “studied in”, “attract”, “receive”, “refer to”, “is that”, “include”, “related to”, “witnessed”, “is”, “has been”, “have been”.

A.3 Regular Expression for Recognizing Citations

As discussed in Chapter 2/Section 2.4.3, a citation can be represented in two single and multiple ways. A multiple way repeats single one many times. Also, a single citation itself has many variants, depending on authors' writing styles (*e.g.* (wilson and wiebe , 2001) or (wiebe et al. , 2001) or wiebe et al. (2001)). In this case, the use of regular expression is robust enough to handle such cases. I defined regular expression for citation recognition using five patterns as shown in Figure A.1.

Pattern 1 (for cases with brackets)
```
sin_pat_1 = "[a-z,\\-]+\\s?(and\\s[a-z,\\-]+|(et\\sal|et\\sal\\.))?\\s?(,|\\()?\\s[1-9][0-9][0-9][0-9][a-z]?\\s?(\\))?"
mul_pat_1 = "\\(\\s" + sin_pat_1 + "\\s" + "(;\\s" + sin_pat_1 + "\\s)*" + "\\s?\\)"
```

Pattern 2 (for cases with square brackets)
```
sin_pat_2 = “[a-z,\\-]+\\s?(and\\s[a-z,\\-]+|(et\\sal|et\\sal\\.))?\\s?(,|\\[)?\\s[1-9][0-9][0-9][0-9][a-z]?\\s?(\\])?”
mul_pat_2 = "\\[\\s" + sin_pat_2 + "\\s" + "(;\\s" + sin_pat_2 + "\\s)*" + "\\s?\\]"
```

Pattern 3 (for cases without any brackets)
```
pat_3 = “[a-z,\\-]+\\s?(and\\s[a-z,\\-]+|(et\\sal|et\\sal\\.))?\\s?\\(\\s[1-9][0-9][0-9][0-9][a-z]?\\s?\\)”
```

Pattern 4 (for cases using numbers only)
```
pat_4 = “\\[\\s[1-9][0-9]?[0-9]?\\s(,\\s[1-9][0-9]?[0-9]?\\s)*\\]”
```

Figure A.1: Regular expression based patterns for citation recognition.

A.4 Sample Outputs of RW Summary

Given the topic hierarchy tree as shown in the Figure 4.2 (in Chapter 4), a list of input referenced articles, and the summary length (set by 1% of the length of referenced articles measured by sentences), four systems (LEAD, MEAD, and two variants of ReWoS system) will produce the following RW summaries (note that the human-written RW summary is also provided for further references):

A.4.1 Human-written RW Summary

The goal of text classification is to classify the topic or theme of a document [10].

Automated text classification is a supervised learning task, defined as automatically assigning pre-defined category labels to documents [23].

It is a well studied task, with many effective techniques.

Feature selection is known to be important.

The purpose of feature selection is to reduce the dimensionality of the term space since high dimensionality may result in the overfitting of a classifier to the training data.

Yang and Pedersen studied five feature selection methods for aggressive dimensionality reduction: term selection based on document frequency (DF), information gain (IG), mutual information, a ?2 test (CIII), and term strength [24].

Using the kNN and Linear Least Squares Fit mapping (LLSF) techniques, they found IG and CIII most effective in aggressive term removal without losing categorization accuracy.

They also found that DF thresholding, the simplest method with the lowest cost in computation could reliably replace IG or CIII when the computations of those measure were expensive.

Popular techniques for text classification include probabilistic classifiers (e.g, Naive Bayes classifiers), decision tree classifiers, regression methods (e.g., Linear Least-Square Fit), on-line (filtering) methods (e.g., perceptron), the Rocchio method, neural networks, example-based classifiers (e.g., kNN), Support Vector Machines, Bayesian inference networks, genetic algorithms, and maximum entropymodelling [18].

Yang and Liu [23] conducted a controlled study of 5 well-known text classification methods: support vector machine (SVM), k-Nearest Neighbor (kNN), a neural network (NNet), Linear Least-Square Fit (LLSF) mapping, and Naive Bayes (NB).

Their results show that SVM, kNN, and LLSF significantly outperform NNet and NB when the number of positive training examples per category are small (fewer than 10).

In monolingual text classification, both training and test data are in the same language.

Cross-language text classification emerges when training data are in some other language.

There have been only a few studies on this issue.

In 1999, Topic Detection and Tracking (TDT) research was extended from English to Chinese [21]. In topic tracking, a system is given several (e.g., 1-4) initial seed documents and asked to monitor the incoming news stream for further documents on the same topic [4], the effectiveness of cross language classifiers (trained on Chinese data and tested on English) was worse than monolingual classifiers.

Bel et al. [2] studied an English-Spanish bilingual classification task for the International Labor Organization (ILO) corpus, which had 12 categories.

They tried two approaches a poly-lingual approach in which both English and Spanish training and test data were available, and cross-lingual approach in which training examples were available in one language.

Using the poly-lingual approach, in which a single classifier was built from a set of training documents in both languages, their Winnow classifier, which, like SVM, computes an optimal linear separator in the term space between positive and negative training examples, achieved F1 of 0.811, worse than their monolingual English classifier (with F1=0.865) but better than their monolingual Spanish classifier (with F1=0.790).

For the cross-lingual approach, they used two translation methodsterminology translation and profile translation.

When trained on English and tested on Spanish translated into English, their classifier achieved F1 of 0.792 using terminology translation and 0.724 using profile translation; when trained on Spanish and tested on pseudo-Spanish, their classifier achieved F1 of 0.618; all worse than their corresponding monolingual classifiers.

Rigutini et al. [17] studied English and Italian cross-language text classification in which training data were available in English and the documents to be classified were in Italian.

They used a Naive Bayes classifier to classify English and Italian newsgroups messages of three categories: Hardware, Auto and Sports.

English training data (1,000 messages for each category) were translated into Italian using Office Translator Idiomax.

Their cross-language classifier was created using Expectation Maximization (EM), with English training data (translated into Italian) used to initialize the EM iteration on the unlabeled Italian documents.

Once the Italian documents were labeled, these documents were used to train an Italian classifier.

The cross-language classifier performed slightly worse than monolingual classifier, probably due to the quality of their translated Italian data.

Gliozzo and Strapparava [5] investigated English and Italian cross-language text classification by using comparable corpora and bilingual dictionaries (MultiWordNet and the Collins English-Italian bilingual dictionary).

The comparable corpus was used for Latent Semantic Analysis which exploits the presence of common words among different languages in the term-by-document matrix to create a space in which documents in both languages were represented.

Their cross-language classifier, either trained on English and tested on Italian, or trained on Italian and tested on English, achieved an F1 of 0.88, worse than their monolingual classifier (with F1 =0.95 for English and 0.92 for Italian).

Olsson et al. [16] classified Czech documents using English training data.

They translated Czech document vectors into English document vectors using a probabilistic dictionary which contained conditional word-translation probabilities for 46,150 word translation pairs.

Their concept label kNN classifier (k = 20) achieved precision of 0.40, which is 73

The main differences of our approach compared with earlier approaches include: (1) classifying document segments into aspects, rather than documents into topics; (2) using few training examples from both languages; (3) using statistical machine translation results to map segment vectors from one language into the other.

A.4.2 Outputs from ReWoS system (with context modeling)

text classification

the automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years , due to the increased availability of documents in digital form and the ensuing need to organize them .

the essential ideas of the dia transforming the classification space by means of abstraction and using a more detailed text representation than the standard bag-of-words approach have not been taken up by other researchers so far .

monolingual text classification

using the same training set , monolingual english classification was run on four similarly partitioned test segments .

automatic text categorization systems based on supervised learning [16] can reach a similar accuracy , so that the (semi) automatic classification of monolingual documents is becoming standard practice .

feature selection

as [16] have training data only in english , they may translate all of the czech data features into english for classification (they refer to this as english sided classification) . alternatively , they may translate all english training features into czech , before classifying in czech . a vectors subscript denotes the language from which the term frequencies were originally drawn (e.g. , ee denotes a feature vector of english term frequencies that were drawn from an english document) .

the approach that [17] propose is based on two steps : first the training set available in the language l1 is translated into the target language l2 using an automatic translation system . the algorithm also requires a proper feature selection technique to avoid to converge to trivial solutions . for reason of simplicity , they reduce the multi-lingual case with k languages to k bi-lingual problems selecting one language as the principal one ; thus studying the bi-lingual case is not restrictive with respect to the multi-lingual problem .

[24] apply feature selection to documents in the preprocessing of knn and llsf . the effectiveness of a feature selection method is evaluated using the performance of knn and llsf on the preprocessed documents . before applying feature selection to documents , they removed the words in a standard stop word list [18] .

[24] use two classifiers which have already scaled to a target space with thousands or tens of thousands of categories . they seek answers to the following questions with empirical evidence : what are the strengths and weaknesses of existing feature selection methods applied to text categorization ? to what extend can feature selection improve the accuracy of a classifier ?

Classifiers

having attained a set of training vectors ee (via normal indexing) and testing vectors e . (via probabilistic word translation) , [16] are free to continue with classification as before in the monolingual case . the base of the probabilistic dictionary is taken from version 1.0 of the prague czech-english dependency treebank (pcedt) [4] , which contains conditional word-translation probabilities for 46,150 word translation pairs .

[16] here confine ourselves to english sided classification , although the concepts may naturally be extended (mutatis mutandis) to the czech and two sided approaches . the matrix e represents a probabilistic dictionary mapping between czech and english terms , such that the (they , j) element represents the probability that an english word ei is the translation of the czech word cj . having attained a set of training vectors ee (via normal indexing) and testing vectors e . (via probabilistic word translation) , they are free to continue with classification as before in the monolingual case .

in the 90s the approach of [18] has increasingly lost popularity (especially in the research community) in favor of the machine learning (ml) paradigm , according to which a general inductive process automatically builds an automatic text classifier by learning , from a set of preclassified documents , the characteristics of the categories of interest .

in all the cases [5] trained on the english part and they classified the italian part , and they trained on the italian and classified on the english part . each graph show the learning curves respectively using a bow kernel (that is considered here as a baseline) and the multilingual domain kernel . analyzing the learning curves , it is worth noting that when the quantity of training increases , the performance becomes better and better for the multilingual domain kernel , suggesting that with more available training it could be possible to improve the results .

multi-lingual text classification

multi-language text classification became an important task .

in this setting , the similarity among texts in different languages could be estimated by exploiting the classical vsm just described .

bilingual text classification

[2] ' translation resources were built using a corpus-driven approach , following a frequency criterion to include nouns , adjectives and verbs with a frequency higher than 30 occurrences in the bilingual lexicon .

in the paper of [5] they have shown that the problem of cross-language text categorization on comparable corpora is a feasible task . in particular , it is possible to deal with it even when no bilingual resources are available . on the other hand when it is possible to exploit bilingual repositories , such as a synset-aligned wordnet or a bilingual dictionary , the obtained performance is close to that achieved for the monolingual task .

in the work of [5] they present many solutions according to the availability of bilingual resources , and they show that it is possible to deal with the problem even when no such resources are accessible . in particular , when bilingual dictionaries are available the performance of the categorization gets close to that of monolingual text categorization .

however , the main disadvantage of the approach of [5] to estimate inter-lingual text similarity is that it strongly terion to decide whether two corpora are comparable is to estimate the percentage of terms in the intersection of their vocabularies . for languages with scarce resources a bilingual dictionary could be not easily available .

cross-lingual text classification

in cltc , [17] can imagine three different scenarios : poly-lingual training : a labeled training set is available for each language and one classifier is trained using training examples from all the different languages . cross-lingual training : the labeled training set is available for only one language and they have to use that to classify documents in other languages .

cross-lingual text categorization is actually easier than cross-lingual information retrieval , for the same reason that lemmatization and term normalization have much less effect in cltc than in clir : the law of large numbers is with [2] . they have found viable solutions for two extreme cases of cross-lingual text categorization , between which all practical cases can be situated . on the one hand they found that poly-lingual training , training one single classifier to classify documents in a number of languages , is the simplest approach to cross-lingual text categorization , provided that enough training examples are available in the respective languages (tens to hundreds) , and the classification algorithm used is immune to the evident disjointedness of the resulting class profile (as is the case for winnow but not for rocchio) .

in sections 5 and 6 [2] propose three different solutions for cross- language classification , implying increasingly smaller (and therefore less costly) translation tasks . when they embarked on this line of research , they did not find any publications addressing the area of cross-lingual text categorization as such . on the other hand , there is a rich literature addressing the related problem of cross-lingual information retrieval (clir) .

in clir , [2] need a relevance model for both the source language and the target language . cross-lingual text categorization (cltc) or cross-lingual classification is a new research subject , about which no previous literature appears to be available .

A.4.3 Outputs from ReWoS system (without context modeling)

text classification

the automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years , due to the increased availability of documents in digital form and the ensuing need to organize them .

the essential ideas of the dia transforming the classification space by means of abstraction and using a more detailed text representation than the standard bag-of-words approach have not been taken up by other researchers so far .

monolingual text classification

using the same training set , monolingual english classification was run on four similarly partitioned test segments .

automatic text categorization systems based on supervised learning [16] can reach a similar accuracy , so that the (semi) automatic classification of monolingual documents is becoming standard practice .

feature selection

as a result , [23] selected 1000 features for nnet , 2000 features for nb , 2415 features for knn and llsf , and 10000 features for svm . [23] applied statistical feature selection at a preprocessing stage for each classifier , using either a x2 statistic or information gain criterion to measure the word-category associations , and the predictiveness of words (features) .

the focus in the paper of [24] is the evaluation and comparison of feature selection methods in the reduction of a high dimensional feature space in text categorization problems .

to assess the effectiveness of feature selection methods [24] used two different m-ary classifiers , a knearest-neighbor classifier (knn) [23] and a regression method named the linear least squares fit mapping (llsf) [27] .

classifiers

having attained a set of training vectors ee (via normal indexing) and testing vectors e . (via probabilistic word translation) , [16] are free to continue with classification as before in the monolingual case .

the matrix e represents a probabilistic dictionary mapping between czech and english terms , such that the ([16] , j) element represents the probability that an english word ei is the translation of the czech word cj .

in the paper of [17] they propose a learning algorithm based on the em scheme which can be used to train text classifiers in a multilingual environment .

in the 90s the approach of [18] has increasingly lost popularity (especially in the research community) in favor of the machine learning (ml) paradigm , according to which a general inductive process automatically builds an automatic text classifier by learning , from a set of preclassified documents , the characteristics of the categories of interest .

multi-lingual text classification

multi-language text classification became an important task .

in the second step , a text classifier for the target language l2 is trained using the em algorithm to take advantage both of the labeled examples obtained from the original language l1 in the first step and of the set of unlabeled data in language l2 .

bilingual text classification

[16] ' goal in cross-language text classification (cltc) is to use english training data to classify czech documents (although the concepts presented here are applicable to any language pair) .

[2] ' translation resources were built using a corpus-driven approach , following a frequency criterion to include nouns , adjectives and verbs with a frequency higher than 30 occurrences in the bilingual lexicon .

in the work of [5] they present many solutions according to the availability of bilingual resources , and they show that it is possible to deal with the problem even when no such resources are accessible .

in [5] ' experiments they exploit two alternative multilingual resources : multiwordnet and the collins english-italian bilingual dictionary .

cross-lingual text classification

cross-lingual training : the labeled training set is available for only one language and [17] have to use that to classify documents in other languages .

cross-lingual text categorization is actually easier than cross-lingual information retrieval , for the same reason that lemmatization and term normalization have much less effect in cltc than in clir : the law of large numbers is with [2] .

on the one hand [2] found that poly-lingual training , training one single classifier to classify documents in a number of languages , is the simplest approach to cross-lingual text categorization , provided that enough training examples are available in the respective languages (tens to hundreds) , and the classification algorithm used is immune to the evident disjointedness of the resulting class profile (as is the case for winnow but not for rocchio) .

[2] describe practical and cost-effective solutions for automatic cross-lingual text categorization , both in case a sufficient number of training examples is available for each new language and in the case that for some language no training examples are available .

A.4.4 Outputs from LEAD system

[16] ' goal in cross-language text classification cltc is to use english training data to classify czech documents although the concepts presented here are applicable to any language pair .

cltc is an off-line problem , and the authors are unaware of any previous work in this area .

an em based training algorithm for cross-language text categorization .

due to the globalization on the web , many companies and institutions need to efficiently organize and search repositories containing multilingual documents .

the management of these heterogeneous text collections increases the costs significantly because experts of different languages are required to organize these collections .

cross-language text categorization can provide techniques to extend existing automatic classification systems in one language to new languages without requiring additional intervention of human experts .
the automated categorization or classification of texts into predefined categories has witnessed a booming interest in the last 10 years , due to the increased availability of documents in digital form and the ensuing need to organize them .
in the research community the dominant approach to this problem is based on machine learning techniques : a general inductive process automatically builds a classifier by learning , from a set of preclassified documents , the characteristics of the categories .
the advantages of the approach of [18] over the knowledge engineering approach consisting in the manual definition of a classifier by domain experts are a very good effectiveness , considerable savings in terms of expert labor power , and straightforward portability to different domains .
the article of [2] deals with the problem of cross-lingual text categorization cltc , which arises when documents in different languages must be classified according to the same classification tree .
[2] describe practical and cost-effective solutions for automatic cross-lingual text categorization , both in case a sufficient number of training examples is available for each new language and in the case that for some language no training examples are available .
topic detection and tracking tdt refers to automatic techniques for discovering , threading , and retrieving topically related material in streams of data .
the paper of [23] reports a controlled study with statistical significance tests on five text categorization methods : the support vector machines svm , a k-nearest neighbor knn classifier , a neural network nnet approach , the linear least-squares fit llsf mapping and a naive bayes nb classifier .
[23] focus on the robustness of these methods in dealing with a skewed category distribution , and their performance as function of the training-set category frequency .
a comparative study on feature selection in text categorization .
the paper of [24] is a comparative study of feature selection methods in statistical learning of text categorization .
[4] investigate important differences between two styles of document clustering in the context of topic detection and tracking .
converting a topic detection system into a topic tracking system exposes fundamental differences between these two tasks that are important to consider in both the design and the evaluation of tdt systems .
exploiting comparable corpora and bilingual dictionaries for cross-language text categorization .
cross-language text categorization is the task of assigning semantic classes to documents written in a target language e.g.
english while the system is trained using labeled documents in a source language e.g.

A.4.5 Outputs from MEAD system

[16] ' goal in cross-language text classification cltc is to use english training data to classify czech documents although the concepts presented here are applicable to any language pair .
the cltc task can be stated as follows : suppose [17] have a good classifier for a set of categories in a language l1 and a large amount of unlabeled data in a different language l2 ; how can they categorize this corpus according to the same categories defined for language l1 without having to manually label any data in l2 ?
in the second step , a text classifier for the target language l2 is trained using the em algorithm to take advantage both of the labeled examples obtained from the original language l1 in the first step and of the set of unlabeled data in language l2 .
cross-lingual training : the labeled training set is available for only one language and [17] have to use that to classify documents in other languages .

the proposed approach is based on the idea that [17] can use a known training set in one language to initialize the em iterations on an unlabeled set of documents written in a different language .

aside from [18] the automatic assignment of documents to a predefined set of categories , which is the main topic of their paper , the term has also been used to mean ii the automatic identification of such a set of categories e.g. , borko and bernick 1963 , or iii the automatic identification of such a set of categories and the grouping of documents under them e.g. , merkl 1998 , a task usually called text clustering , or iv any activity of placing text items into groups , a task that has thus both tc and text clustering as particular instances manning and sch utze 1999 .

other applications [18] do not explicitly discuss are speech categorization by means of a combination of speech recognition and tc myers et al. 2000 ; schapire and singer 2000 , multimedia document categorization through the analysis of textual captions sable and hatzivassiloglou 2000 , author identification for literary texts of unknown or disputed authorship forsyth 1999 , language identification for texts of unknown language cavnar and trenkle 1994 , automated identification of text genre kessler et al. 1997 , and automated essay grading larkey 1998 .

there are two distinct ways of viewing dr , depending on whether the task is performed locally i.e. , for each individual category or globally : local dr : for each category ci , a set t ' of terms , with it 'i ¡¡ iti , is chosen for classification under ci see apt e et al. 1994 ; lewis and ringuette 1994 ; li and jain 1998 ; ng et al. 1997 ; sable and hatzivassiloglou 2000 ; sch utze et al. 1995 , wiener et al. 1995 .

other more sophisticated information-theoretic functions have been used in the literature , among them the dia association factor fuhr et al. 1991 , chi-square caropreso et al. 2001 ; galavotti et al. 2000 ; sch utze et al. 1995 ; sebastiani et al. 2000 ; yang and pedersen 1997 ; yang and liu 1999 , ngl coefficient ng et al. 1997 ; ruiz and srinivasan 1999 , information gain caropreso et al. 2001 ; larkey 1998 ; lewis 1992a ; lewis and ringuette 1994 ; mladeni c 1998 ; moulinier and ganascia 1996 ; yang and pedersen 1997 , yang and liu 1999 , mutual information dumais et al. 1998 ; lam et al. 1997 ; larkey and croft 1996 ; lewis and ringuette 1994 ; li and jain 1998 ; moulinier et al. 1996 ; ruiz and srinivasan 1999 ; taira and haruno 1999 ; yang and pedersen 1997 , odds ratio caropreso et al. 2001 ; mladeni c 1998 ; ruiz and srinivasan 1999 , relevancy score wiener et al. 1995 , and gss coefficient galavotti et al. 2000 .

an interesting evaluation has been carried out by dumais et al. 1998 , who have compared five different learning methods along three different dimensions , namely , effectiveness , training efficiency i.e. , the average time it takes to build a classifier for category ci from a training set tr , and classification efficiency i.e. , the average time it takes to classify a new document dj under category ci .

[2] describe practical and cost-effective solutions for automatic cross-lingual text categorization , both in case a sufficient number of training examples is available for each new language and in the case that for some language no training examples are available .

automatic text categorization systems based on supervised learning 16 can reach a similar accuracy , so that the semi automatic classification of monolingual documents is becoming standard practice .

by means of a number of experiments , [2] shall test the following hypotheses : poly-lingual training : simultaneous training on labeled documents in languages a and b will allow they to classify both a and b documents with the same classifier cross-lingual training : a mono lingually trained classifier for language a plus a translation of the most important terms from language b to a allows to classify documents written in b. lessons from clir for cltc ?

rocchio is in all cases much worse than for monolingual classification .

on the one hand [2] found that poly-lingual training , training one single classifier to classify documents in a number of languages , is the simplest approach to cross-lingual text categorization , provided that enough training examples are available in the respective languages tens to hundreds , and the classification algorithm used is immune to the evident disjointedness of the resulting class profile as is the case for winnow but not for rocchio .

once again , [21] see that techniques work comparably well in monolingual tasks training and testing in the same language .

as in monolingual segmentation or tracking , monolingual detection results are reassuringly similar .

in particular , when bilingual dictionaries are available the performance of the categorization gets close to that of monolingual text categorization .

for instance the classical monolingual text categorization tc problem can be reformulated as a cross language text categorization cltc task , in which the system is trained using labeled examples in a source language e.g.

[5] can observe that the cltc results are quite close to the performance obtained in the monolingual classification tasks .

on the other hand when it is possible to exploit bilingual repositories , such as a synset-aligned wordnet or a bilingual dictionary , the obtained performance is close to that achieved for the monolingual task .