Topic Detection and Segmentation in Automatic Text Summarization*

Elena Lloret

December 13, 2009

1 Topic: Definition

A topic or a theme is what discourse, a discourse fragment or a sentence is about. It is the shortest summary of a discourse, the main proposition of a paragraph or what is commented on in a sentence. The term *topic* is usually defined as the aboutness of a unit of discourse [20].

1.1 Topic Structure

Topic structure within a text can be tackled from two different points of view: from a Document Level (*discourse topic*) or a Sentence Level perspective (*sentence topic*). If a text is considered as a whole, it usually talks about a single topic, but in a more deeper analysis, several sub-topics can be identified, giving additional information about the main topic. On the other hand, taking into account sentence structure, we can find that every sentence has a topic (the part of the structure that is being presenting) and a comment (what is being asserted about the topic). Next, these two-way perspectives are going to be explained in more detail.

1.1.1 Document Level

- **Main Topic and Topic Shifts**

  Normally, a text deals with a single main topic, which is developed through the rest of the document, exposing several subtopics as well. It is important to detect these parts of text where a change of topic occurs. This leads to a hierarchical organization of a text into topics and subtopics, topic concatenation, and semantic return [1].

- **Topic Sentences**

  Text documents are usually subdivided in different paragraphs. A paragraph can be defined as a coherent text segment focused on a single topic. Every paragraph needs a topic sentence. The topic sentence is usually the first sentence of the paragraph [13]. It gives the reader an idea of what the paragraph is going to be about.

1.1.2 Sentence Level

- **Topic-Focus**

  The dicotomy of topic and focus is relevant not only for a possible placement of the sentence in a context, but also for its semantic interpretation [9]. The *topic*, also known as theme, can be understood as the part of the sentence structure that is being presented by the speaker as readily available in the hearer's memory; in other words, it is the "given" information, while the *focus* (comment or rheme) reflects

aspects of the topic that adds new or unpredictable information. This leads to the concept of *thematic progression* (or *thematic structure*) [13], that is the way in which "new" information is conveyed in the context of what is already known or what has already been expressed. There are thus two basic types of thematic progression: (1) parallel, where the topic (theme, T) is constant, and (2) linear, where topics (themes) relate to preceding focuses (rhemes, R). Figures 1 and 2, show both types of thematic progression: parallel and linear, respectively.

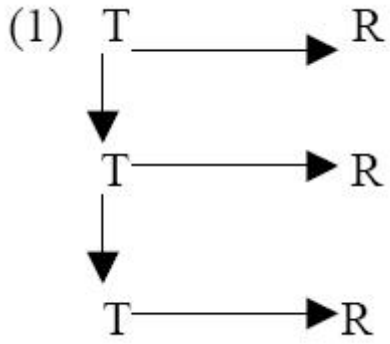


Figure 1: Parallel Thematic Progression.

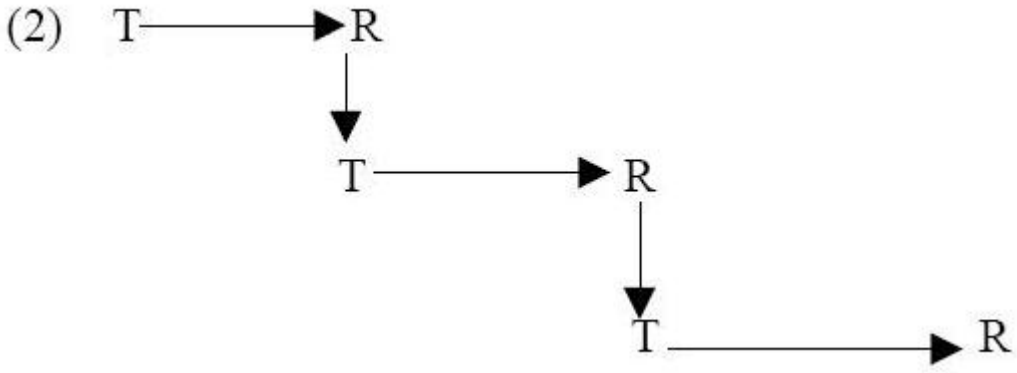


Figure 2: Linear Thematic Progression.

Being able to determine the thematic progession of a text through its sentences, would lead to a modular and flexible summarization process, because we would identify this features: (1) document's topic, (2) document's subtopics, (3) document's new information, (4) document type. As a result,

we could model the summary production process depending on what we want to give more importance to.

2 Topic Structure: Applicability to Automatic Summarization

The process of summarization requires natural language understanding, abstraction and generation. That is the reason why is so difficult for computers to generate summaries but, at the same time, it is a very challenging task which has been studied for more than 50 years. Since the abstractive summarization is even a harder task, what most systems do is to extract the most salient text units (usually sentences) from the source input. Apart from distinguishing between extracts and abstracts, summaries also differ according to what function they are intended for, and to whom they are targeted [8]. For example, a summary can be:

Indicative summaries, if enough content to alert users to relevant sources is provided, so that they can decide wthere to read the whole document or not.

Informative summaries, if they act as substitutes for the source, mainly by assembling relevant or novel factual information in a concise structure.

According to [16], the summarization process can be decomposed into three main subtasks: *topic identification*, *topic interpretatio*n, and *summary generation*. Topic analysis, which consists of two main tasks: (1) topc identification, and (2) text segmentation, is used to determine a text's topic structure, that is a representation indicating what topics are included in a text and how those topics change within the text.

As it is also stated in [19], Topic Detection is a very important step in the summarisation process, as it identifies the most salient themes in a text. There are a wide range of methods for performing this depending on the application. Moreover, not only is topic detection an advantageous stage on summarization, but also topic segmentation, according to the text topic structure, can be also suitable in this Natural Language Processing task. In the literature, systems such as [3], [18], [1], [10] , or [5] take profit of the advantages of combining topics' identification and segmentation in the Automatic Summarization task. In this section, some of the methods that exist, either for detecting a topic or segmenting according its topic structure are going to be described, giving examples of how they work, and how can be combined with summarization. Regarding the topic structure level previously mention in Section 1.1, only the last Section 2.7 (Topic-Focus Identification) tackles the problem from a sentence level perspective, whereas the rest ones are focused on the document level topic structure.

2.1 TextTiling

TextTiling is a method for partioning full-length text into coherent multi-paragraph units which represents subtopics [11, 12]. The goal of TextTiling is to segment written expository text into contiguous, non-overlaping discourse units that correspond to the pattern of subtopics in a text. However, the concept of topic used here differs from the topic/comment distinction found within individual sentences (this will be explained later in Section 2.7). It simply describes that one portion of texts is about something and the next portion is about something else.

As it is stated in [11], the capability to automate the recognition of this kind of structure in a full-text document should be useful for improving a variety of computational task, such as summarization. Automatic Summarization approaches that have taken profit of this segmentation method can be found in [18], [17], [10]. Furthermore, this method served as the basis for improving new ways of topic segmentation, like in the ones shown in [4],[15].

The discourse cues that TextTiling uses for identifying major subtopics shifts are patterns of lexical co-occurrence and distributions. Words that occur frequently throughout the text are often indicative of the main topic(s) of the text. On the other hand, words that are less frequent but more uniform in distribution are considered to be neutral and do not provide much information. The remainding words are of interest for the TextTiling algorithm, because it is assumed that they are indicative of the subtopic structure [12].

The idea behind this algorithm is to discover subtopic structure using term repetition (tf·idf measure), rather than semantic relationships between concepts. This is justified by the fact that term repetition alone can be a very useful indicator of subtopic structure as it is proven in [12]. Nevertheless, other attempts have been proposed for detecting text segments in narrative documents, by exploring semantic within concepts [14].

The TextTiling algorithm has three main parts: tokenization, lexical score determination and boundary identification. Next, a brief description of how it works is going to be explained:

- **Tokenization:** first, the input text is divided into individual lexical units (words). Stop words are filtered out, and the remainding words are reduced to their root, respectively. Then, the text is subdivided into pseudosentences (token-sequences) of predefined size (w, which is a parameter of the algorithm) to allow equal comparison of blocks.

- **Lexical Score Determintation:** once the first step has been performed, similarity between token-sequences is carried out by comparing adjacents blocks of text to see how similar they are, according to haow many words they have in common. Thus, if adjacents blocks

share many terms, and those shared terms are weightled heavily, there is strong evidence that they cohere with one another.

- **Boundary Identification:** this is done by assigning a depth score to each token-sequence, which correspond to how strongly the cues for a subtopic changed on both sides of a given token-sequence. Finally, boundaries are determined by locating the lowermost portions of these valleys.

Figure 3 depicts a visual of the idea of valleys and peaks used in the TextTiling algorithm to detect topic boundaries.

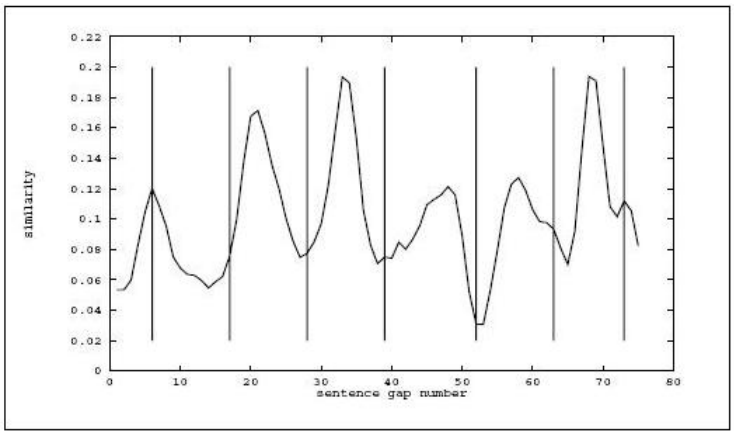


Figure 3: TextTiling visual idea.

In the Appendix (Section 3), it can be seen two examples of how TextTiling would partition a text. The first example is based on text AP880720-0262 of DUC[1] 2002 documents in which topics have also been detected in the Layered Topic Segmentation (Section 2.5), whereas the second one is based on the text used also in the Topic Analysis (Section 2.3). We considered appropriate to show an example of how the TextTiling algorithm would run taking into account texts that have been executed with other segmentation and topic detection mechanisms.

A free downloadable implementation of the algorithm can be found in *http://myweb.tiscali.co.uk/freddyyychoi/software/software.htm#jtexttile*.

2.2 C99: A Linear Text Segmentation

In [4] the same idea as in the TextTiling algortihm is described, taking into account that long documents typically address several topics or different aspects of the same topic. The aim of linear text segmentation is to discover

[1]Document Understanding Conferences: http://duc.nist.gov/

Range of sentences	3-11	3-5	6-8	9-11
C99	13%	18%	10%	10%
TT	46%	44%	43%	48%

Table 1: Error rate of C99 and TextTiling (TT) algorithms.

the topic boundaries, but in contrast to the approach taken in TextTiling, a domain independent method for segmenting text is shown here.

This segmentation algorithm, called C99, has two key elements: a similarity measure and a clustering strategy. This algorithm takes tokenized sentences as input, and a word stem frequencies dictionary is built. The similarity between each pair of sentences is computed by using the cosine measure, and as a result a similarity matrix is constructed. Because of the short text segments, this similarity value is unreliable, so an order of similarity between sentences is estimated, by means of a ranking matrix. The rank for a sentence will be the number of neighbouring elements with a lower similarity value, normalised by the number of elements examined. Clustering is the final process, and its aim is to determine the location of the topic boundaries (a text segment is defined by two sentences). Moreover, the number of segments to generate is determine automatically.

From the experiments reported in [4], it can be seen that the algorithm proposed there has better accuracy than the TextTiling algorithm, as it has lower error rate. In Table 1, the comparison between both algorithms, C99 and TextTiling (TT), is shown.

Following the same idea as in the previous section, in order to show a descriptive example of the performance of the algorithm, we can see in the Appendix (Section 3) the same two texts than before, but segmented with this algorithm.

A free available tool that implements this algorithm can be download from *http://myweb.tiscali.co.uk/freddyyychoi/software/software.htm#c99*.

2.3 Topic Analysis

The two methods previously described perform the segmentation of text through topic detection, but in none of them, labelling of topics or subtopics of each segment is provided. In [15], this problem is addressed, by performing both topic identication and text segmentation.

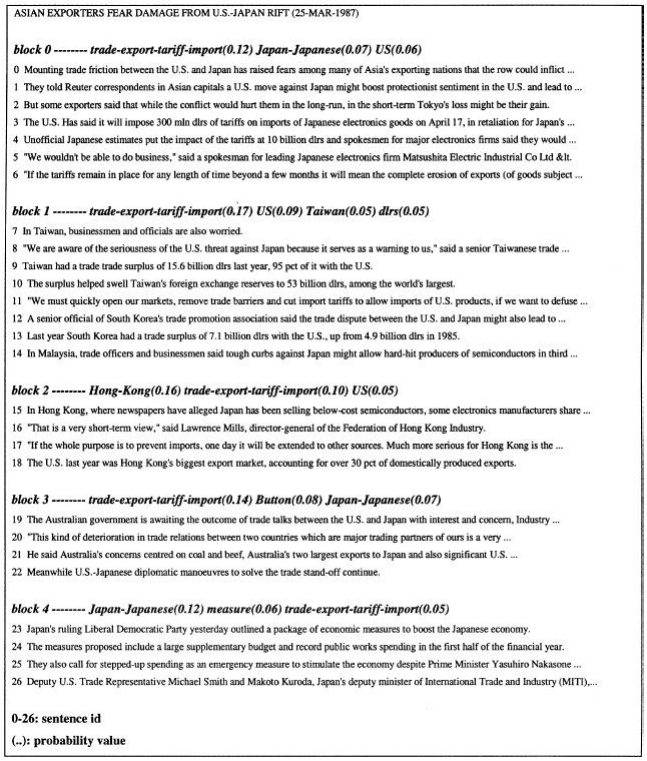
The ideas underlying this approach (the Stochastic Topic Model, STM) are mainly two. The first one is to represent a topic by means of a cluster of words that are closely related to the topic. This word clustering process is independent of topic analysis. The second idea deals with topic analysis and it involves the use of a stochastic model, called a finite mixture model to represent a word distribution within a text. This topic analysis consists

of three processes:

1. Topic Spotting.
2. Text Segmentation.
3. Topic Identification.

In the topic spotting stage, topics discussed in the given text are selected, so that STMs can be constructed on the basis of these topics. Further on, in the text segmentation process, text is split according the STMs, assuming that each block is generated by an individual STM. In the last step, the parameters of the STM for each segmented block are estimated, and topics with high probabilities for the block are selected, to perform the topic identifcation.

Finally, a text's topic structure representation is obtained, which consists of segmented blocks and their topics. Figure 4 shows the topic structure of a text. For a detailed description of the algorithm we encourage readers to see [15].


ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT (25-MAR-1987)

block 0 -------- trade-export-tariff-import(0.12) Japan-Japanese(0.07) US(0.06)

0 Mounting trade friction between the U.S. and Japan has raised fears among many of Asia's exporting nations that the row could inflict ...
1 They told Reuter correspondents in Asian capitals a U.S. move against Japan might boost protectionist sentiment in the U.S. and lead to ...
2 But some exporters said that while the conflict would hurt them in the long-run, in the short-term Tokyo's loss might be their gain.
3 The U.S. Has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics goods on April 17, in retaliation for Japan's ...
4 Unofficial Japanese estimates put the impact of the tariffs at 10 billion dlrs and spokesmen for major electronics firms said they would ...
5 "We wouldn't be able to do business," said a spokesman for leading Japanese electronics firm Matsushita Electric Industrial Co Ltd &lt.
6 "If the tariffs remain in place for any length of time beyond a few months it will mean the complete erosion of exports (of goods subject ...

block 1 -------- trade-export-tariff-import(0.17) US(0.09) Taiwan(0.05) dlrs(0.05)

7 In Taiwan, businessmen and officials are also worried.
8 "We are aware of the seriousness of the U.S. threat against Japan because it serves as a warning to us," said a senior Taiwanese trade ...
9 Taiwan had a trade trade surplus of 15.6 billion dlrs last year, 95 pct of it with the U.S.
10 The surplus helped swell Taiwan's foreign exchange reserves to 53 billion dlrs, among the world's largest.
11 "We must quickly open our markets, remove trade barriers and cut import tariffs to allow imports of U.S. products, if we want to defuse ...
12 A senior official of South Korea's trade promotion association said the trade dispute between the U.S. and Japan might also lead to ...
13 Last year South Korea had a trade surplus of 7.1 billion dlrs with the U.S., up from 4.9 billion dlrs in 1985.
14 In Malaysia, trade officers and businessmen said tough curbs against Japan might allow hard-hit producers of semiconductors in third ...

block 2 -------- Hong-Kong(0.16) trade-export-tariff-import(0.10) US(0.05)

15 In Hong Kong, where newspapers have alleged Japan has been selling below-cost semiconductors, some electronics manufacturers share ...
16 "That is a very short-term view," said Lawrence Mills, director-general of the Federation of Hong Kong Industry.
17 "If the whole purpose is to prevent imports, one day it will be extended to other sources. Much more serious for Hong Kong is the ...
18 The U.S. last year was Hong Kong's biggest export market, accounting for over 30 pct of domestically produced exports.

block 3 -------- trade-export-tariff-import(0.14) Button(0.08) Japan-Japanese(0.07)

19 The Australian government is awaiting the outcome of trade talks between the U.S. and Japan with interest and concern, Industry ...
20 "This kind of deterioration in trade relations between two countries which are major trading partners of ours is a very ...
21 He said Australia's concerns centred on coal and beef, Australia's two largest exports to Japan and also significant U.S. ...
22 Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the trade stand-off continue.

block 4 -------- Japan-Japanese(0.12) measure(0.06) trade-export-tariff-import(0.05)

23 Japan's ruling Liberal Democratic Party yesterday outlined a package of economic measures to boost the Japanese economy.
24 The measures proposed include a large supplementary budget and record public works spending in the first half of the financial year.
25 They also call for stepped-up spending as an emergency measure to stimulate the economy despite Prime Minister Yasuhiro Nakasone ...
26 Deputy U.S. Trade Representative Michael Smith and Makoto Kuroda, Japan's deputy minister of International Trade and Industry (MITI),...

0-26: sentence id
(..): probability value

Figure 4: Topic structure of text.

2.4 Topic Themes

According to [10], the topic structure is characterized in terms of topics themes, which are representations of events or states that are reiterated throughout the document collection, and therefore represent repetitive information. They considered five different topic representation: (1) representing topics via *topic signatures*. This idea comes from [16], where it is assumed that the topic of a document can be represented using a set of terms; (2)representing topics via *enhanced topic signatures* differs from the previuos one in the fact that the aim now is to discover relevant relations between two topic concepts; (3) representing topics via *thematic signatures*, which is carried out by segmenting documents using the TextTiling algorithm [12] first, and then assigning labels to themes to be able to rank them later; (4) representing topics by modeling the *content structure* of documents. The assumption here is that all texts describing a given topic are generated by a single content model (in this case a Hidden Markov Model). Finally, (5) the last method to represent topics within a text is to use *templates*, following the idea of the Information Extraction task.

They applied these methods to perform multi-document summarization. In Figure 5, it is shown an example of the topics detected through the different sorts of methods that have been suggested.

	T1=PINOCHET TRIAL		*T2=LEONID METEOR SHOWER*	
	Topic Signature TS1		*Topic Signature TS1*	
(a)	pinochet(N)	562	satellite(N)	444
	chile(N)	203	meteor(N)	392
	arrest(V)	104	storm(N)	160
	immunity(N)	100	leonid(N)	119
	argentina(N)	57	comet(N)	87
	garzon(N)	54	shower(N)	69
	dictator(N)	49	earth(N)	65
	london(N)	48	tuttle(N)	59
	crime(N)	46	sky(N)	49
	castellon(N)	43	kelly(N)	46
	request(N)	43	meteorid(N)	43
	Seed Relation for T1		*Seed Relation for T2*	
(b)	arrest – Pinochet		Leonids – shower	
	Topic Signature TS2 for T1		*Topic Signature TS2 for T2*	
(c)	arrest – Pinochet	483	shed – particle	395
	charge – murder	412	predict – shower	321
	react – Chile	216	position – satellites	147
	warrant – judge	114	study – scientists	102
	Topic Signature TS3 for T1		*Topic Signature TS3 for T2*	
(d)	Arrest	483	Meteor	392
	Reaction	216	Shower	321
	Warrant	114	Satellite	147
	Immunity	100	Scientists	102

Figure 5: Topic representation for topics T1= *Pinochet Trial* and T2= *Leonid Meteor Shower* (a) topic signature; (b) seed relations for T1 andT2; (c) enhanced topic signature; (d) thematic signatures.

Al the experiments performed and the results obtained are described in

detail in [10].

2.5 Layered Topic Segmentation

In [1], a topic segmentation algorithm is developed, and it serves as the basis for producing single and multi-document summaries. The topic segmentation approach proposed (Layered Topic Segmentation) associates key terms with each topic or subtopic, and outputs a tree-like table of content, which is called TOC. What they suggest here is that the text structure trees reflect the most important terms at general and more specific levels of topicality, indicating topically coherent segments from which sentences are mined for inclusion into summaries.

In the construction of the topic hierharchy, three different processes are involved. As a preprocessing step, the input text has to be tagged and chunked. Afterwards, in the first optional step, lexical chains[2] are built for the nouns of the text, by means of synonymy relations. In the next step, the main topic of each sentence is determined. This is computed through two heuristics: the initial position of noun phrases, and the persistency of the topic term. This is motivated by the fact that in languages that have SVO order, such as English, noun phrases in a clause-initial position tend to be indicative of the topic of the sentence and of its most important information. Moreover, the main topic of a sentence usually occurs persistently in consecutive sentences [7]. They proposed to study the definiteness of a noun phrase or the noun phrase embedding in future work. The third step takes into account the distribution of topic terms in the text. The detection of the main sentence topics and the main term distribution identifies topically coherent segments and aids in detecting topic shifts, nested topics and semantic returns, as well as in finding the most appropriate segmentation.

Basically, the way this TOC is used in the summarization process is the following: by restricting the number of levels of the TOC, the first sentence of each topical segment at the chosen level of detail would be included in the summary. The results obtained are acceptable [1].

In Figure 6 an example of the TOC structure is shown. Starting from this TOC, the summary generated can be seen in Figure 7.

2.6 Topic Identification through Lexical Chains

Lexical chains can model the lexical cohesion structure of a text. The first time that lexical chains were introduced within a text summarization system was performed was done in [2]. With the assumption that cohesion relations could provide good results in text summarization, lexical chains were used to detect and represent topics. To generate summaries, the first sentences of the strongest lexical chains were selected. However, in this case, topics

[2]This algorithm follows the one developed in [2].

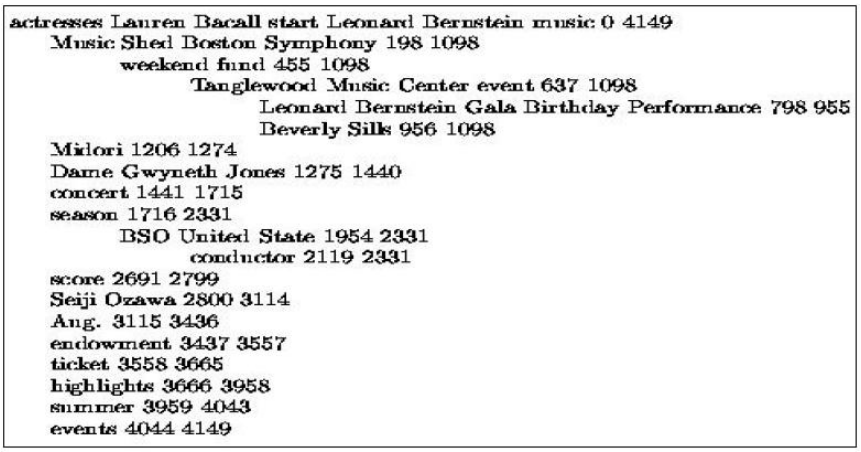


Figure 6: Example of a table of content.

Actresses Lauren Bacall, Betty Comden and Phyllis Newman are among performers in the birthday bash being thrown in August for conductor Leonard Bernstein at the music center where he got his start. The Tanglewood Music Center also is celebrating the 50th anniversary of its Music Shed this summer with a special concert by the resident Boston Symphony Orchestra featuring the same music, Beethoven's Ninth Symphony, that inaugurated the building in 1938. Violin soloist Midori will play two movements from his ``Serenade''. Dame Gwyneth Jones and Frederica von Stade will be among those performing highlights from ``Fidelio'' and other works to honor Bernstein's landmark opera recordings.

Figure 7: Example of a summary generated from a TOC.

were treated as single lexical chains. The motivation of [5] is based on the observation that a topic is formed of a group of lexical chains, and not from a single lexical chain on its own. Therefore, three lexical chains could correspond to a topic as their members could correspond to *what*, *when*, and *where* portions of the topic. Their lexical chaining implementations follows the ideas in [6], but using also meronym and holonym relations.

The summarization algorithm they suggested can be briefly described in these stages:

1. Sentence Detection
2. Part of Speech Tagging
3. Noun Phrase Detection
4. Lexical Chaining
5. Filtering Weak Lexical Chains (those which are below a certain strength criterion)

6. Clustering Lexical Chains Based on Co-occurrence
7. Extracting Sentences / Segmenting the Text Regard to Clusters

Clustering of lexical chains is used in order to represent topics in the text. The underlying assumption is that if two lexical chains tend to appear in same sentences, then there may be a relation between two sets in the given context. The similarity is computed by means of the cosine formula. After the identification of all clusters, text is segmented according to each of the identified clusters. In oder words, for each cluster, connected sequences of sentences are extracted as segments. To extract the final sentences to generate the summary, the first sentence of each segment is selected to become part of the summary, because it is assumed that humans tend to first explain the topic more generally, and then more details are given in the following sentences. Therefore, first sentences are general descriptions of the topic and these general descriptions does contain enough information to represent the text segment in the summary.

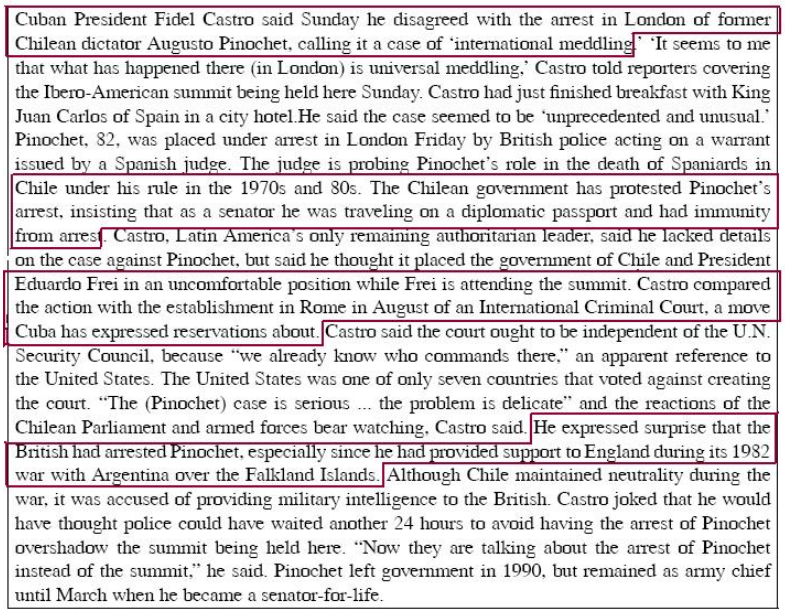

Cuban President Fidel Castro said Sunday he disagreed with the arrest in London of former Chilean dictator Augusto Pinochet, calling it a case of 'international meddling.' 'It seems to me that what has happened there (in London) is universal meddling,' Castro told reporters covering the Ibero-American summit being held here Sunday. Castro had just finished breakfast with King Juan Carlos of Spain in a city hotel.He said the case seemed to be 'unprecedented and unusual.' Pinochet, 82, was placed under arrest in London Friday by British police acting on a warrant issued by a Spanish judge. The judge is probing Pinochet's role in the death of Spaniards in Chile under his rule in the 1970s and 80s. The Chilean government has protested Pinochet's arrest, insisting that as a senator he was traveling on a diplomatic passport and had immunity from arrest. Castro, Latin America's only remaining authoritarian leader, said he lacked details on the case against Pinochet, but said he thought it placed the government of Chile and President Eduardo Frei in an uncomfortable position while Frei is attending the summit. Castro compared the action with the establishment in Rome in August of an International Criminal Court, a move Cuba has expressed reservations about. Castro said the court ought to be independent of the U.N. Security Council, because "we already know who commands there," an apparent reference to the United States. The United States was one of only seven countries that voted against creating the court. "The (Pinochet) case is serious ... the problem is delicate" and the reactions of the Chilean Parliament and armed forces bear watching, Castro said. He expressed surprise that the British had arrested Pinochet, especially since he had provided support to England during its 1982 war with Argentina over the Falkland Islands. Although Chile maintained neutrality during the war, it was accused of providing military intelligence to the British. Castro joked that he would have thought police could have waited another 24 hours to avoid having the arrest of Pinochet overshadow the summit being held here. "Now they are talking about the arrest of Pinochet instead of the summit," he said. Pinochet left government in 1990, but remained as army chief until March when he became a senator-for-life.

Figure 8: Example of sentences selected for a summary from the clustering of lexical chains.

2.7 Topic-Focus Identification

When dealing with *sentence topic* detection, it is essential to bear in mind and understand the concepts **topic-focus** (also known as *topic-comment*, *theme-rheme*, or *given-new information*). The topic can be understood as

what is "under discussion" in a given situation (often is the subject of a sentence, but not always). The focus is what it is said about the topic, usually the predicate of a sentence. Although there are no unequivocal criteria for determining the topic of a sentence, some tendencies can be given. A topic is [20]:

1. more likely to be definite than indefinite;
2. sooner pronoun than noun;
3. sooner subject than object.

There also appears to be a slight tendency in the order: first the topic and then the comment.

In [9], an automatic procedure for topic and focus identification is provided and it is stated that the topic-focus articulation can be relevant not only for a possible placement of a sentence in a context, but also for its semantic interpretation. They assume that all sentences must have a topic and focus, and although their procedure does not cover all possibilities occurring in English sentences, it gives a set of common rules and an algorithm to compute the topic-focus analysis of a sentence. For a deeper explanation of the algorithm, see [9].

Figure 9 shows an example of how this algorithm would identified the topic and focus of the sentence *"A painter arrived at a French village on a nice September day"* [3].

A painter arrived at a French village on a nice September day.

```
verb(topic(f),touch(0),sem(interm),label(arrived),
ltree(act(topic(f),touch(0),def(0),so(01),surf(np),
label(painter),ltree(det(a)))),rtree(loc(topic(f),touch(0),
def(0),sem(gen),so(012345678),surf(np),label(village),
ltree(prep(at),det(a),generic(french))),temp(topic(f),
touch(0),def(0),sem(gen),so(0),surf(np),label(day),
ltree(prep(on),det(a),generic(nice),generic(september)))))
```

Figure 9: Sentence analysis.

Finally, the output of the topic-focus algorithm would lead to the following analysis in Figure 10.

[3]The symbol topic denotes here whether the given item belongs to the topic or to the focus, touch stores the information if the complementation has been already determined, sem is the semantic information about the verb (general, specific, intermediate), and Itree and rtree are the left and right subtrees in the dependency tree. The word form is saved under label, so contains information about the position of the complementation in systemic ordering, and surf is the surface form (noun group, personal pronoun, indexical word, etc.). The other symbols are self-explanatory.

```
a painter(t/f) arrived(t/f) at a french village(t) on a nice
september day(f)
```

Figure 10: Topic-focus detection.

3 Appendix

- **TextTiling on text AP880720-0262:**

  Actresses Lauren Bacall, Betty Comden and Phyllis Newman are among performers in the birthday bash being thrown in August for conductor Leonard Bernstein at the music center where he got his start. The Tanglewood Music Center also is celebrating the 50th anniversary of its Music Shed this summer with a special concert by the resident Boston Symphony Orchestra featuring the same music, Beethoven's Ninth Symphony, that inaugurated the building in 1938. The Bernstein weekend alone, with tickets as high as $5,000 each, may bring in more than $1 million for the festival's endowment fund, according to Caroline Smedvig of the symphony. "Because Tanglewood is so special for him and, conversely, he's so special to us, the combination will produce what we hope will be a unique event," she said. The Leonard Bernstein Gala Birthday Performance is a four-day affair beginning on the composer's 70th birthday, Aug. 25, to raise money for the music center. Beverly Sills, who announced last month that she will step down as general director of the New York City Opera, will be the host of the event. Bacall and soprano Barbara Hendricks will perform a movement from Bernstein's Symphony No. 3, "Kaddish." Violin soloist Midori will play two movements from his "Serenade." Dame Gwyneth Jones and Frederica von Stade will be among those performing highlights from "Fidelio" and other works to honor Bernstein's landmark opera recordings.
  ==========
  The concert will also celebrate Bernstein's accomplishments in popular music with excerpts from such works as "West Side Story" and "On the Town." The orchestra opened the nine-week season with the special July 1 concert marking the first half-century of the Music Shed. The season of classical music will be expanded with performances by such popular rock, jazz and folk artists as The Beach Boys, George Benson and Leo Kottke.
  ==========
  But the highlight of the new season is the celebration of Bernstein's birthday. The BSO is even billing the birthday party as "one of the music world's great occasions" and is negotiating the television rights

in the United States and Europe. Among the conductor's friends who have been invited to the bash are pop star Michael Jackson and British Princess Diana, although Tanglewood officials say the have not yet received an answer from the celebrities. Bernstein was a member of the original Tanglewood Music Center class of 1940, where he became a protege of Tanglewood founder Serge Koussevitzky. When Koussevitzky retired in 1949, he tried to have the young Bernstein named as his successor, but the BSO's trustees chose Charles Munch because they considered Bernstein too young and tainted by popular music.
==========
He later wrote the score for the musical "West Side Story" and returned every summer to conduct and teach. Conductors Seiji Ozawa, John Williams, John Mauceri and Michael Tillson Thomas will share the poduim featuring some of his works on the night of his birth. The following night, Ozawa and special guests will appear in a Music Shed concert for a performance intended to show Bernstein's influence on other musicians. Aug. 27 will be given over to a performance of Bernstein's "Mass" by a 250-member troupe from the Indiana University school of music.
==========
The four-day celebration concludes with Bernstein conducting the annual Serge and Ola Koussevitzky Memorial Concert, including a performance of Tchaikovsky's Symphony No. 5, on Aug. 28. All proceeds go to the music center's endowment fund, and to a fund Bernstein established for aspiring young conductors. Ticket prices start at $20 for a space on the lawn and go up to $5,000 for front row seats for benefactors. Other highlights of the new season include several major debuts, including that of Soviet painist Vladimir Feltsman, who arrived in the United States last August after an an eight-year battle to emigrate; a return visit by the Israel Philharmonic; and a recital by Irish flutist James Galway. This summer is the second in a four-year jubilee of the festival's 50th anniversary. Events climax in 1990 with observances marking the 50th year of the music center, the festival's school.
==========

- **TextTiling on the same text used in Section 2.3 [15]:**

Mounting trade friction between the U.S. And Japan has raised fears among many of Asia's exporting nations that the row could inflict far-reaching economic damage, businessmen and officials said. They told Reuter correspondents in Asian capitals a U.S. Move against Japan might boost protectionist sentiment in the U.S. And lead to curbs on American imports of their products. But some exporters said that while the conflict would hurt them in the long-run, in the short-term

Tokyo's loss might be their gain. The U.S. Has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics goods on April 17, in retaliation for Japan's alleged failure to stick to a pact not to sell semiconductors on world markets at below cost. Unofficial Japanese estimates put the impact of the tariffs at 10 billion dlrs and spokesmen for major electronics firms said they would virtually halt exports of products hit by the new taxes. "We wouldn't be able to do business," said a spokesman for leading Japanese electronics firm Matsushita Electric Industrial Co Ltd <MC.T¿. "If the tariffs remain in place for any length of time beyond a few months it will mean the complete erosion of exports (of goods subject to tariffs) to the U.S.," said Tom Murtha, a stock analyst at the Tokyo office of broker <James Capel and Co¿.

==========

In Taiwan, businessmen and officials are also worried. "We are aware of the seriousness of the U.S. Threat against Japan because it serves as a warning to us," said a senior Taiwanese trade official who asked not to be named. Taiwan had a trade trade surplus of 15.6 billion dlrs last year, 95 pct of it with the U.S. The surplus helped swell Taiwan's foreign exchange reserves to 53 billion dlrs, among the world's largest. "We must quickly open our markets, remove trade barriers and cut import tariffs to allow imports of U.S. Products, if we want to defuse problems from possible U.S. Retaliation," said Paul Sheen, chairman of textile exporters <Taiwan Safe Group¿. A senior official of South Korea's trade promotion association said the trade dispute between the U.S. And Japan might also lead to pressure on South Korea, whose chief exports are similar to those of Japan. Last year South Korea had a trade surplus of 7.1 billion dlrs with the U.S., Up from 4.9 billion dlrs in 1985.

==========

In Malaysia, trade officers and businessmen said tough curbs against Japan might allow hard-hit producers of semiconductors in third countries to expand their sales to the U.S. In Hong Kong, where newspapers have alleged Japan has been selling below-cost semiconductors, some electronics manufacturers share that view. But other businessmen said such a short-term commercial advantage would be outweighed by further U.S. Pressure to block imports. "That is a very short-term view," said Lawrence Mills, director-general of the Federation of Hong Kong Industry. "If the whole purpose is to prevent imports, one day it will be extended to other sources. Much more serious for Hong Kong is the disadvantage of action restraining trade," he said. The U.S. Last year was Hong Kong's biggest export market, accounting for over 30 pct of domestically produced exports.

==========

The Australian government is awaiting the outcome of trade talks between the U.S. And Japan with interest and concern, Industry Minister John Button said in Canberra last Friday. "This kind of deterioration in trade relations between two countries which are major trading partners of ours is a very serious matter," Button said. He said Australia's concerns centred on coal and beef, Australia's two largest exports to Japan and also significant U.S. Exports to that country. Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the trade stand-off continue. Japan's ruling Liberal Democratic Party yesterday outlined a package of economic measures to boost the Japanese economy. The measures proposed include a large supplementary budget and record public works spending in the first half of the financial year. They also call for stepped-up spending as an emergency measure to stimulate the economy despite Prime Minister Yasuhiro Nakasone's avowed fiscal reform program. Deputy U.S. Trade Representative Michael Smith and Makoto Kuroda, Japan's deputy minister of International Trade and Industry (MITI), are due to meet in Washington this week in an effort to end the dispute.
==========

- **C99 on text AP880720-0262:**

Actresses Lauren Bacall, Betty Comden and Phyllis Newman are among performers in the birthday bash being thrown in August for conductor Leonard Bernstein at the music center where he got his start. The Tanglewood Music Center also is celebrating the 50th anniversary of its Music Shed this summer with a special concert by the resident Boston Symphony Orchestra featuring the same music, Beethoven's Ninth Symphony, that inaugurated the building in 1938. The Bernstein weekend alone, with tickets as high as $5,000 each, may bring in more than $1 million for the festival's endowment fund, according to Caroline Smedvig of the symphony. "Because Tanglewood is so special for him and, conversely, he's so special to us, the combination will produce what we hope will be a unique event," she said. The Leonard Bernstein Gala Birthday Performance is a four-day affair beginning on the composer's 70th birthday, Aug. 25, to raise money for the music center. Beverly Sills, who announced last month that she will step down as general director of the New York City Opera, will be the host of the event. Bacall and soprano Barbara Hendricks will perform a movement from Bernstein's Symphony No. 3, "Kaddish." Violin soloist Midori will play two movements from his "Serenade." Dame Gwyneth Jones and Frederica von Stade will be among those performing highlights from "Fidelio" and other works to honor Bernstein's landmark opera recordings. The concert will also celebrate Bernstein's

accomplishments in popular music with excerpts from such works as "West Side Story" and "On the Town." The orchestra opened the nine-week season with the special July 1 concert marking the first half-century of the Music Shed. The season of classical music will be expanded with performances by such popular rock, jazz and folk artists as The Beach Boys, George Benson and Leo Kottke. But the highlight of the new season is the celebration of Bernstein's birthday. The BSO is even billing the birthday party as "one of the music world's great occasions" and is negotiating the television rights in the United States and Europe. Among the conductor's friends who have been invited to the bash are pop star Michael Jackson and British Princess Diana, although Tanglewood officials say the have not yet received an answer from the celebrities. Bernstein was a member of the original Tanglewood Music Center class of 1940, where he became a protege of Tanglewood founder Serge Koussevitzky. When Koussevitzky retired in 1949, he tried to have the young Bernstein named as his successor, but the BSO's trustees chose Charles Munch because they considered Bernstein too young and tainted by popular music. He later wrote the score for the musical "West Side Story" and returned every summer to conduct and teach. Conductors Seiji Ozawa, John Williams, John Mauceri and Michael Tillson Thomas will share the poduim featuring some of his works on the night of his birth. The following night, Ozawa and special guests will appear in a Music Shed concert for a performance intended to show Bernstein's influence on other musicians. Aug. 27 will be given over to a performance of Bernstein's "Mass" by a 250-member troupe from the Indiana University school of music. The four-day celebration concludes with Bernstein conducting the annual Serge and Ola Koussevitzky Memorial Concert, including a performance of Tchaikovsky's Symphony No. 5, on Aug. 28. All proceeds go to the music center's endowment fund, and to a fund Bernstein established for aspiring young conductors.

==========

Ticket prices start at $20 for a space on the lawn and go up to $5,000 for front row seats for benefactors. Other highlights of the new season include several major debuts, including that of Soviet painist Vladimir Feltsman, who arrived in the United States last August after an an eight-year battle to emigrate; a return visit by the Israel Philharmonic; and a recital by Irish flutist James Galway. This summer is the second in a four-year jubilee of the festival's 50th anniversary. Events climax in 1990 with observances marking the 50th year of the music center, the festival's school.

==========

- **C99 on the same text used in Section 2.3 [15]:**

Mounting trade friction between the U.S. And Japan has raised fears among many of Asia's exporting nations that the row could inflict far-reaching economic damage, businessmen and officials said. They told Reuter correspondents in Asian capitals a U.S. Move against Japan might boost protectionist sentiment in the U.S. And lead to curbs on American imports of their products. But some exporters said that while the conflict would hurt them in the long-run, in the short-term Tokyo's loss might be their gain.

==========

The U.S. Has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics goods on April 17, in retaliation for Japan's alleged failure to stick to a pact not to sell semiconductors on world markets at below cost. Unofficial Japanese estimates put the impact of the tariffs at 10 billion dlrs and spokesmen for major electronics firms said they would virtually halt exports of products hit by the new taxes. "We wouldn't be able to do business," said a spokesman for leading Japanese electronics firm Matsushita Electric Industrial Co Ltd <MC.T¿. "If the tariffs remain in place for any length of time beyond a few months it will mean the complete erosion of exports (of goods subject to tariffs) to the U.S.," said Tom Murtha, a stock analyst at the Tokyo office of broker <James Capel and Co¿. In Taiwan, businessmen and officials are also worried.

==========

"We are aware of the seriousness of the U.S. Threat against Japan because it serves as a warning to us," said a senior Taiwanese trade official who asked not to be named. Taiwan had a trade trade surplus of 15.6 billion dlrs last year, 95 pct of it with the U.S. The surplus helped swell Taiwan's foreign exchange reserves to 53 billion dlrs, among the world's largest. "We must quickly open our markets, remove trade barriers and cut import tariffs to allow imports of U.S. Products, if we want to defuse problems from possible U.S. Retaliation," said Paul Sheen, chairman of textile exporters <Taiwan Safe Group¿. A senior official of South Korea's trade promotion association said the trade dispute between the U.S. And Japan might also lead to pressure on South Korea, whose chief exports are similar to those of Japan. Last year South Korea had a trade surplus of 7.1 billion dlrs with the U.S., Up from 4.9 billion dlrs in 1985. In Malaysia, trade officers and businessmen said tough curbs against Japan might allow hard-hit producers of semiconductors in third countries to expand their sales to the U.S.

==========

In Hong Kong, where newspapers have alleged Japan has been selling below-cost semiconductors, some electronics manufacturers share

that view. But other businessmen said such a short-term commercial advantage would be outweighed by further U.S. Pressure to block imports. "That is a very short-term view," said Lawrence Mills, director-general of the Federation of Hong Kong Industry. "If the whole purpose is to prevent imports, one day it will be extended to other sources. Much more serious for Hong Kong is the disadvantage of action restraining trade," he said. The U.S. Last year was Hong Kong's biggest export market, accounting for over 30 pct of domestically produced exports. The Australian government is awaiting the outcome of trade talks between the U.S. And Japan with interest and concern, Industry Minister John Button said in Canberra last Friday. "This kind of deterioration in trade relations between two countries which are major trading partners of ours is a very serious matter," Button said. He said Australia's concerns centred on coal and beef, Australia's two largest exports to Japan and also significant U.S. Exports to that country. Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the trade stand-off continue.
==========
Japan's ruling Liberal Democratic Party yesterday outlined a package of economic measures to boost the Japanese economy. The measures proposed include a large supplementary budget and record public works spending in the first half of the financial year. They also call for stepped-up spending as an emergency measure to stimulate the economy despite Prime Minister Yasuhiro Nakasone's avowed fiscal reform program. Deputy U.S. Trade Representative Michael Smith and Makoto Kuroda, Japan's deputy minister of International Trade and Industry (MITI), are due to meet in Washington this week in an effort to end the dispute.
==========

References

[1] Roxana Angheluta, Rik De Busser, and Marie francine Moens. The use of topic segmentation for automatic summarization. In *In Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization*, pages 66–70, 2002.

[2] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Inderjeet Mani and Mark Maybury, editors,* Advances in Automatic Text Summarization, pages 111–122. MIT Press, 1999.

[3] Branimir K. Boguraev and Mary S. Neff. Discourse segmentation in aid of document summarization. In *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 3*, page 3004, 2000.

[4] Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics (NAACL)*, pages 26–33, 2000.

[5] Gonenc Ercan and Ilyas Cicekli. Lexical cohesion based topic modeling for summarization. In *Proceedings of the 9th International Conference in Computational Linguistics and Intelligent Text Processing, CICLing 2008, Haifa, Israel, February 17-23*, pages 582–592, 2008.

[6] Michel Galley and Kathleen McKeown. Improving word sense disambiguation in lexical chaining. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI 2003), Acapulco, Mexico, August 9-15*, pages 1486–1488, 2003.

[7] Talmy Givón. *Syntax: A functional-typological introduction, II.* John Benjamins, 1990.

[8] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. *Computer*, 33(11):29–36, 2000.

[9] Eva Hajičová, Petr Sgall, and Hana Skoumalová. An automatic procedure for topic-focus identification. *Computational Linguistics*, 21(1):81–94, 1995.

[10] Sanda Harabagiu and Finley Lacatusu. Topic themes for multi-document summarization. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209, 2005.

[11] Marti A. Hearst. Texttiling: A quantitative approach to discourse. Technical report, Berkeley, CA, USA, 1993.

[12] Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[13] John Hutchins. Summarization: Some problems and methods. In *Meaning: The Frontier of Informatics*, pages 151–173. Aslib, 1987.

[14] Hideki Kozima. Text segmentation based on similarity between words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 286–288, 1993.

[15] Hang Li and Kenji Yamanishi. Topic analysis using a finite mixture model. *Infomation Processing & Management*, 39(4):521–541, 2003.

[16] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, 2000.

[17] Manuel Maña López. *Generación automática de resúmenes de texto para el acceso a la información*. PhD thesis, 2003. Adviser-Manuel de Buenaga Rodríguez.

[18] Joel Larocca Neto, Alexandre Santos, Celso A. A. Kaestner, and Alex Alves Freitas. Generating text summaries through the relative importance of topics. In *IBERAMIA-SBIA '00: Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI*, pages 300–309, 2000.

[19] Eamonn Newman, William Doran, Nicola Stokes, Joe Carthy, and John Dunnion. Comparing redundancy removal techniques for multi?document summarisation. In *Proceedings of STAIRS, August*, pages 223–228, 2004.

[20] Jan Renkema. *Introduction to Discourse Studies*. John Benjamins Publishing Company, 2004.