Lecture 11
Automatic Abstracting

陳光華
台灣大學圖書資訊學系
khchen@ntu.edu.tw

摘要的功能

- 宣示功能（Announcement）
  - 宣示原始文件的存在性
- 篩檢功能（Screening）
  - 判定原始文件的相關性
- 取代功能（Substitution）
  - 取代原始文件
- 回溯功能（Retrospection）
  - 查詢原始文件

摘要的類型

- 指示性摘要
  - 宣示功能、篩檢功能、回溯功能
- 訊息性摘要
  - 取代功能、回溯功能
- 評論性摘要
  - 回溯功能
- 摘錄
  - 宣示、篩檢、取代以及回溯功能

摘要

- 摘要是文件的精緻版，亦即以較少的文字表述原始文件所欲傳達的訊息
- 摘要的長度
  - 「研究報告及專論，摘要宜少於250字，附錄及簡訊性質之資料，以100字為佳，至於社論或讀者來函只需要一個句子即可，長篇論著，如：技術報告、學位論文，其摘要以一頁以內，且以500字為限」

自動摘要

- 以自動化的程序製作原始文件的精緻版
- 從自動摘要模型的角度出發，自動摘要分為兩種作法：
  - 由文件中挑選適當的段落或句子構成摘要，亦即製作所謂的「摘錄」
  - 分析原始文件，抽取文件的「概念表意」（Conceptual Representation），再進行「摘要的產生」（Summary Generation）

TIPSTER

- MUC
  - Message Understanding Conference
- TREC
  - Text Retrieval Conference
- SUMMAC
  - Summarization Conference
- MET
  - Multilingual Entity Task

SUMMAC

- 評比三種不同用途的文件摘要
- Categorization Task
- Adhoc Task
- Q&A Task

Categorization Task

- 評估自動摘要系統對於文件關鍵概念的掌握能力
- 參與競賽的團隊取得大會準備的500篇文件，其中每100篇各與某一Topic相關，總共有五個不同的Topic
- 製作完成的摘要送回大會，大會要求評估人員設定該摘要的屬於哪一個Topic，如果無法決定則設定為第六個Topic
- 依據評估人員的評估結果，計算參賽系統的績效

Topics

- Topic並不是主題
- 是對於資訊需求的描述
- 目前TIPSTER已經製作350個Topic

A Topic in TIPSTER

<top>
<head> Tipster Topic Description
<num> Number: 001
<dom> Domain: International Economics
<title> Topic: Antitrust Cases Pending
<desc> Description:
Document discusses a pending antitrust case.
<narr> Narrative:
To be relevant, a document will discuss a pending antitrust case and will identify the alleged violation as well as the government entity investigating the case. Identification of the industry and the companies involved is optional. The antitrust investigation must be a result of a complaint, NOT as part of a routine review.
...

Adhoc Task

- 評估自動摘要系統能否提供使用者找尋的資訊
- 製作使用者導向（User-directed）的文件摘要
- 大會提供20個Topic，每一個Topic有50篇文章，共計1000篇文章
- 參賽系統必須視Topic為使用者資訊需求的描述，依據Topic建構每一文件的摘要
- 當大會接獲參賽者製作完成的文件摘要，評估人員必須閱讀每一篇摘要，並且判定摘要是否與Topic相關

最佳摘要與定常摘要

- 參與前述兩類Task競賽的團隊，可以選擇製作定長摘要（Fixed-length Summary）或最佳摘要（Best Summary），或是兩者皆予以製作。
- 定長摘要，其長度不可超過原文的10%
- 最佳摘要則無長度限制
  - 文件摘要的長度為評比的項目
  - 評估人員閱讀摘要的時間也是評比的項目
  - 因此，過長的摘要是參賽團隊必須極力避免的

Q&A Task

- SUMMAC將Q&A Task產生的摘要假想為撰寫報告過程中所需的資訊
- 亦即為了撰寫有價值的報告，撰寫人員必須取得某些特定問題的相關資訊
- 難度相對較高
- SUMMAC認為這個競賽項目並不容易，因此聲明這個競賽項目仍處於初期設計的階段。

一個自動摘要模型

- 採取的作法是製作「摘錄」型摘要
- 直接由文件擷取重要的句子，製作文件的摘要

模型的假設

- 一般而言，組織完善、意念完整的文件，其名詞與名詞以及名詞與動詞的關係相當密切
- 本模型的建構是基於下列的假設
  - 名詞與動詞共存於述語參數結構
  - 名詞間的關係是建構於言談層次

詞彙的統計值

- 詞彙的重要性
- 詞彙的重複性
- 詞彙的共現性
- 詞彙的距離

詞彙的重要性

- 當詞彙出現於文件時，做為作者意念核心的機會，
- 並非所有的詞彙都一樣重要
  - 將文件中的冠詞、副詞、以及介系詞等詞彙刪除，仍然知道這份文件的梗概，這說明了上述的詞彙並不十分重要
  - 名詞與動詞十分重要
- IDF代表詞彙對文件的重要程度

$$IDF(w) = \log((P-O(w))/O(w))$$

詞彙的共現性

- 意念一致的文件資料，作者使用的詞彙必然趨向某一個語意範疇
- 從統計的觀點，這表示該語意範疇的詞彙一起出現的機率比較大
- 判斷那些詞彙屬於同樣的語意範疇是相當困難的工作
- 由大規模語料庫計算詞彙的共現程度就很簡單

詞彙的距離

- 詞彙的位置也很重要
- 基於文件是有生命的文字組合的觀點，相關的詞彙其出現的距離必定不會太長
- 一旦相隔太遠，彼此之間的相乘效果就大打折扣
- 引入距離的因素，比較能夠忠實反應寫作的行為。距離的計算

距離的計算

蘇聯$_1$ 許多 製造$_2$ 民生$_3$ 日用品$_4$ 的 工業$_5$ 得到$_6$ 政策性$_7$ 的 補貼$_8$，其 目的$_9$ 是 保持$_{10}$ 物價$_{11}$ 的 平穩$_{12}$。但 補貼$_{13}$ 勢 難 普及$_{14}$ 於 各行各業$_{15}$，因此 又 造成$_{16}$ 某 些 日用品$_{17}$ 不足$_{18}$ 或 完全 缺乏$_{19}$ 的 後遺症$_{20}$ 。 現在 既然 要 引進$_{21}$ 市場$_{22}$ 經濟$_{23}$，補貼$_{24}$ 政策$_{25}$ 又 勢 難 繼續$_{26}$，一旦，放棄$_{27}$，許多 民生$_{28}$ 物資$_{29}$ 的 價格$_{30}$ 必然 上漲$_{31}$，於是 又 引出$_{32}$ 民間$_{33}$ 屯積$_{34}$ 物資$_{35}$ 與 通貨膨脹$_{36}$ 的 壓力$_{37}$。

計算模型

$$SNN(n_i) = \sum_j \frac{IDF(n_i) \times IDF(n_j) \times f(n_i, n_j)}{f(n_i) \times f(n_j) \times D(n_i, n_j)}$$

$$SNV(n_i) = \sum_j \frac{IDF(n_i) \times IDF(v_j) \times f(n_i, v_j)}{f(n_i) \times f(v_j) \times D(n_i, v_j)}$$

$$CS(n) = pn \times SNN(n) + pv \times SNV(n)$$

消去內差法

$$S_N = \sum_i \frac{pn \times SNN(n_i)}{pn \times SNN(n_i) + pv \times SNV(n_i)}$$

$$S_V = \sum_i \frac{pv \times SNV(n_i)}{pn \times SNN(n_i) + pv \times SNV(n_i)}$$

$$pn = \frac{S_N}{S_N + S_V} \qquad pv = \frac{S_V}{S_N + S_V}$$

摘錄強度

- 一旦求得每一個名詞的聯結強度，便能夠進而得到每一個句子的重要性
- 假設某一個句子s_i有m個不同的名詞，該句子被摘錄的可能性度量，若以摘錄強度（Extraction Strength，簡稱ES）稱之，可以用下列數學式度量：

$$ES(s_i) = \sum_{j=1}^{m} CS(n_{ij}) / m$$

摘要的產生

- 文件的句子可以排成有序集合（Ordered Set）
- 文件的摘要就可以由該有序集合擷取數量適當的句子組成
- 若要製作定長摘要與最佳摘要，則可以設定一個門檻值（Threshold），刪除有序集合中摘錄強度小於門檻值的句子即構成最佳摘要
- 從最佳摘要再刪除部份的句子，使得有序集合中句子數小於原文的10%即構成定長摘要

其他考量

- 句子的位置事實上扮演重要的角色
  - 第一段的第一個句子
  - 最後一段的第一個句子
- 線索詞彙（Cue Word）具有舉足輕重的份量
  - 增益詞（Bonus Word）:重要、顯著
  - 損益詞（Stigma Word）不可能、幾乎不
- 修正模型

$$ES(s_i) = w_1 \times \sum_{j=1}^{m} CS(n_{ij}) / m + w_2 \times POS(s_i) + w_3 \times CW(s_i)$$

結論

- 網際網路更加速了資訊的流通，縮短了資訊形成知識所需要的時間
- 網際網路膨脹地過於快速，資訊累積太快造成雜訊過多，卻又干擾了知識的形成
- 加值型服務業者為處於世紀末的人們提供搜尋引擎、主題指引以及其他特殊的服務
- 目前的網際網路服務無法提供啟發性的資訊
- 21世紀需要的服務是資訊擷取與自動摘要

結論 （續）

- 文件摘要是新資訊時代重要的啟發性資訊服務
- 透過文件摘要可快速地判讀文件的相關性，不必取得完整文件後才發覺文件不符合需求
- 文件摘要的服務能夠有效降低網際網路的流量
- 對於文件的使用者與整體網際網路環境，文件摘要都是值得期待的服務

結論 （續）

- 人工製作文件摘要具有高品質的特性，但是緩不濟急
- 審視網際網路文件數量極為龐大的事實，自動製作文件摘要是無法避免的作法
- 一個可能的模型
  - 建構於詞彙的重要性、詞彙的重複性、詞彙的共現性、詞彙的距離等四項要素
  - 考慮文件中句子的位置以及線索詞彙

努力的方向

- 因應網際網路的特性，模型有待進一步的實驗與修正，以適應各種不同類型的文件
- 必須更加縮短所需要的計算時間

Other Approaches

- Cut and Paste Summarization
- Lexical Chain Summarization
- Summarization as Sentence retrieval

Cut and Paste Summarization

Hongyan Jing and Kathleen R. Mckeown
Department of Computer Science
Columbia University, New York, USA
(NAACL 2000)

Cut and Paste Summarization

- Reuse the text in an article to generate the summary
- Instead of simply extracting sentences as current summaries do, this system will "smooth" the extracted sentences by editing them
- Uses operations derived from an analysis of human written abstracts
- The summarizer edits extracted sentences, using *reduction* to remove inessential phrases and *combination* to merge resulting phrases together as coherent sentences

Sentence Reduction

- Document sentence

  When it arrives sometime next year in new TV sets, the V-chip will give parents a new and potentially revolutionary device to block out programs they don't want their children to see.

- Summary sentence

  The V-chip will give parents a device to block out programs they don't want their children to see.

Sentence Combination

- Text sentence 1

  But it also raises serious questions about the privacy of such highly personal information wafting about the digital world.

- Text sentence 2

  The issue thus fits squarely into the broader debate about privacy and security on the internet, whether it involves protecting credit card number or keeping children from offensive information

- Summary sentence

  But it also raises the issue of privacy of such personal information and this issue hits the head on the nail in the broader debate about privacy and security on the internet.

Other Operations

- syntactic transformation
  - The position of the subject in a sentence may be moved from the end to the front
- lexical paraphrasing
  - Replace phrase with their paraphrases.
  - For example,
    - Point out -> note
    - Fits squarely into -> hit the head on the nail

Other Operations (Continued)

- generation and specification
  - Replace phrase or clauses with more general or specific descriptions.

    e.g., the White House's top Drug official -> Gen. Barrey R. McCaffrey, the White House's top Drug official
- reordering
  - Change the order of extracted sentences.

Lexical Chain Summarization

Regina Barzilay and Michael Elhadad
Mathematics and Computer Science Dept.
Ben Gurion University in the Negev, Israel
(ACL/EACL 1997)

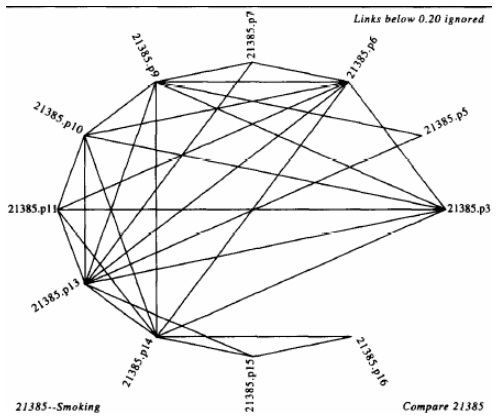
Lexical Chain Summarization

- Produce a summary of an original text without requiring its full semantic interpretation, but instead relying on a model of the topic progression in the text derived from lexical chains.
- Lexical chains in a text were computed by merging several robust knowledge sources:
  - WordNet thesaurus
  - a POS tagger
  - shallow parser for identifying of nominal groups
  - a segmentation algorithm

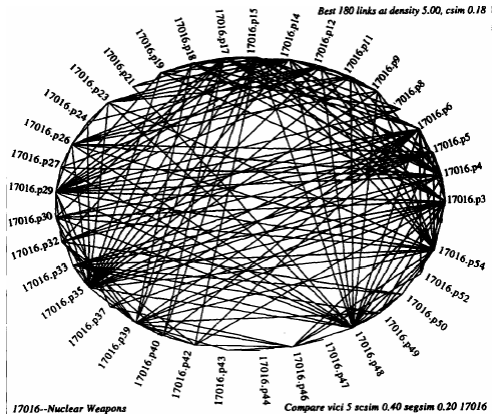
Summarization in 3 Steps

- the original text is first segmented
- lexical chains are constructed
- strong chains are identified and significant sentences are extracted from the text.

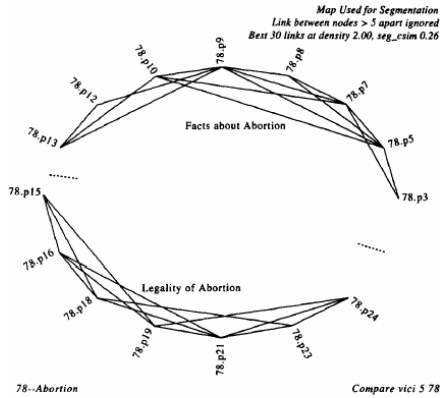
Lexical Chain (1/3)



Lexical Chain (2/3)



Lexical Chain (3/3)



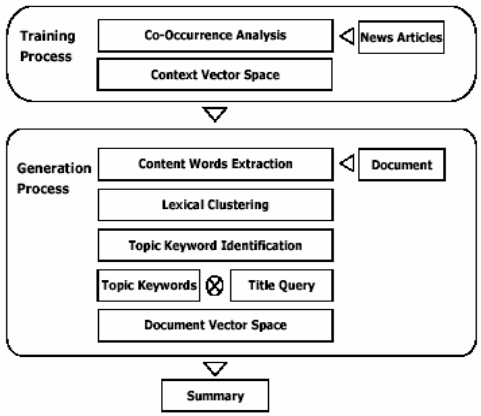
Some Pending Problems

- the need for anaphora resolution
- a model for reconstructing a coherent summary out of selected sentences
- a method to handle long sentences
- a method to control the degree of condensation.

Summarization as Sentence Retrieval

Source: “Topic keyword Identification for Text Summarization using Lexical Cluster”

DOCUSUM Summarization System



Effectiveness of Various Features

Method	30%	10%	4 Sentences
DOCUSUM	**51.1**	**51.2**	**53.6**
Title	48.6	43.3	51.6
Location	49.4	46.6	51.6
DOCUSUM*	44.6	39.6	47.1
Frequency	35.9	14.8	38.4

DOCUSUM: Topical Words + Title
DOCUSUM*: Topical Words

Source: "Topic keyword Identification for Text Summarization using Lexical Cluster"

MMR Concept

- Maximal marginal relevance
  - A document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents

$$MMR \overset{def}{=} \arg\max_{D_i \in R \setminus S} [\lambda sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} sim_2(D_i, D_j)]$$

Where C is a document collection, Q is a query.
R=IR(C,Q,θ) is the ranked list of an IR system, given C and Q and relevance threshold θ.
S is the subset of documents in R already selected
R\S is the set difference

Statistical Characters of Abstracts

Source: "A Case Against some Features Used in Content Selection" In the Proceedings of Workshop on Multilingual Summarization and Question Answering.

Test Data

- 57 journal articles selected from *Behavioral Ecology and Sociobiology* (bes) and *Oecologia* (oec) 生態學

Statistics of Author-Tailored Abstracts

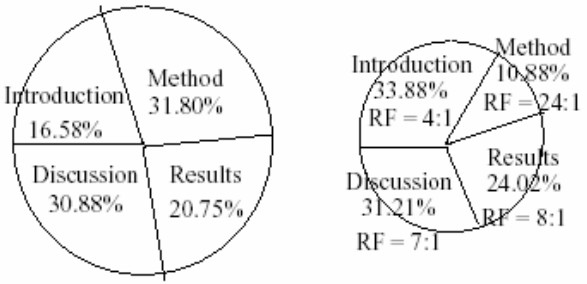
	full-text (ft)	abstract(ab)	RF
Corpus size	7938 sn; 175,613 wd	534 sn; 11,975 wd	15:1; 15:1
Size of article	62–269 sn; 1,552-6,333wd	5–21 sn; 109-415 wd	7:1–31:1; 7:1–31:1
Av. size of article	139 sn; 3,081 wd	9 sn; 210 wd	15:1; 15:1
Range of sn length	4–129 wd	7–80 wd	

sn = sentence; wd = word; Av. sn length = 22 wds; Reduction factor, RF = No. ft-sn (or wd): No. ab-sn (or wd).

Source: "A Case Against some Features Used in Content Selection" In the Proceedings of Workshop on Multilingual Summarization and Question Answering.

Distribution of Sentences in Sections

(a) Full-text sentences (b) Selected sentence



Source: "A Case Against some Features Used in Content Selection" In the Proceedings of Workshop on Multilingual Summarization and Question Answering.

Cue Words

Table 1. Sentences Cued (i.e. with any of the following lexemes: CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST) and Not Cued Selected by Author for Abstracting

cor-pora	selected by author		not selected by author		total
	with cues	without cues	with cues	without cues	
bes1	60	161	264	1418	1903
bes2	58	167	258	1454	1937
oec1	48	200	239	1188	1675
oec2	76	204	363	1780	2423
Total	242 (3 %)	732 (9 %)	1124 (14 %)	5840 (74 %)	7938

Init/Fin Sentences Used in Abstracts

Sub-corpus	Section							
	I		M		R		D	
	Init+Fin+	Para	Init+Fin	Para	Init+Fin	Para	Init+Fin	Para
bes1	37	73	13	117	27	99	27	100
bes2	36	59	13	101	19	69	29	130
oec1	50	70	14	100	30	69	39	97
oec2	38	66	12	141	33	106	41	137
total	161	258	52	459	109	343	136	464
proba-bility	0.31		0.06		0.16		0.15	

+Init+Fin = no. of initial/final position sentences that were used by an author in abstracting;
Para. = no. of paragraphs;
probability = Init+Fin / (Para x 2);

Source: "A Case Against some Features Used in Content Selection" In the Proceedings of Workshop on Multilingual Summarization and Question Answering.

Effectiveness of Cued Sentences

	Section				Tot. I + R + D (excl. M)
	I	M	R	D	
Init/Fin + Cued + Selected	47	2	20	49	116
Init/Fin + Cued	130	91	106	235	471
Init/Fin + Selected	161	52	109	136	406
Init/Fin	516	918	686	928	2130

Init/Fin = Initial or Final sentence;
Selected = sentence selected for abstracting by an author;
Cued = sentence cued with any of the following lexemes: CONCLUDE, INDICATE, RESULT, SHOW, STUDY, SUGGEST;

Source: "A Case Against some Features Used in Content Selection" In the Proceedings of Workshop on Multilingual Summarization and Question Answering.

Multi-Doc Summarization

- Useful
- How
- Discussion