Introduction to Automatic Summarization

Li Sujian
Institute of Computational Linguistics,
Peking University
lisujian@pku.edu.cn

- In the last decade there has been a surge of interest in automatic summarizing. … There has been some progress, but there is much to do.

  Karen Sparck Jones

Karen Sparck Jones, **automatic summarising: the state of the art**, **Information Processing & Management, 43(6): 1449--1481, 2007.**

大纲

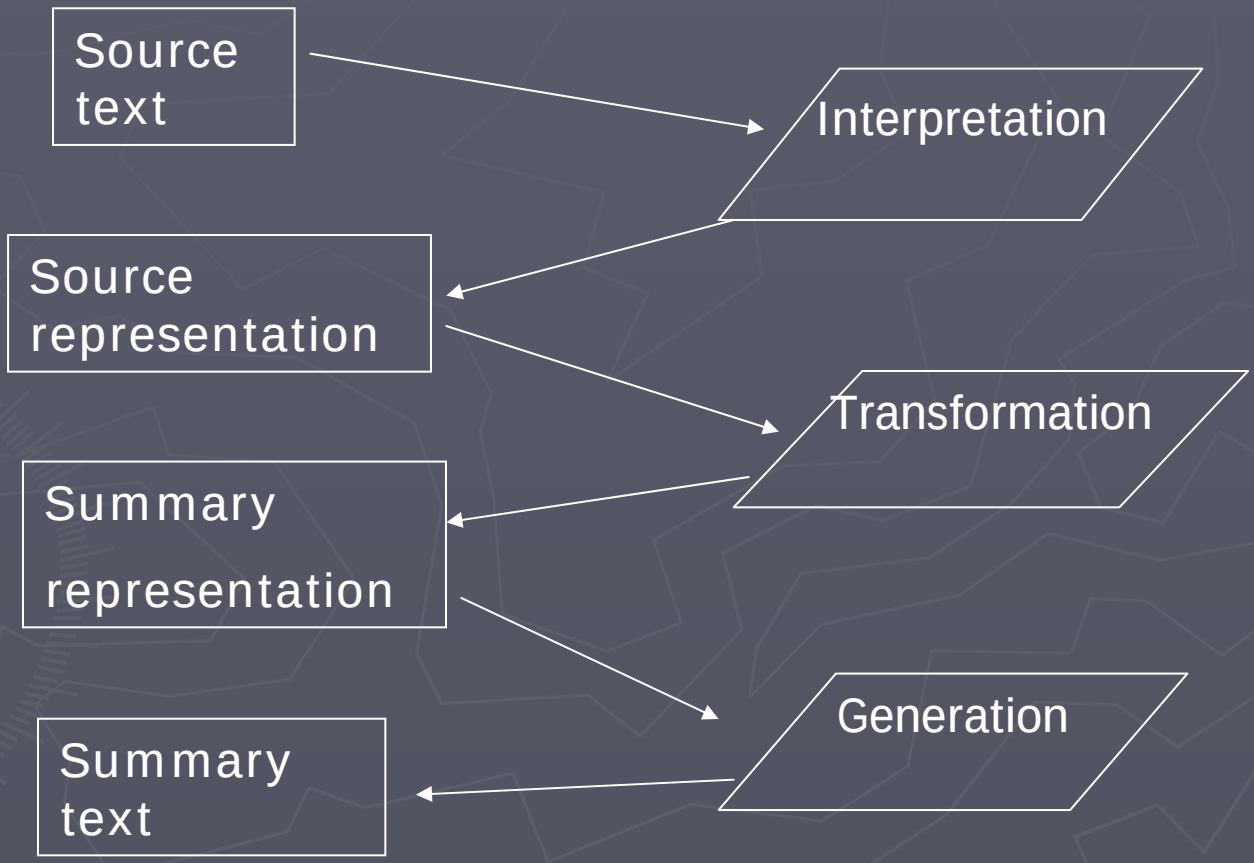
- 自动文摘技术简介
  - 背景
  - 评测
    - Manual
    - Automatic
- 一个多文档自动摘要系统的搭建
  - 语句特征选择
  - 语句打分 – 机器学习方法的使用
- 其他
  - Graph ranking
  - Sentence Compression
  - Discourse Structure

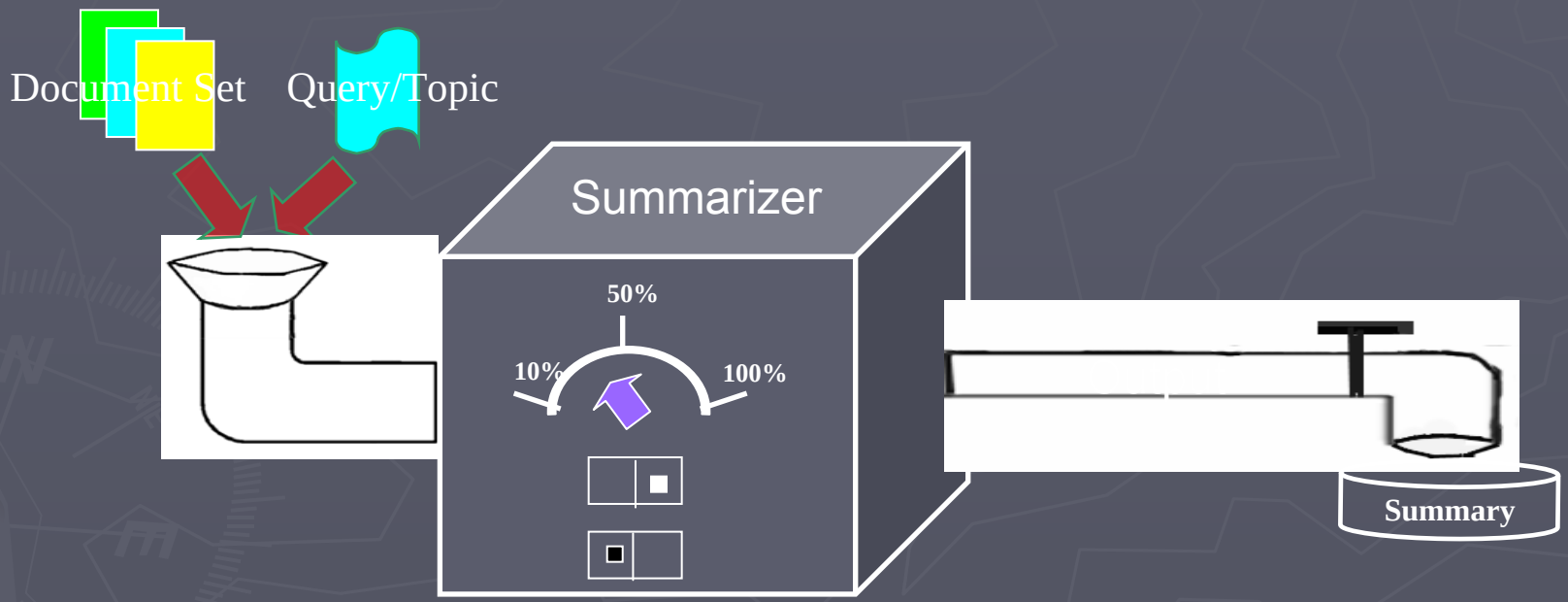
Part 1: 自动文摘技术简介

- Summary definition by Sparck Jones (1999)

  “a reductive transformation of source text to summary text through content condensation by selection and/or generalization on what is important in the source.”

Schematic summary processing model




Document Set
Query/Topic
Summarizer
50%
10%
100%
Summary

Context factors affecting summarizing

- input factors
  - form - language, medium, structure, genre, length
  - subject type
  - Unit
  - Author
  - header (metadata)

  [contrasted examples: archaeological paper, children's tale]
- purpose factors
  - use
  - audience
  - envelope - time, location, formality, trigger, destination

  [contrasted examples: emergency alert, literary review]
- output factors
  - material - coverage, reduction, derivation, specialty
  - style
  - format - language, medium, structure, genre

  [contrasted examples: bullet item list, prose paragraph]

文本自动摘要技术的发展

- *20世纪50年代* 基于分析位置信息和关键词的方法 (Luhn, 1958; Edmundson, 1969)
- *80年代* 基于文档内容的方法 (Schank and Abelson, 1977; DeJong; 1979)
- *90年代* 基于信息提取的方法 (Hovy and Lin, 1997; Kupiec et al., 1995)
- *当前* 语义分析、图模型、自动问答、机器学习……

文档自动摘要方法的分类

- 按照摘要生成方法分类
  - 句子抽取：从原文中抽取出完整的句子组成
    - 摘要的语法性、流畅性较好
    - 性能受原文质量影响
    - 目前的主流方法
  - 句子生成：从原文中摘取句子片断组成摘要
    - 生成摘要更灵活
    - 较难保证句子的可读性
    - 尚处于实验阶段

文本自动摘要任务的发展

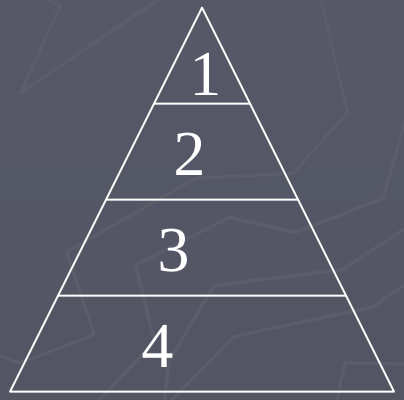
- 科技领域 → 一般领域
- 单文档 → 多文档
- 单语言 → 多语言
- 查询无关 → 面向查询
- 静态摘要 → 动态摘要

Document Understanding Conference

- NIST从2001年开始组织的大型文档摘要评测会议
- DUC定义的摘要任务的发展
  - 2001-2002 100字的查询无关的单文档和多文档摘要
  - 2003 基于查询问题的文档摘要
  - 2004 多语言文档摘要
  - 2005-2007 基于主题查询的多文档摘要

Evaluation

- Manual
  - Linguistic Quality(readability) – Responsiveness(content)
    - *Grammaticality*
    - *Non-redundancy*
    - *Referential clarity*
    - *Focus*
    - *Structure*
- Pyramid: SCU



– BE(basic element)

- Automatic
  - Rouge
    - ROUGE 2
    - ROUGE SU4

- Gold standards
  - Gold-standard comparison has nevertheless been seen as sufficiently attractive, and operationally viable, for it to have been widely used in both evaluation programs and individual tests during the last decade.
  - Model summary
  - Peer (system) summary

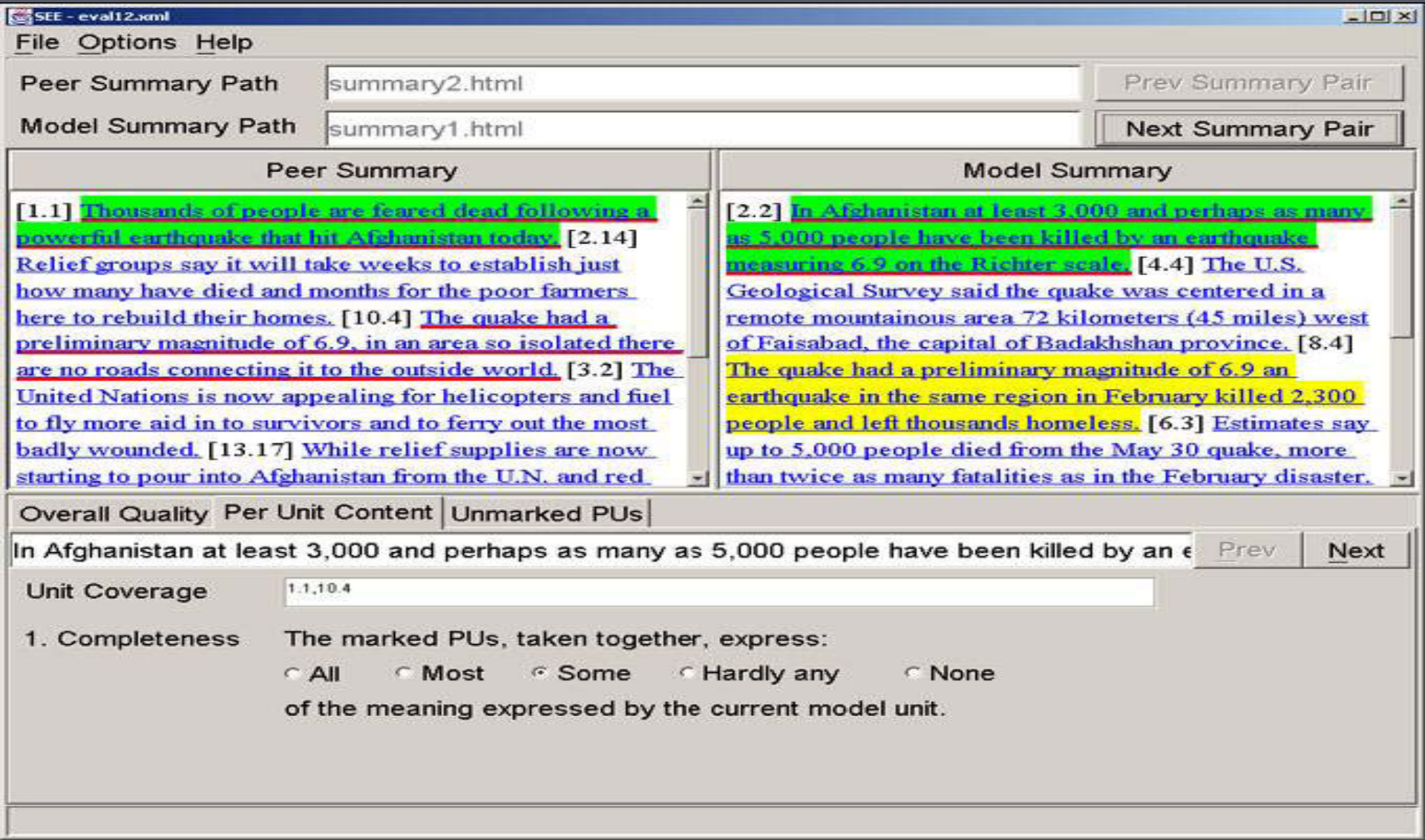
- Baseline
  - Lead baseline: e.g. take the first *n* words in the last document in the collection.
  - Coverage baseline: take the first sentence in the first document, the first sentence in the second document and so on until it had a summary of *n* words length.

Summary Evaluation Environment

- Using SEE, NIST assessors who created the ‘ideal’ written summaries did pairwise comparisons (both the content and the quality) of their summaries (the *model* text) to the system-generated summaries (the *peer* text).

- Each text was decomposed into a list of units and displayed in separate windows. The sentence was used as the smallest unit of evaluation.

SEE


SEE - eval12.xml
File Options Help
Peer Summary Path
summary2.html
Prev Summary Pair
Model Summary Path
summary1.html
Next Summary Pair
Peer Summary
[1.1] Thousands of people are feared dead following a powerful earthquake that hit Afghanistan today. [2.14] Relief groups say it will take weeks to establish just how many have died and months for the poor farmers here to rebuild their homes. [10.4] The quake had a preliminary magnitude of 6.9, in an area so isolated there are no roads connecting it to the outside world. [3.2] The United Nations is now appealing for helicopters and fuel to fly more aid in to survivors and to ferry out the most badly wounded. [13.17] While relief supplies are now starting to pour into Afghanistan from the U.N. and red
Model Summary
[2.2] In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an earthquake measuring 6.9 on the Richter scale. [4.4] The U.S. Geological Survey said the quake was centered in a remote mountainous area 72 kilometers (45 miles) west of Faisabad, the capital of Badakhshan province. [8.4] The quake had a preliminary magnitude of 6.9 an earthquake in the same region in February killed 2,300 people and left thousands homeless. [6.3] Estimates say up to 5,000 people died from the May 30 quake, more than twice as many fatalities as in the February disaster.
Overall Quality
Per Unit Content
Unmarked PUs
In Afghanistan at least 3,000 and perhaps as many as 5,000 people have been killed by an e
Prev
Next
Unit Coverage
1.1,10.4
1. Completeness
The marked PUs, taken together, express:
All
Most
Some
Hardly any
None
of the meaning expressed by the current model unit.

Rouge basics

- Rouge(Recall-Oriented Understudy for Gisting Evaluation) Recall-oriented, within-sentence word overlap with model(s)
- Models - no theoretical limit to number
  - compared system output to 4 models
  - compared manual summaries to 3 models
- Using n-gram
  - Correlate reasonably with human coverage judgements
  - Not address summary discourse characteristics, and suffer from lack of text cohesion or coherence
- ROUGE v1.2.1 measures
  - ROUGE-1,2,3,4: N-gram matching where N = 1,2,3,4
  - ROUGE-LCS: Longest common substring
  - ROUGE-W-1.2: Favors LCS with least intervening material

Rouge(2)

$$R_n(X) = \frac{\sum_{j=1}^{h} \sum_{i \in N_n} \min(X_n(i), M_n(i, j))}{\sum_{j=1}^{h} \sum_{i \in N_n} M_n(i, j)}$$

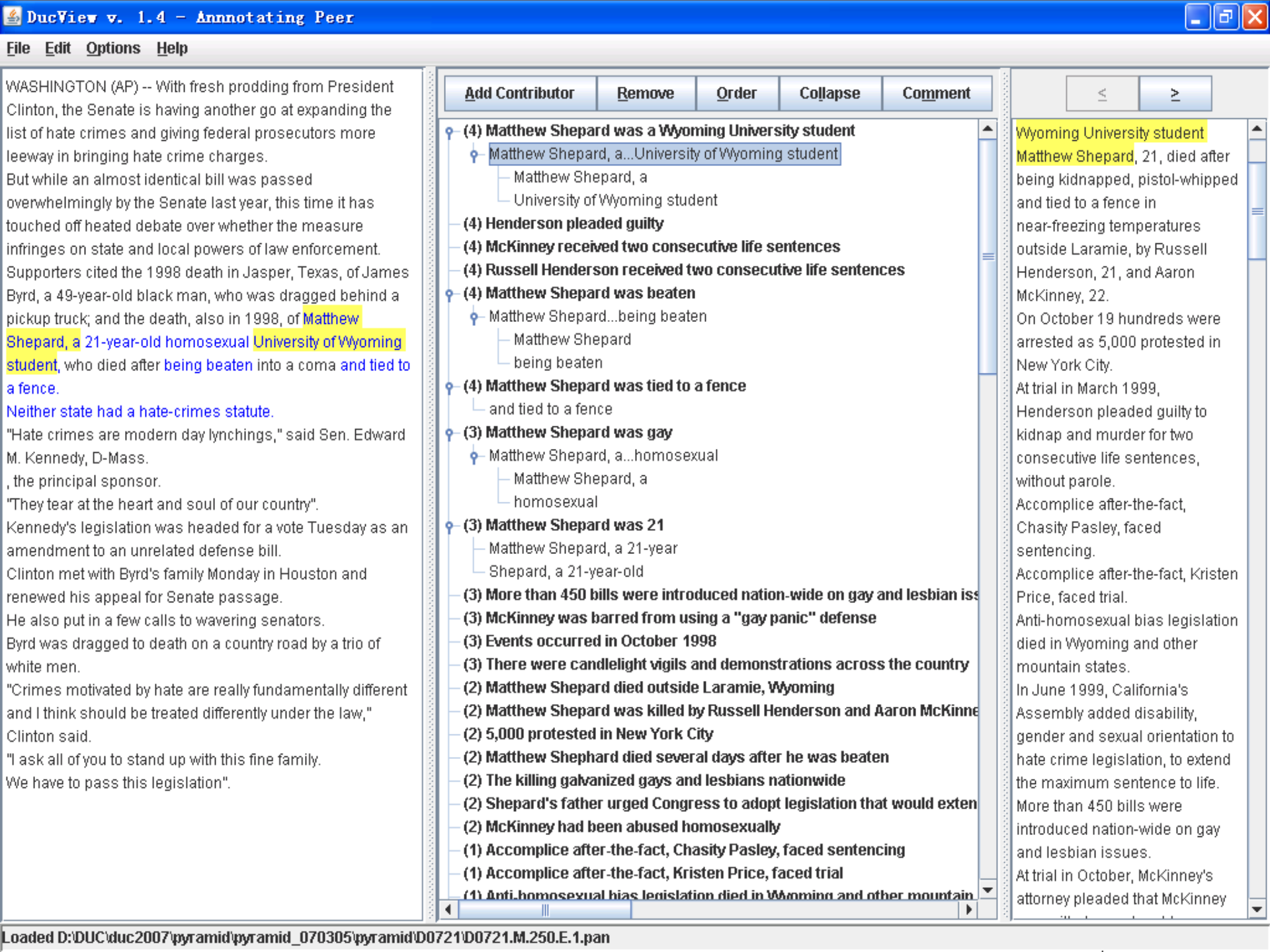
- Where *Nn* represents the set of all *n*-grams and *i* is one member from *Nn*. *Xn(i)* is the number of times the n-gram *i* occurred in the summary and *Mn(i,j)* is the number of times the n-gram *i* ocurred in the *j*-th model reference(human) summary. There are totally *h* human summaries.

Pyramid(1)

- The pyramid content annotation and evaluation method is an approach designed to capitalize on an observation: summaries from different humans always **have partly overlapping content**.
- The pyramid method includes a manual annotation method to represent **Summary Content Units** (SCUs) and to quantify the proportion of model summaries that express this content.
- All SCUs have a weight representing the number of models they occur in, thus from 1 to max_n, where max_n is the total number of models
  - There are very few SCUs expressed in all models (i.e., weight=maxn), and increasingly many SCUs at each lower weight, with the most SCUs at weight=1.

Pyramid(2)

- The approach involves two phases of manual annotation:
  - pyramid construction
  - annotation against the pyramid to determine which SCUs in the pyramid have been expressed in the peer summary.


DucView v. 1.4 - Annnotating Peer
File Edit Options Help
WASHINGTON (AP) -- With fresh prodding from President Clinton, the Senate is having another go at expanding the list of hate crimes and giving federal prosecutors more leeway in bringing hate crime charges.
But while an almost identical bill was passed overwhelmingly by the Senate last year, this time it has touched off heated debate over whether the measure infringes on state and local powers of law enforcement.
Supporters cited the 1998 death in Jasper, Texas, of James Byrd, a 49-year-old black man, who was dragged behind a pickup truck; and the death, also in 1998, of Matthew Shepard, a 21-year-old homosexual University of Wyoming student, who died after being beaten into a coma and tied to a fence.
Neither state had a hate-crimes statute.
"Hate crimes are modern day lynchings," said Sen. Edward M. Kennedy, D-Mass.
, the principal sponsor.
"They tear at the heart and soul of our country".
Kennedy's legislation was headed for a vote Tuesday as an amendment to an unrelated defense bill.
Clinton met with Byrd's family Monday in Houston and renewed his appeal for Senate passage.
He also put in a few calls to wavering senators.
Byrd was dragged to death on a country road by a trio of white men.
"Crimes motivated by hate are really fundamentally different and I think should be treated differently under the law," Clinton said.
"I ask all of you to stand up with this fine family.
We have to pass this legislation".
Add Contributor
Remove
Order
Collapse
Comment
(4) Matthew Shepard was a Wyoming University student
Matthew Shepard, a...University of Wyoming student
Matthew Shepard, a
University of Wyoming student
(4) Henderson pleaded guilty
(4) McKinney received two consecutive life sentences
(4) Russell Henderson received two consecutive life sentences
(4) Matthew Shepard was beaten
Matthew Shepard...being beaten
Matthew Shepard
being beaten
(4) Matthew Shepard was tied to a fence
and tied to a fence
(3) Matthew Shepard was gay
Matthew Shepard, a...homosexual
Matthew Shepard, a
homosexual
(3) Matthew Shepard was 21
Matthew Shepard, a 21-year
Shepard, a 21-year-old
(3) More than 450 bills were introduced nation-wide on gay and lesbian iss
(3) McKinney was barred from using a "gay panic" defense
(3) Events occurred in October 1998
(3) There were candlelight vigils and demonstrations across the country
(2) Matthew Shepard died outside Laramie, Wyoming
(2) Matthew Shepard was killed by Russell Henderson and Aaron McKinne
(2) 5,000 protested in New York City
(2) Matthew Shephard died several days after he was beaten
(2) The killing galvanized gays and lesbians nationwide
(2) Shepard's father urged Congress to adopt legislation that would exten
(2) McKinney had been abused homosexually
(1) Accomplice after-the-fact, Chasity Pasley, faced sentencing
(1) Accomplice after-the-fact, Kristen Price, faced trial
(1) Anti-homosexual bias legislation died in Wyoming and other mountain
≤
≥
Wyoming University student Matthew Shepard, 21, died after being kidnapped, pistol-whipped and tied to a fence in near-freezing temperatures outside Laramie, by Russell Henderson, 21, and Aaron McKinney, 22.
On October 19 hundreds were arrested as 5,000 protested in New York City.
At trial in March 1999, Henderson pleaded guilty to kidnap and murder for two consecutive life sentences, without parole.
Accomplice after-the-fact, Chasity Pasley, faced sentencing.
Accomplice after-the-fact, Kristen Price, faced trial.
Anti-homosexual bias legislation died in Wyoming and other mountain states.
In June 1999, California's Assembly added disability, gender and sexual orientation to hate crime legislation, to extend the maximum sentence to life.
More than 450 bills were introduced nation-wide on gay and lesbian issues.
At trial in October, McKinney's attorney pleaded that McKinney
Loaded D:\DUC\duc2007\pyramid\pyramid_070305\pyramid\D0721\D0721.M.250.E.1.pan

Part 2:
基于主题查询的多文档摘要任务

- 任务描述
- 自动文摘系统实现
  - 特征驱动的文摘系统设计
  - 基于机器学习算法的改进
  - 摘要系统后处理算法介绍
- 实验设计和结果
- 总结

任务描述

- 对由一个主题查询和25篇相关文档组成的新闻主题给出不超过250字的摘要，综合了文档摘要和自动问答的需求
- 新闻主题搜集了近年发生的世界重大事件，新闻文档语料来自美联社、纽约时报和新华社
- ROUGE、Pyramid等自动评测方法和人工评测方法，对每个主题给出了4个人工摘要用于评测

DUC新闻主题样例(一)

- 每个新闻主题由主题查询和25篇相关新闻文档组成
- 主题查询样例：

<topic>
<num> D0601A </num>
<title> Native American Reservation System - pros and cons </title>
<narr>
Discuss conditions on American Indian reservations or among Native American communities. Include the benefits and drawbacks of the reservation system. Include legal privileges and problems.
</narr>
</topic>

DUC新闻主题样例(二)

新闻文档样例：

<DOC>
<DOCNO> APW20000416.0024 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 2000-04-16 12:56 </DATE_TIME>
<HEADLINE> Clinton To Visit Navajo Community </HEADLINE>
By CHRIS ROBERTS, Associated Press Writer
<TEXT>
<P>ALBUQUERQUE, N.M. (AP) -- At Mesa Elementary School in far northwest New Mexico, Navajo children line up to use the few computers connected to the Internet. Their time online must be short for everyone to get a chance.
</P>
</P>
… …
… …
<P>
Navajo Nation: http://www.navajo.org/nnhomepg.html
</P>
</TEXT>
</DOC>

基于特征的抽取文摘技术

- 通过设计一系列基于词法、句法、语义的特征，构建了基于句子抽取的自动摘要系统
- 引入机器学习算法计算特征权重来改进摘要系统，自动生成训练语料来解决语料匮乏的问题
- 后处理算法设计，包括冗余处理、句子精简、句子重排序等

系统流程

- 预处理
  - 将原始文档转化为后续步骤所需的输入格式。
- 句子打分
  - 估算句子的重要度以决定抽取哪些句子来生成摘要。
- 后处理
  - 对抽取出的句子结果进行句子精简、冗余处理、重排序等后续处理来提高摘要可读性和紧凑性。

预处理模块

- 输入：DUC语料中的XML格式文档
- 处理流程
  1) 分析原始XML文档提取出标题和正文部分
  2) 去除标题和正文中多余的XML标签得到其纯文本
  3) 用基于规则的算法对正文进行切句，得到独立的句子
- 输出：切好句的文档

预处理结果样例

<?xml version='1.0' encoding='UTF-8'?>

<!DOCTYPE DOCSENT SYSTEM "/home/lsjian/mead/dtd/docsent.dtd">

<DOCSENT DID='APW20000416.0024' DOCNO='APW20000416.0024' LANG='ENG' CORR-DOC='FT.c'>

<BODY>

<HEADLINE>

<S PAR='1' RSNT='1' SNO='1'> Clinton To Visit Navajo Community </S>

</HEADLINE>

<TEXT>

<S PAR='2' RSNT='1' SNO='2'> ALBUQUERQUE, N.M. </S>

<S PAR='2' RSNT='2' SNO='3'> At Mesa Elementary School in far northwest New Mexico, Navajo children line up to use the few computers connected to the Internet. </S>

… …

<S PAR='2' RSNT='27' SNO='28'> Clinton's plans Monday including speaking at Shiprock Boys/Girls Club, then joining an evening Webcast at Dine Tribal College in Shiprock that will involve high school student online at Lake Valley Navajo School, about 55 miles away. ------ On the Net: Navajo Nation: http://www.navajo.org/nnhomepg.html </S>

</TEXT></BODY></DOCSENT>

句子打分模块

- 特征驱动
  - 从词法、句法、语义三个层次上进行分析，设计一系列特征来估计句子的重要度得分
  - 特征又可分为查询相关和查询无关两类
- 线性加权打分函数
  - 用线性函数来加权所有特征进行打分
  - 人工指定权重，领域知识+经验尝试

特征设计(一)：查询相关的特征

- 查询相关的特征刻画句子与主题查询之间的相关度 (自动问答技术)
- 特征设计
  - 基于词频的特征：考虑句子和查询间的词汇重叠度
  - 基于命名实体的特征：考虑句子和查询之间的命名实体的重叠度
  - 基于WordNet的特征：考虑句子和查询之间的语义相关度

特征设计(二)：查询无关的特征

- 查询无关的特征刻画句子在文档集合中的重要度 (文档摘要技术)
- 特征设计
  - Centroid特征：
  - 基于TfIdf的特征：句子中所有词的TfIdf之和
  - 基于实体的特征：句子中的命名实体的个数
  - 基于停用词的特征：句子中的停用词的个数
  - 基于位置的特征：句子在全文中的位置
  - 基于长度的特征：句子的总词数

Centroid

- Centroid-based summarization: use as input the centroids of the clusters to **identify which sentences are central to the topic of the cluster, rather than the individual articles**.
- Centroids consist of words which are central not only to *one* article in a cluster, but to *all* the articles.
- A centroid is a pseudo-document which consists of words which have Count*IDF scores above a predefined threshold in the documents that constitute the cluster.
  - Count: average number of occurrences of a word across the entire cluster
  - IDF: computed from a large corpus.
- Hypothesize that sentences that contain the words from the centroid are more indicative of the topic of the cluster.

用回归模型训练权重

- 人工权重的缺点
  - 无法保证权重的最优性
  - 权重搜索的时间复杂度指数级上升
- 解决方法
  - 句子打分过程就是寻找特征向量与句子得分之间的映射的过程
  - 将句子打分问题看作回归问题，利用回归模型来训练权重

基于回归模型的句子打分方法

- 打分流程
  - 用给出了句子得分的训练语料 $\{(score(s), V(s)) \mid s \in D\}$ 来训练回归函数
  - 用训练出的回归函数 $f_0 : V(s) \rightarrow score(s)$ 对新的句子得分进行估计

$$\hat{score}(s^*) = f_0(V(s^*))$$

- 关键问题
  - 模型选择：支持向量回归(SVR)模型
  - 训练语料的匮乏：基于人工摘要自动生成

自动生成训练语料(一)

- 指导思想
  - 利用DUC提供的人工摘要来得到近似的训练语料
  - 计算人工摘要与句子之间的相似度来作为“真实”句子得分
- 计算过程

1) 计算单个N-gram在单个人工摘要下的重要度

$$p(t|H_i) = freq(t)/|H_i| \quad \text{或} \quad p(t|H_i) = appr(t)/|H_i|$$

2) 计算单个N-gram在所有人工摘要下的重要度

$$p_{Max}(t|H) = \max_{H_i \in H}(p(t|H_i)) \quad \text{或} \quad p_{Avg}(t|H) = \frac{1}{|H|}\sum_{H_i \in H} p(t|H_i)$$

3) 计算句子在所有人工摘要下的重要度

$$score(s|H) = \sum_{t_j \in s} p(t_j|H)$$

自动生成训练语料(二)

- 计算过程

Bigram

Unigram

词频

出现

1) 计算单个N-gram在单个人工摘要下的重要度

$$p(t|H_i) = freq(t)/|H_i| \quad \text{或} \quad p(t|H_i) = appr(t)/|H_i|$$

最大化

平均值

2) 计算单个N-gram在所有人工摘要下的重要度

$$p_{Max}(t|H) = \max_{H_i \in H}(p(t|H_i)) \quad \text{或} \quad p_{Avg}(t|H) = \frac{1}{|H|}\sum_{H_i \in H} p(t|H_i)$$

3) 计算句子在人工摘要下的重要度

$$score(s|H) = \sum_{t_j \in s} p(t_j|H)$$

自动生成训练语料(三)

- 8种不同的自动生成训练语料的方法
  - N-gram类型
    - Unigram 或 Bigram
  - 统计类型
    - 频率统计 或 出现统计
  - 整合所有人工摘要的策略
    - Maximum 或 Average

后处理模块

- 简单的后处理模块
  - 将句子按照得分由高至低排列，依次抽取高分句子直到总字数达到目标要求
- 可能存在的问题
  - 问题1：句子之间的冗余
  - 问题2：句子中的无意义词汇
  - 问题3：摘要的通顺

句子冗余处理

- 由于句子打分算法中的特征大部分基于词汇计算，所以词汇上相似的句子的特征值相近，那么得分也就相近，所以相似的句子常被同时抽取，如主题D0623中的
  - “The top fine for smoking at work or in public places would be 200 rand (dlrs 35). ”
  - “The maximum fine for smoking at work or in public places would be 200 rand (dlrs 35).
- 基于MMR(最大边缘相关度)的句子冗余处理
  - 每次选择新句子时，首先计算当前句子与已抽取句子之间的相似度，如果超过了阈值，则认为其为冗余句子，不予抽取
  - 相似度阈值为句子向量的夹角余弦，阈值为0.7

- MMR:
  - The Maximal Marginal Relevance (MMR) criterion strives to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passage for text summarization.

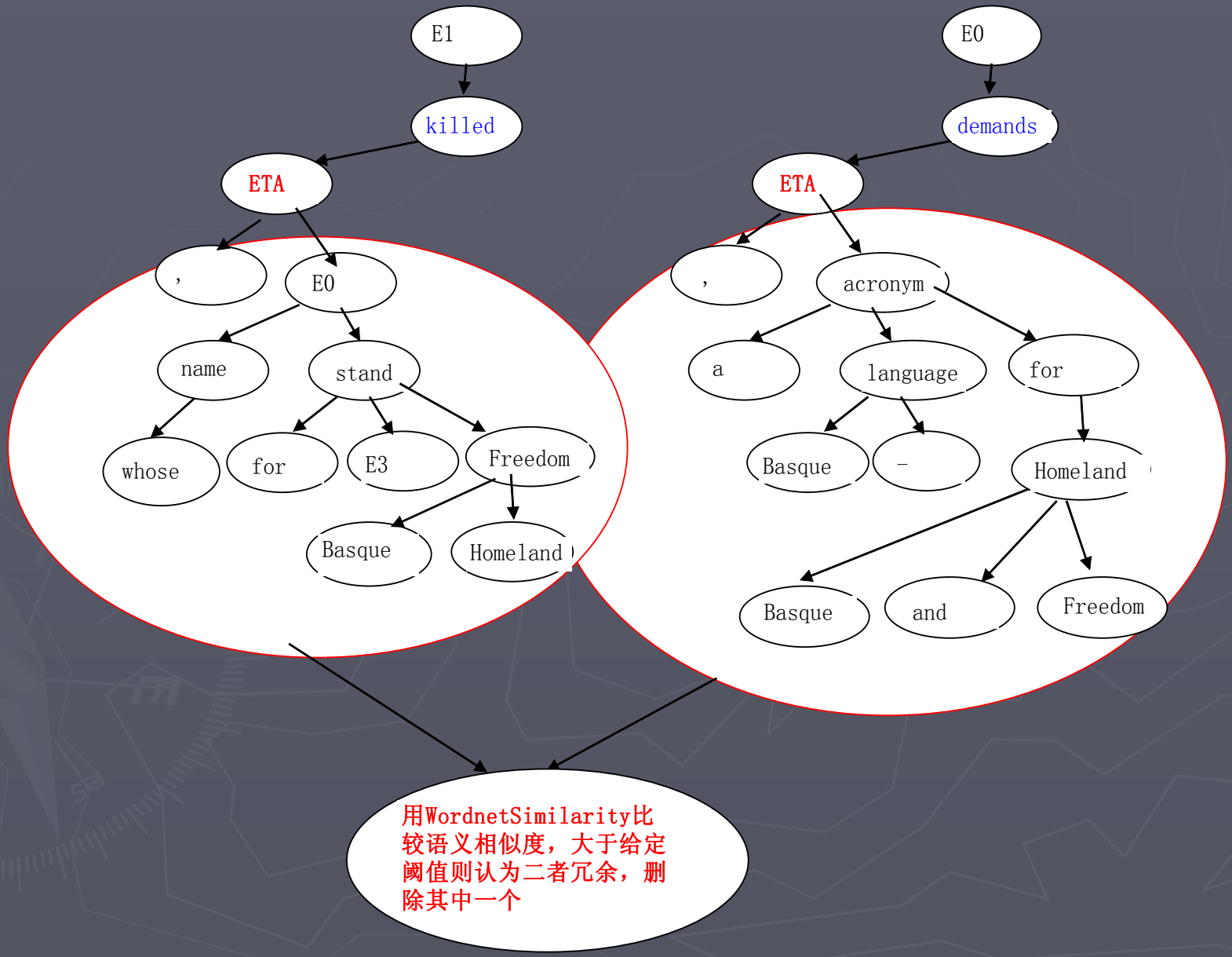
$$MMR \overset{def}{=} Arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j))]$$

句子精简

- 通过去除新闻文档中包括的无意义词汇来提高系统性能
  - 新闻的噪音词汇
  - 内容无关的词汇
- 基于规则的句子精简方法
  - 去除如“ALBUQUERQUE, N.M. (AP) --”形式的新闻开头；
  - 去除句首的“and”，“also”，“besides，”，“though，”，“in addition”，“somebody(代词) said”，“somebody(代词)says”等词汇；
  - 去除句首的“somebody (代词)/It is said/reported/noticed/thought that”等词汇；
  - 内容全是大写字母的括号部分
  - …

文本简化

- 基于句法分析的文本简化：
  - 去掉冗余信息，无关的背景信息等，应用于后处理部分
  - 使用句法分析器分析句子的句法结构，得到该句子的依存树。
  - 具有相同主语的句子，记录其主语同位语，补语等主语修饰成分保存在一张表中。计算其语义相似度，只有相似度小于给定阈值的成分才会出现在摘要中，并且只出现一次。


E1
killed
ETA
,
E0
name
stand
whose
for
E3
Freedom
Basque
Homeland
E0
demands
ETA
,
acronym
a
language
for
Basque
-
Homeland
Basque
and
Freedom
用WordnetSimilarity比较语义相似度，大于给定阈值则认为二者冗余，删除其中一个

句子重排序

- 依据分数高低依次抽取出来的句子组成的摘要的句子组织缺少逻辑性，摘要可读性较低

- 基于时间信息的句子重排序
  - DUC的XML文档中已包含文档时间信息，分析XML标签获得文档时间信息
  - 按照所在文档的时间先后次序来排列句子；对于同一文档中的句子，按照在文档中出现的先后排列

实验环境介绍

- 语料：DUC2005、2006、2007三年的语料，每年50个新闻主题，每个主题4个不同的人工摘要结果
- 评测方法：ROUGE评测方法，基于待评测摘要和人工摘要的匹配度计算
  - 基于比较和匹配待评测摘要和人工摘要的内容的一种评测方法
- 评测指标：ROUGE-1、2、SU4的召回率的平均值和95%置信区间
  - 只使用了召回率而未使用准确率的原因是：DUC任务中限定了摘要的总字数不能超过250字，任务目标是在250字的摘要中尽可能地多包含人工摘要中的信息。

实验一：训练语料的合理性

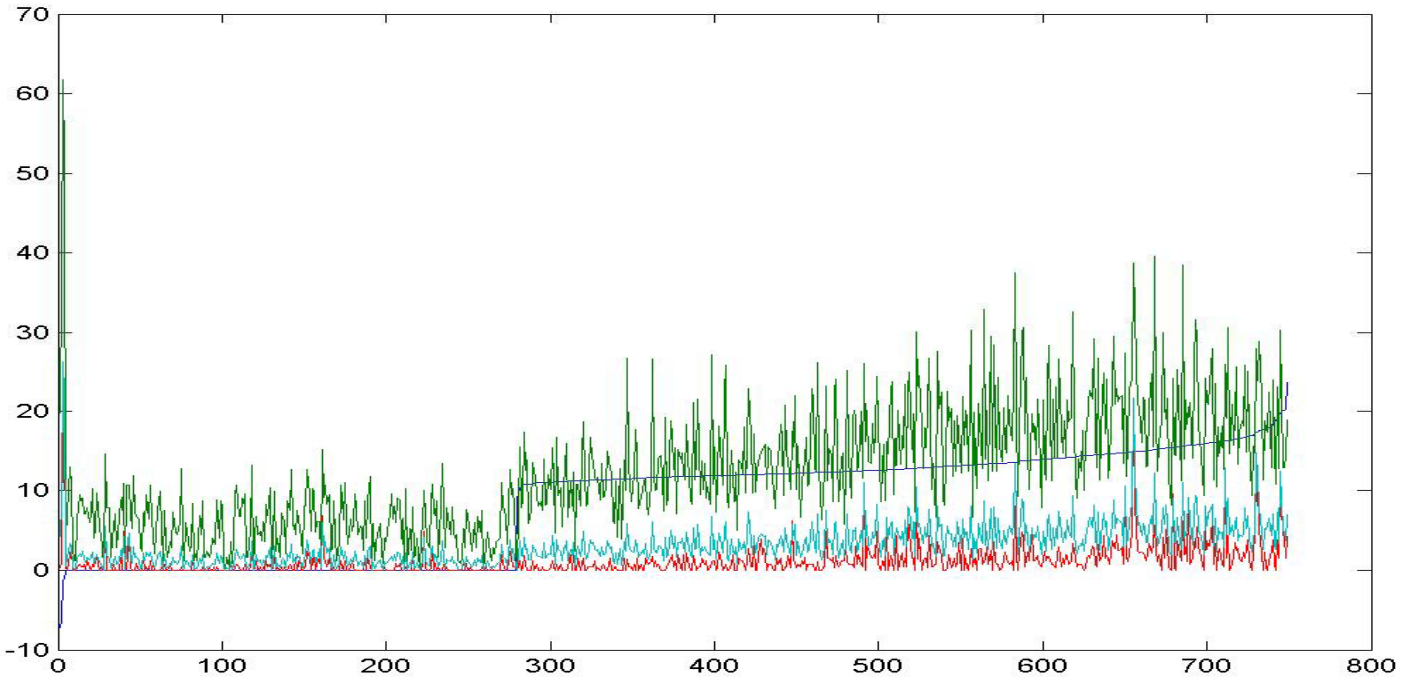
- 实验条件
  - 用基于人工摘要估计出的“真实”句子得分来指导句子抽取
- 实验结果
  - 得到的摘要结果在ROUGE上的得分超过了人工摘要本身
- 结论
  - 基于人工摘要的句子打分方法的合理性
  - 基于句子抽取的摘要方法的潜力

实验一：训练语料实验结果

Submission	Average Rouge-2 and CI	Submission	Average Rouge-2 and CI
Bi+Appr+Max	**0.1711** (0.1608, 0.1830)	Uni+Appr+Avg	0.1468 (0.1358, 0.1584)
Bi+Appr+Avg	0.1666 (0.1563, 0.1772)	Best human Summary	**0.1326** (0.1160, 0.1520)
Uni+Appr+Max	0.1603 (0.1503, 0.1710)	Uni+Freq+Avg	0.1149 (0.1058, 0.1245)

相关性分析

我们将生成摘要中的每个句子用Rouge进行打分，并将打分与我们的系统打分进行比较，研究表明系统打分与Rouge得分较为相关



实验二：特征驱动的摘要系统

- 实验条件
  - 特征驱动的摘要系统，采用了人工指定权重的线性函数打分方法，由不同的人给出多组权重
- 实验结果
  - 摘要结果远高于Baseline，具有较好的性能
  - 不同的人工权重性能相差较大

实验二：特征驱动的摘要系统

Submission	Average Rouge-1 (CI)	Average Rouge-2 (CI)	Average Rouge-SU4(CI)
Best human summary	0.4582 (0.4496, 0.4682)	**0.1036** (0.0926, 0.1162)	0.1683 (0.1604, 0.1773)
Best submitted summary	0.4111 (0.4049, 0.4171)	**0.0956** (0.0914, 0.0998)	0.1553 (0.1513, 0.1591)
Our best summary	0.3838 (0.3769, 0.3904)	**0.0817** (0.0774, 0.0862)	0.1357 (0.1316, 0.1397)
Baseline summary	0.3022 (0.2923, 0.3119)	**0.0495** (0.0456, 0.0538)	0.0979 (0.0934, 0.1021)

实验三：用回归模型训练权重

- 实验条件
  - 用回归模型和自动生成的语料来训练句子打分函数
- 实验结果
  - 2005年语料上超过了所有提交系统，2006年语料上排名第二
  - 用回归模型训练出的权重的性能远远高于人工权重的性能
  - 不同的训练语料生成方法之间的性能差别，总体说来，基于Unigram及Maximum的方法性能较好

实验三：用回归模型训练权重

Submission	Average Rouge-1 (CI)	Average Rouge-2 (CI)	Average Rouge-SU4
…	…	…	…
A	0.4582 (0.4496, 0.4682)	**0.1036** (0.0926, 0.1162)	0.1683 (0.1604, 0.1773)
24	0.4111 (0.4049, 0.4171)	**0.0956** (0.0914, 0.0998)	0.1553 (0.1513, 0.1591)
Uni+Freq+Max	0.4018 (0.3959, 0.4078)	**0.0926** (0.0883, 0.0969)	0.1485 (0.1443, 0.1525)
Human Weights	0.3744 (0.3668, 0.3813)	**0.0751** (0.0709, 0.0792)	0.1299 (0.1257, 0.1340)

实验四：后处理算法对性能的影响

- 实验条件
  - 比较后处理算法前后的摘要结果的性能
- 实验结果
  - 句子精简算法和冗余处理算法提高了系统的ROUGE得分
  - 对于有明显时间跨度的新闻主题，句子重排序算法能按照事件发生的顺序组织摘要，提高了摘要的可读性

实验四：后处理算法对性能的影响

System	Average Rouge-1 (CI)	Average Rouge-2 (CI)	Average Rouge-SU4(CI)
Final System	0.4018 (0.3959, 0.4078)	**0.0926** (0.0883, 0.0969)	0.1485 (0.1443, 0.1525)
Without MMR	0.3990 (0.3930, 0.4050)	**0.0909** (0.0867, 0.0952)	0.1467 (0.1427, 0.1506)
Without Simplification	0.3882 (0.3804, 0.3933)	**0.0883** (0.0827, 0.0943)	0.1376 (0.1339, 0.1426)

系统总结

- 设计了特征驱动的句子打分方法和基于句子抽取的摘要系统框架
- 用回归模型训练打分函数的权重，与人工权重的性能进行了比较
- 设计了自动生成训练语料的算法，比较不同的训练语料生成方法对系统性能的影响
- 有效提高摘要系统性能的简单后处理算法

Part 3

- Graph ranking
- Sentence compression
- Discourse structure
- QA, entity recognition,
- …

Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction

Author: Xiaojun Wan, Jianwu Yang, Jianguo Xiao

文本摘要&关键词提取

- 文本摘要：根据给定的文本生成摘要

  分类：问题相关/问题无关

  单文档/多文档摘要

  …

- 关键词提取：从指定的文本中提取关键词，要求关键词能够反映原文信息

  分类：问题相关/问题无关

  单文档/多文档摘要

  …

文本摘要&关键词提取

- 相同点：

二者的目标都是从原文中提取简练的，有代表性的信息。

- 不同之处：

抽取信息的单位不一样：　句子 / 词

文本摘要需要对摘要结果重组织

Question: 能不能同时完成文本摘要和关键词提取？
两个任务有没有互补性？

Graph-based ranking algorithm

- 被成功地运用在文本摘要和关键词提取

  TextRank LexPageRank
- 原理：根据全局信息(图的结构)而不是局部信息来对节点排序
- Popular Graph-based ranking algorithm
  - Google PageRank
  - HITS

PageRank

- 原理：

  1. 一个网页被其它网页链接的次数越多，则该网页越重要

  2.链接一个网页的网页越重要，则该网页越重要

Simplified Pagerank:

$$R(u) = c\sum_{v \in B_u} \frac{R(v)}{N_v} = cA^T R$$

$$R(u) = c\sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u)$$

PageRank Implementation

- 给定邻接矩阵E，记$|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots$, q_1是属于λ_1的特征向量
- 初始化向量p_0，使得$|p_0|=1$
- 对于k = 1, 2, …,执行如下步骤
  - $x = E^T p_{k-1}$，　　基本迭代
  - $p_k = x/||x||$，　　规格化步骤
- 可以证明（收敛速度）
  - $|p_k - q_1| = O(|\lambda_2/\lambda_1|^k)$

Application of Graph ranking in NLP(1)

- 一个通用的算法：

1.确定完成任务所需要的“文本单元” (如句子，单个的词，短语等)，将这些单元作为图中的顶点。

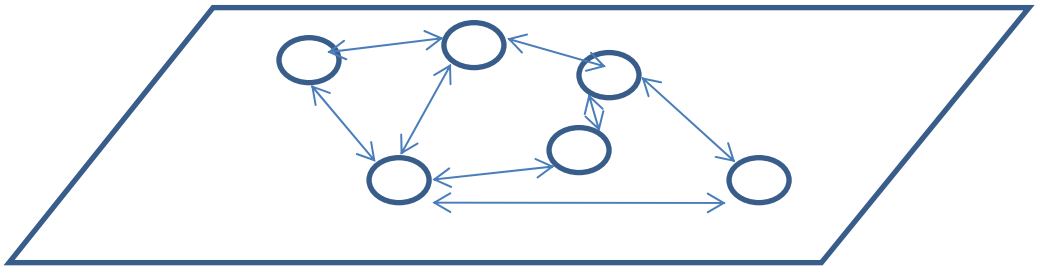
2.确定顶点之间的“关系”。如顶点之间存在某种关系，则在这两个点之间连一条边。边可为有向/无向，带权重/不带权重

3.运用图排序算法进行运算，直到结果收敛于给定阈值

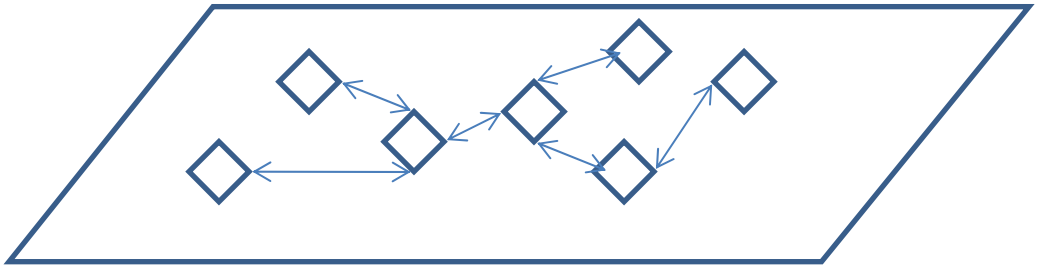
4.根据计算出的得分对“文本 单元”排序，选取高得分的组成结果。

Application of Graph ranking in NLP(2)

- 句子-句子关系图



- 词-词关系图



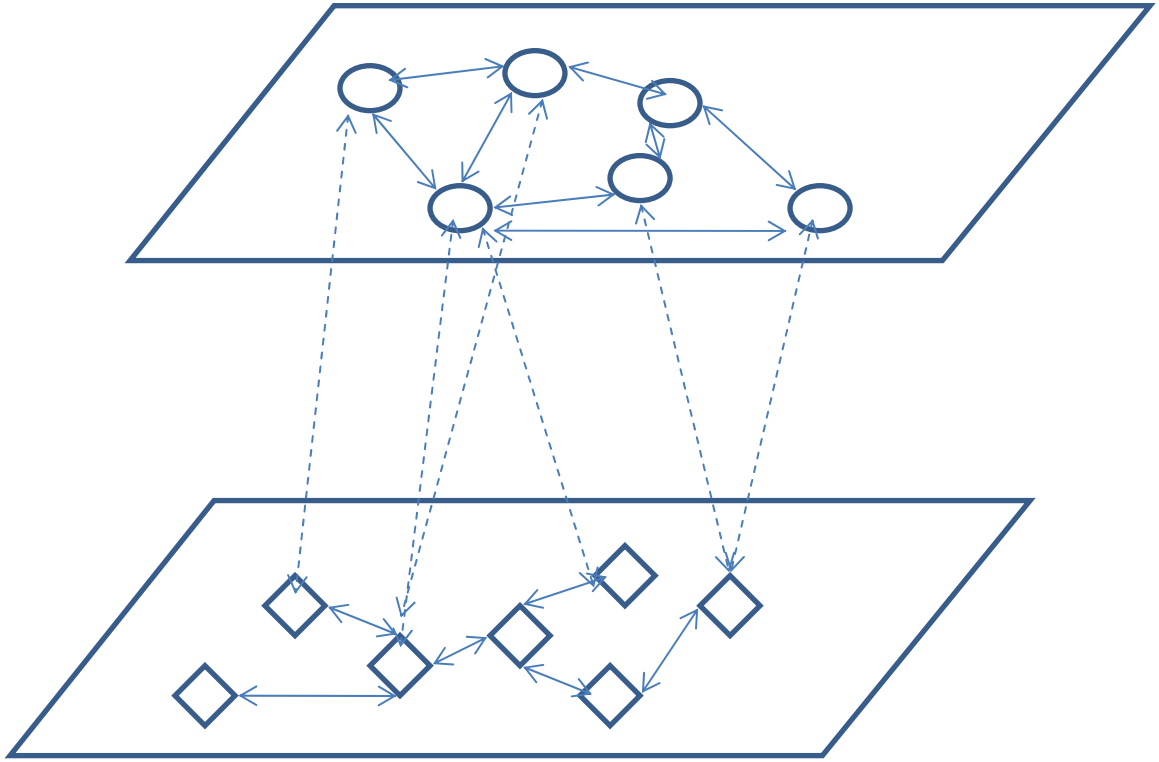
Document Model

Assumption 1

一个句子很重要，那么与该句子有紧密联系的句子也很重要；一个词很重要，那么与该词紧密联系的词也很重要

Assumption 2

一个句子中重要词的出现越多，则该句子越重要，一个词在重要的句子中出现次数越多，则该词越重要



- 三种关系
  - SS-Relationship
  - WW-Relationship
  - SW-Relationship

Build Sentence-Sentence Graph

- 句子之间的关系——句子相似度
- 如果两个句子之间的相似度大于0，那么在代表这两个句子的节点之间连一条带权重的无向边，把相似度作为权重
- 句子相似度计算

$$U_{ij} = sim(t_i, t_j), i \neq j$$

$$U_{ij} = 0, i = j$$

- $\vec{S_i}$ 是句子 S_i 的term vector

$$s_i =< wt_1, wt_2, ...wt_n >$$

wt_i 为term t_i 的权重，isf_{t_i} 为 t_i 的倒排句子频率

$$wt_i = tf_{t_i} * isf_{t_i}$$

$$isf_{t_i} = 1 + \log(N / n_{t_i})$$

- 这样我们得到句子-句子图的矩阵表示

$$U = [U_{ij}]_{n \times n}$$

- 其中，$U_{ij} = sim(s_i, s_j), i \neq j$

$$U_{ij} = 0, i = j$$

然后将该矩阵正则化

Build Word-Word Graph

- 词之间的关系——词相似度
- 如果两个词之间的相似度大于0，那么在这代表这两个词之间的点之间连一条带权重的无向边，把相似度作为权重
- 词相似度计算：两种方法
  - 基于知识库
  - 基于语料

基于知识库的词相似度计算

- 利用Semantic Networks（语义网）中的信息计算相似度
- WordNet

  a large lexical database of English

  Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

  Each synset has a gloss that defines the concept that it represents.

  - 利用网状结构信息（如点之间的距离，点的稀疏程度）来计算相似度

基于语料库的词相似度计算

- 利用互信息

$$sim(t_i, t_j) = \log \frac{N \times p(t_i, t_j)}{p(t_i) \times p(t_j)}$$

$p(t_j)$ t_i 在语料中出现的概率

$p(t_j)$ t_j 在语料中出现的概率

$p(t_i, t_j)$ 二者共同出现的概率（二者间隔的距离不超过一个给定值即认为共同出现）

- 这样我们得到Word-Word图的矩阵表示

$$V = [V_{ij}]_{n \times n}$$

其中 $V_{ij} = sim(t_i, t_j), i \neq j$

$V_{ij} = 0, i = j$

然后将该矩阵正则化

Build Sentence-Word Graph

- 句子与词之间的关系
- 不是“同源”的关系，因此不能用相似度来计算

  如果一个词在某个句子中出现，则我们认为它们之间存在关系

$$T = \{t_j \mid 1 \le j \le n\} \quad S = \{s_i \mid 1 \le i \le m\}$$

$$aff(s_i, t_j) = \frac{tf_{t_j} \times isf_{t_j}}{\sum_{t \in S_i} tf_t \times isf_t}$$

- 这样我们得到句子-词图的矩阵表示

$$W = [W_{ij}]_{m \times n}$$

其中 $W_{ij} = aff(s_i, t_j)$

然后将该矩阵正则化

Reinforcement Algorithm

- 我们用列向量 $u=[u(s_i)]_{mX1}$ 和$v=[v(t_j)]_{nX1}$ 分别表示句子和词的得分

根据前面两个assumptions,

$$u(s_i) \propto \sum_j U_{ji} u(s_j)$$

$$v(t_j) \propto \sum_i V_{ij} v(t_i)$$

$$u(s_i) \propto \sum_j W_{ji} v(t_j)$$

$$v(t_j) \propto \sum_i W_{ij} u(s_i)$$

$$u(s_i) = \alpha \sum_{j=1}^{m} U_{ji} u(s_j) + \beta \sum_{j=1}^{n} W_{ji} v(t_j)$$

$$v(t_j) = \alpha \sum_{i=1}^{n} V_{ij} v(t_i) + \beta \sum_{i=1}^{m} W_{ij} u(s_i)$$

matrix form:

$$u = aU^T u + \beta W \; v$$

$$v = aV^T v + \beta W^T u$$

总结：

图排序算法可以很好地被运用在文本摘要，关键词提取中。

文本摘要和关键词提取这两个任务因其相似性，可以同时完成，并且二者可以互相促进

- 一些待讨论的问题

  上面讨论的是问题无关，单文档摘要，对于问题有关，多文档摘要，如何处理？

  问题有关：问题包含的句子作为一个节点，是否应该和图中其他节点平等看待？

  多文档摘要：如何将不同文档的词，句子区分开来？

K&M Model

Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139, 91–107.

The K&M model(1)

- Sentence compression is a task of creating a short grammatical sentence by removing extraneous words or phrases from an original sentence while preserving its meaning.
- K&M model presents a noisy-channel model for sentence compression and learns statistics on trimming context-free grammar rules.

The K&M model(2)

- In the sentence compression model, the short string is the original sentence and someone adds noise, resulting in the longer sentence.
- Using this framework, the end goal is, given a long sentence *l*, to determine the short sentence *s* that maximizes *P(s|l)*.

$$P(s \mid l) = \frac{P(l \mid s)P(s)}{P(l)}$$

- *P(s)* is the source model: the probability that s is the original sentence. *P(l|s)* is channel model: the probability the long sentence is the expanded version of the short. This framework independently models the grammaticality of *s* and whether *s* is a good compression of *l*.

The K&M model(3)

- The K&M model uses parse trees for the sentences. These allow it to better determine the probability of the short sentence and to obtain alignments from the training data.
- In the K&M model, the sentence probability is determined by combining a probabilistic context free grammar. The joint rules used to create the compression are generated by aligning the nodes of the short and long trees in the training data to determine expansion probabilities $P(l/s)$.

The K&M model(4)

- It obtains these probabilities by aligning nodes in the parsed parallel training corpus, and counting the nodes that align as "joint events." For example, there might be *S -> NP VP PP* in the long sentence and *S -> NP VP* in the short sentence; we count this as one joint event. Non-compressions, where the long version is the same as the short, are also counted. The expansion probability, as used in the channel model, is given by

$$P_{\exp and}(l \mid s) = \frac{count(joint(l, s))}{count(s)}$$

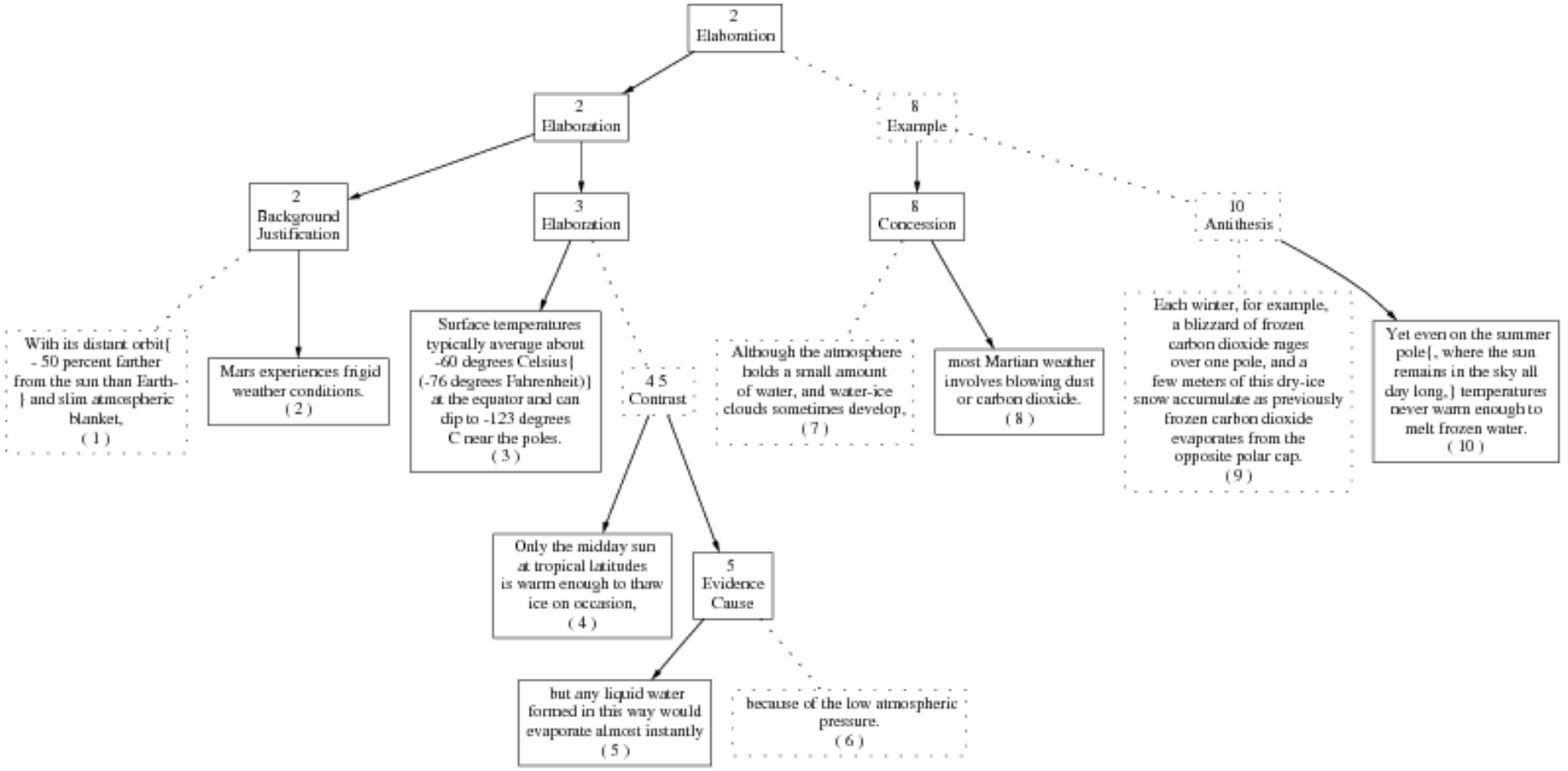
- The K&M model creates a packed parse forest of all possible compressions that are grammatical with respect to the Penn Treebank.

Discourse structure

- RST (rhetorical structure theory)

  Marcu, D. "*Discourse trees are good indicators of importance in text*" In Advances in Automatic Text Summarization, pages 123-136, 1999.

  - The mapping between discourse structures and importance scores can be used effectively for determining the most important units in a text.
  - There is a strong correlation between the nuclei of a discourse structure of a text and what readers perceive to be the most important units in a text.
- CT (centering theory)


2
Elaboration
2
Elaboration
8
Example
2
Background
Justification
3
Elaboration
8
Concession
10
Antithesis
With its distant orbit{
- 50 percent farther
from the sun than Earth-
} and slim atmospheric
blanket,
(1)
Mars experiences frigid
weather conditions.
(2)
Surface temperatures
typically average about
-60 degrees Celsius{
(-76 degrees Fahrenheit)}
at the equator and can
dip to -123 degrees
C near the poles.
(3)
4 5
Contrast
Although the atmosphere
holds a small amount
of water, and water-ice
clouds sometimes develop,
(7)
most Martian weather
involves blowing dust
or carbon dioxide.
(8)
Each winter, for example,
a blizzard of frozen
carbon dioxide rages
over one pole, and a
few meters of this dry-ice
snow accumulate as previously
frozen carbon dioxide
evaporates from the
opposite polar cap.
(9)
Yet even on the summer
pole{, where the sun
remains in the sky all
day long,} temperatures
never warm enough to
melt frozen water.
(10)
Only the midday sun
at tropical latitudes
is warm enough to thaw
ice on occasion,
(4)
5
Evidence
Cause
but any liquid water
formed in this way would
evaporate almost instantly
(5)
because of the low atmospheric
pressure.
(6)

Reference

- Karen Sparck Jones, **automatic summarising: the state of the art**, **Information Processing & Management, 43(6): 1449--1481, 2007.**
- **Lin C. Y. , Hovy E., Manual and automatic evaluation of summaries, Proceedings of the Workshop on Automatic Summarization (including DUC 2002), Philadelphia, July 2002, pp. 45-51. Association for Computational Linguistics.**
- **http://haydn.isi.edu/ROUGE/**
- **Passonneau, R.J. et. al (2005), applying the Pyramid method in DUC 2005, DUC 2005, 25-32.**
- **Jaime Carbonell, Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 1998.**
- **Radev, D.R., Jing, H. and Budzikowska, M. (2000) 'Centroid-based summarisation of mul-tiple documents: sentence extraction, utility-based evaluation, and user studies', ANLP/NAACL-00, 2000, 21-30.**
- **J.M. Conroy, et. al., Topic-focused multi-document summarization using an approximate oracle score, ACL 2006.**

Conclusions

- Have confidence to construct a simple summarization system?
- Which research points you are interested in?

Thanks & QA