第9章 文本自动摘要

Long Article
Summarized
before
after

主要内容

- 文本摘要概述
- 文本摘要分类
- 文本摘要方法
- 文本摘要评价与评测

文本摘要概述

- 文本摘要的定义
- 文本摘要的需求
- 文本摘要的发展简史
- 代表系统

文本摘要的定义

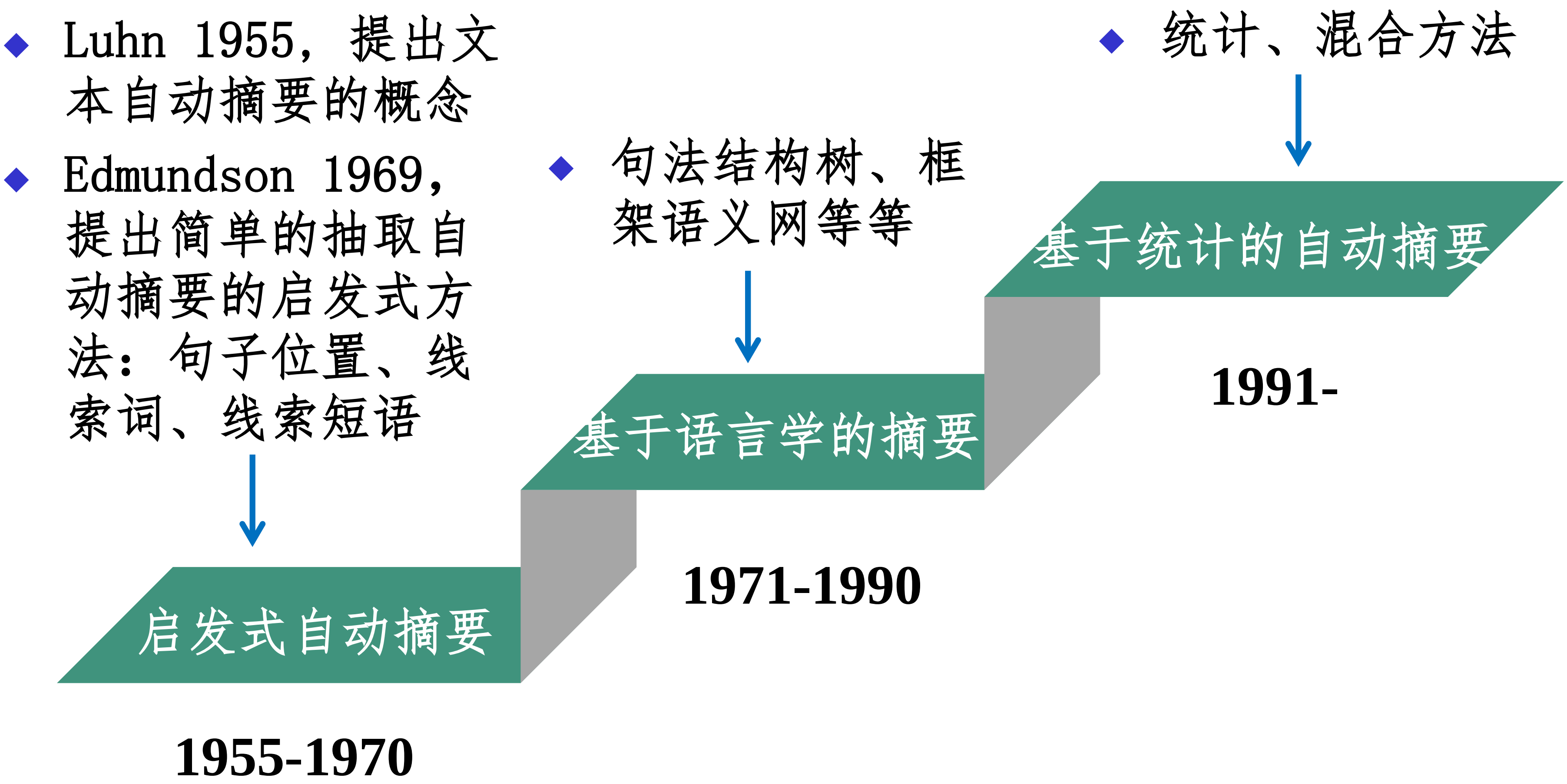
- **定义：**
  - 文本自动摘要是利用计算机按照某类应用自动地将文本（或文本集合）转换生成简短摘要的一种信息压缩技术

- **要求：**
  - 信息量足、覆盖面广、冗余度低和可读性高

文本摘要的需求

- 快速的信息获取
  - 搜索引擎
  - 一句话标题
  - 科技文献摘要
- 适应特定的环境
  - 广告
  - 手机、平板等屏幕受限的设备

文本摘要的发展简史



- Luhn 1955，提出文本自动摘要的概念
- Edmundson 1969，提出简单的抽取自动摘要的启发式方法：句子位置、线索词、线索短语

启发式自动摘要

1955-1970

- 句法结构树、框架语义网等等

基于语言学的摘要

1971-1990

- 统计、混合方法

基于统计的自动摘要

1991-

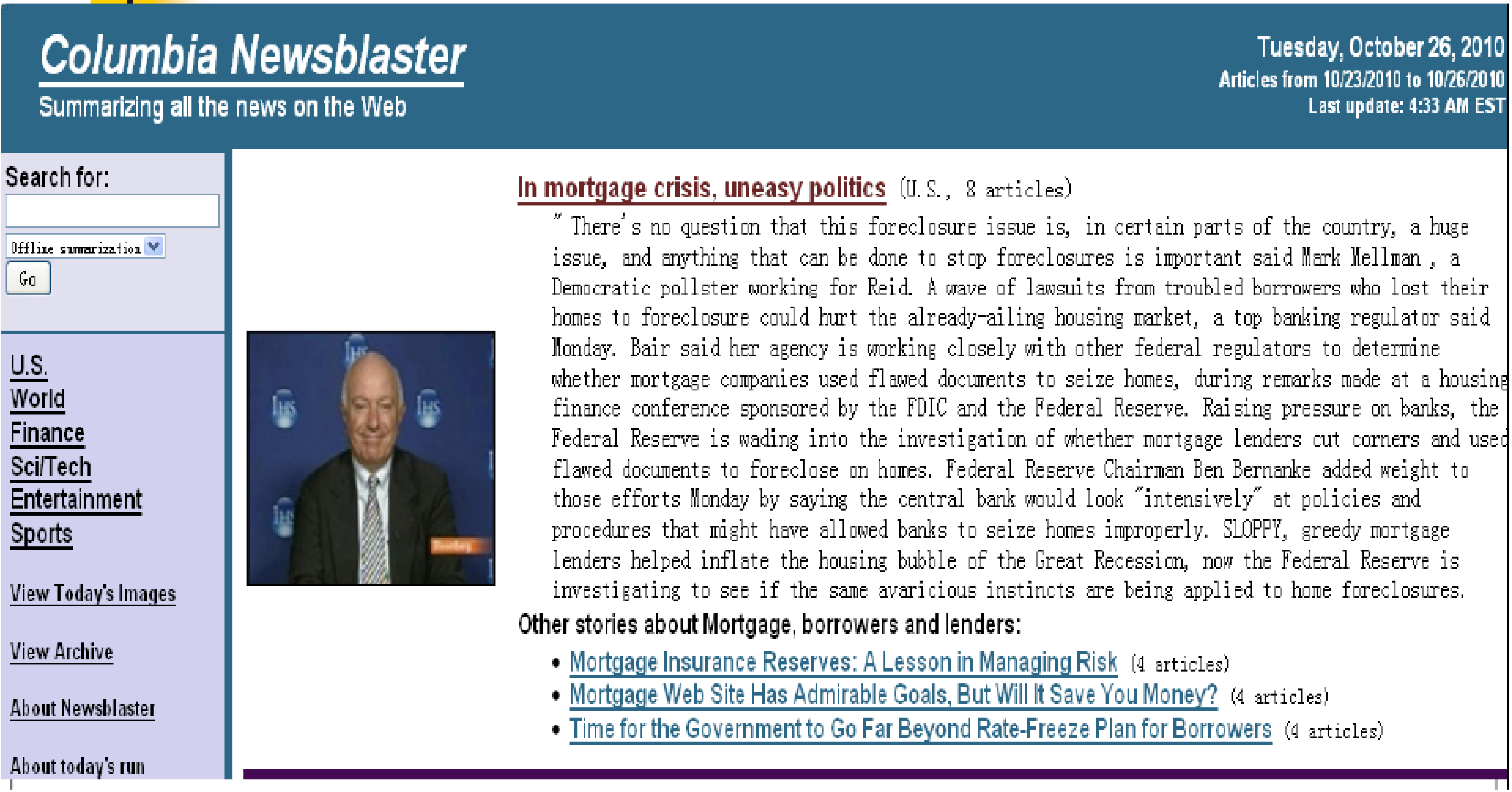
代表系统

- **NewsInEssence**
  - 密歇根大学开发
  - Dragomir Radev et al., 2005. Newsinessence: Summarizing online news topics, *ACM Communications*
- **NewsBlaster**
  - 哥伦比亚大学开发
  - David Kirk Evans et al., 2004. Columbia Newsblaster: Multilingual News Summarization on the Web, *In HLT-NAACL.*

NewsInEssence

- 抽取式方法的代表系统
  - 利用词汇计算文本中心
  - 抽取与文本中心临近的句子作为摘要
- 应用范围
  - 单文档摘要
  - 多文档摘要
  - 基于用户查询的摘要
  - …

NewsBlaster



Columbia Newsblaster

Summarizing all the news on the Web

Tuesday, October 26, 2010

Articles from 10/23/2010 to 10/26/2010

Last update: 4:33 AM EST

In mortgage crisis, uneasy politics (U.S., 8 articles)

" There's no question that this foreclosure issue is, in certain parts of the country, a huge issue, and anything that can be done to stop foreclosures is important said Mark Mellman , a Democratic pollster working for Reid. A wave of lawsuits from troubled borrowers who lost their homes to foreclosure could hurt the already-ailing housing market, a top banking regulator said Monday. Bair said her agency is working closely with other federal regulators to determine whether mortgage companies used flawed documents to seize homes, during remarks made at a housing finance conference sponsored by the FDIC and the Federal Reserve. Raising pressure on banks, the Federal Reserve is wading into the investigation of whether mortgage lenders cut corners and used flawed documents to foreclose on homes. Federal Reserve Chairman Ben Bernanke added weight to those efforts Monday by saying the central bank would look "intensively" at policies and procedures that might have allowed banks to seize homes improperly. SLOPPY, greedy mortgage lenders helped inflate the housing bubble of the Great Recession, now the Federal Reserve is investigating to see if the same avaricious instincts are being applied to home foreclosures.

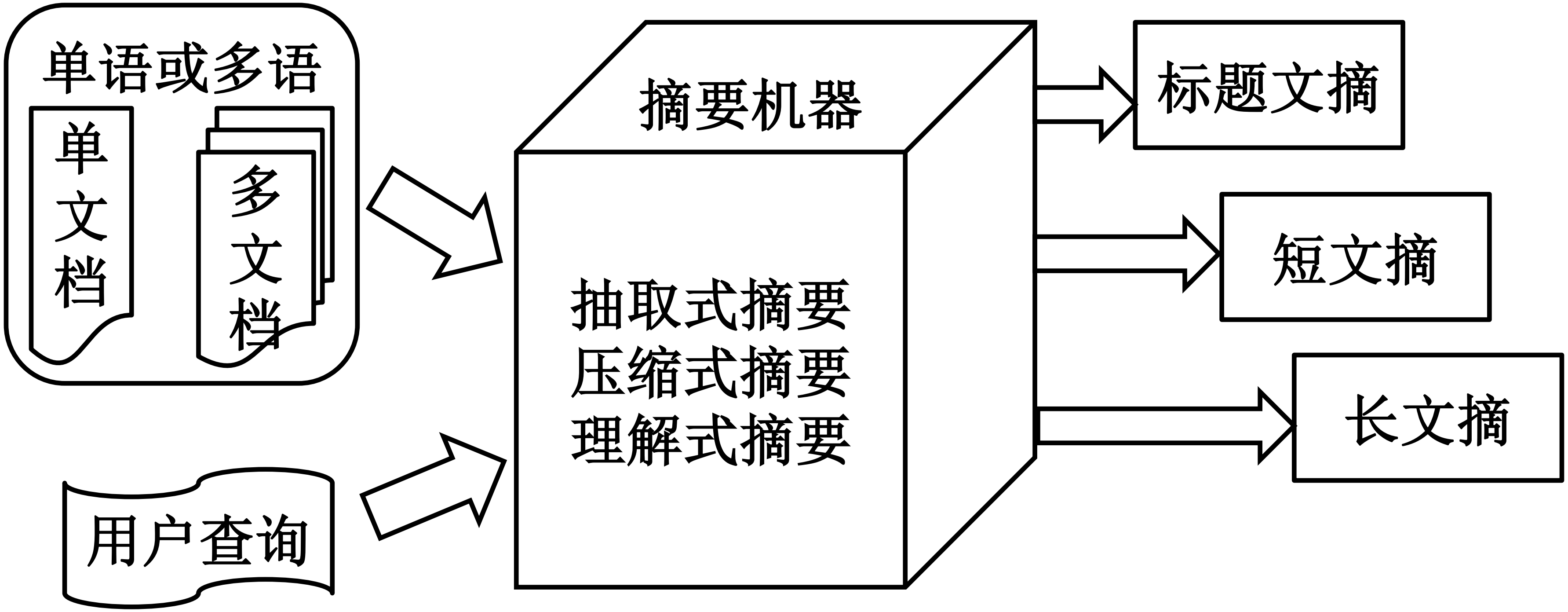
Other stories about Mortgage, borrowers and lenders:

- Mortgage Insurance Reserves: A Lesson in Managing Risk (4 articles)
- Mortgage Web Site Has Admirable Goals, But Will It Save You Money? (4 articles)
- Time for the Government to Go Far Beyond Rate-Freeze Plan for Borrowers (4 articles)

主要内容

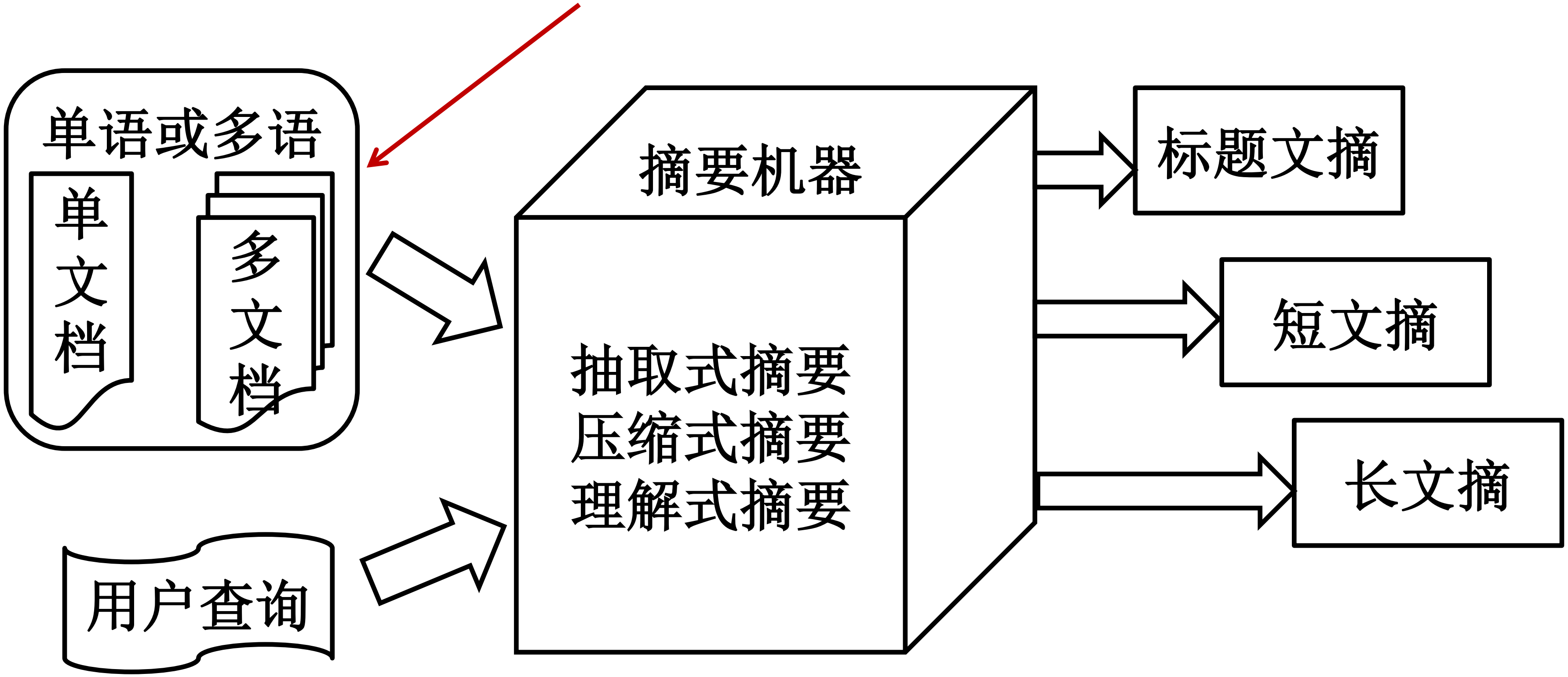
- 文本摘要概述
- 文本摘要分类
- 文本摘要方法
- 文本摘要评价与评测

文本摘要分类

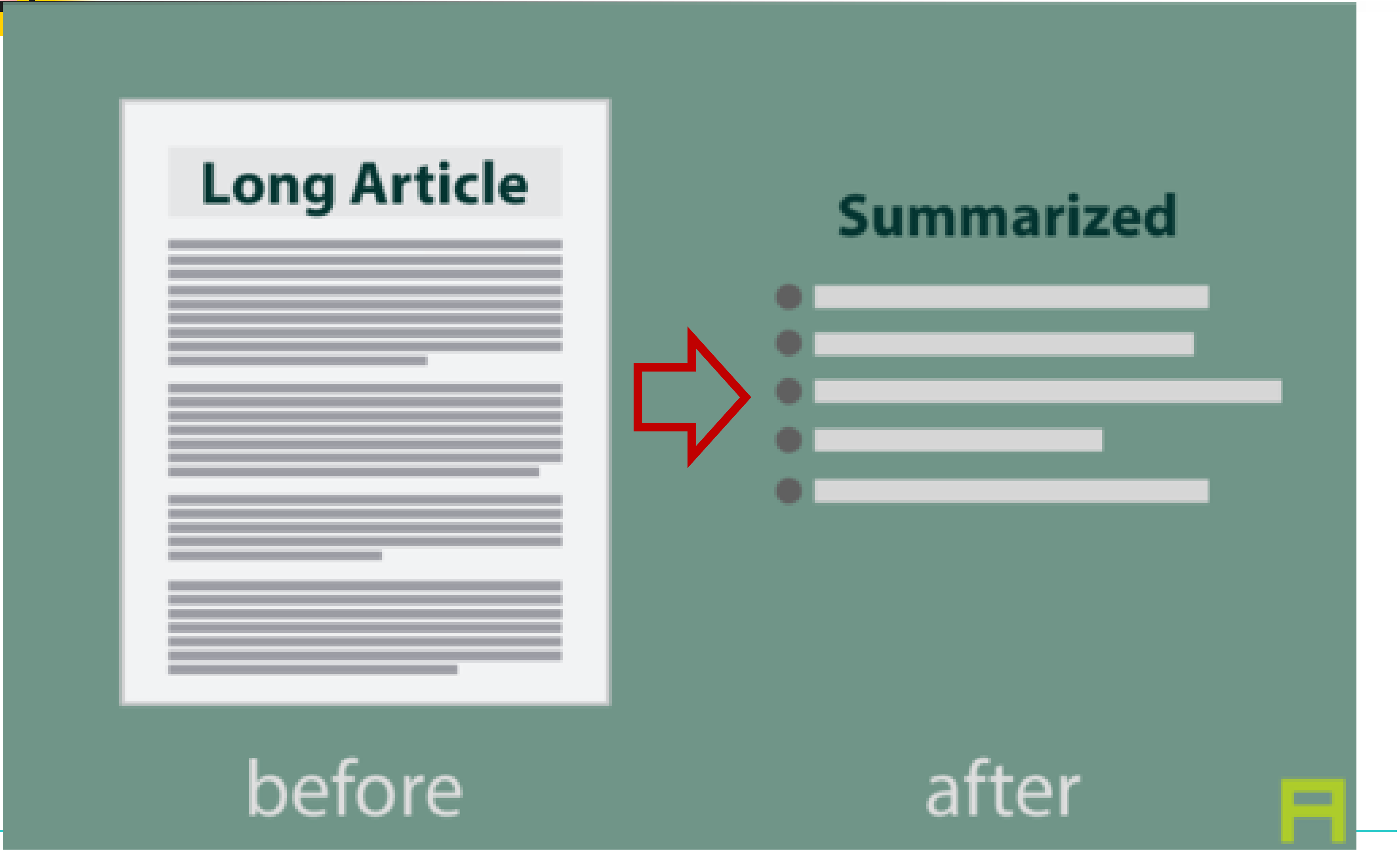


文本摘要分类

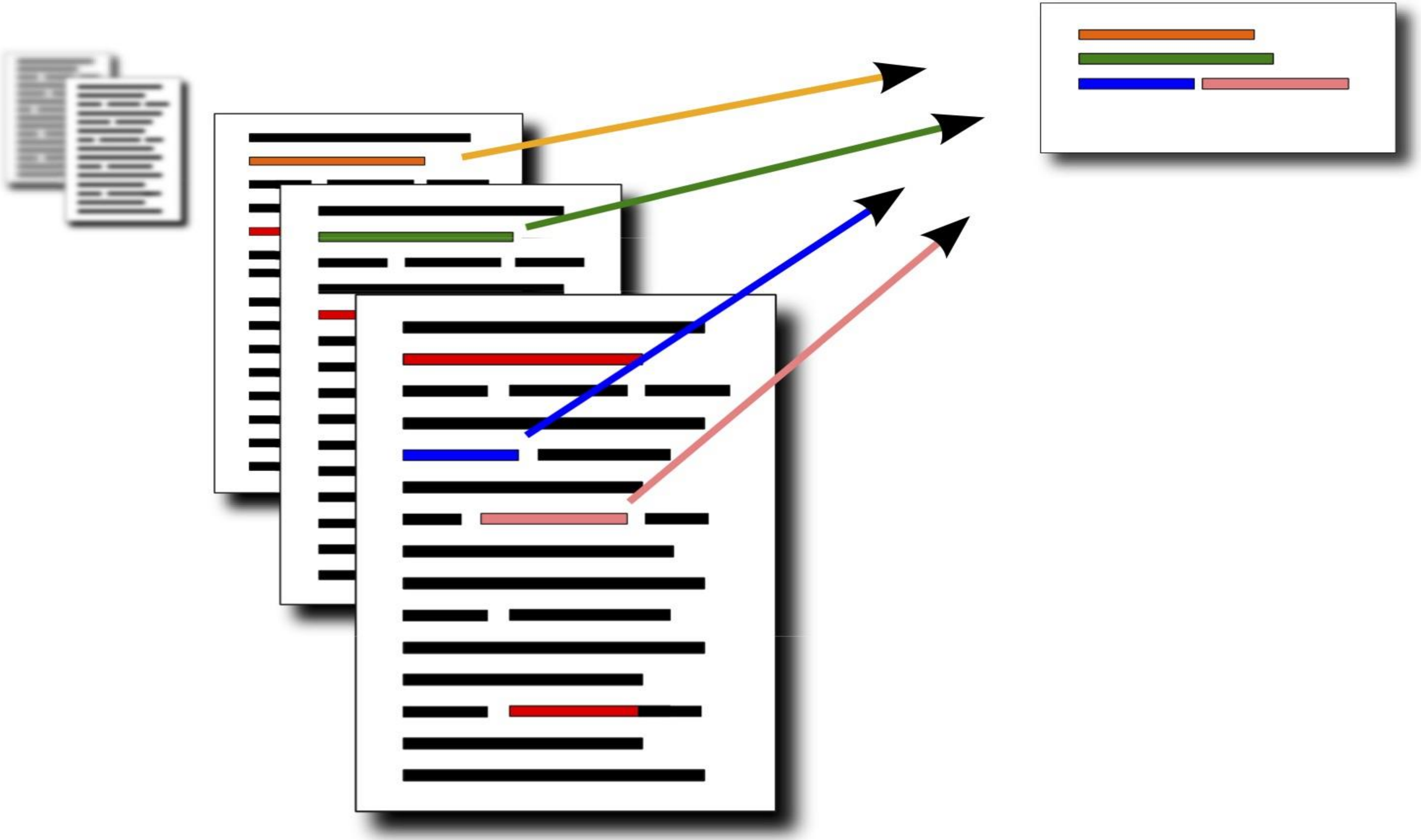
①文档数目：单文档摘要、多文档摘要



单文档摘要



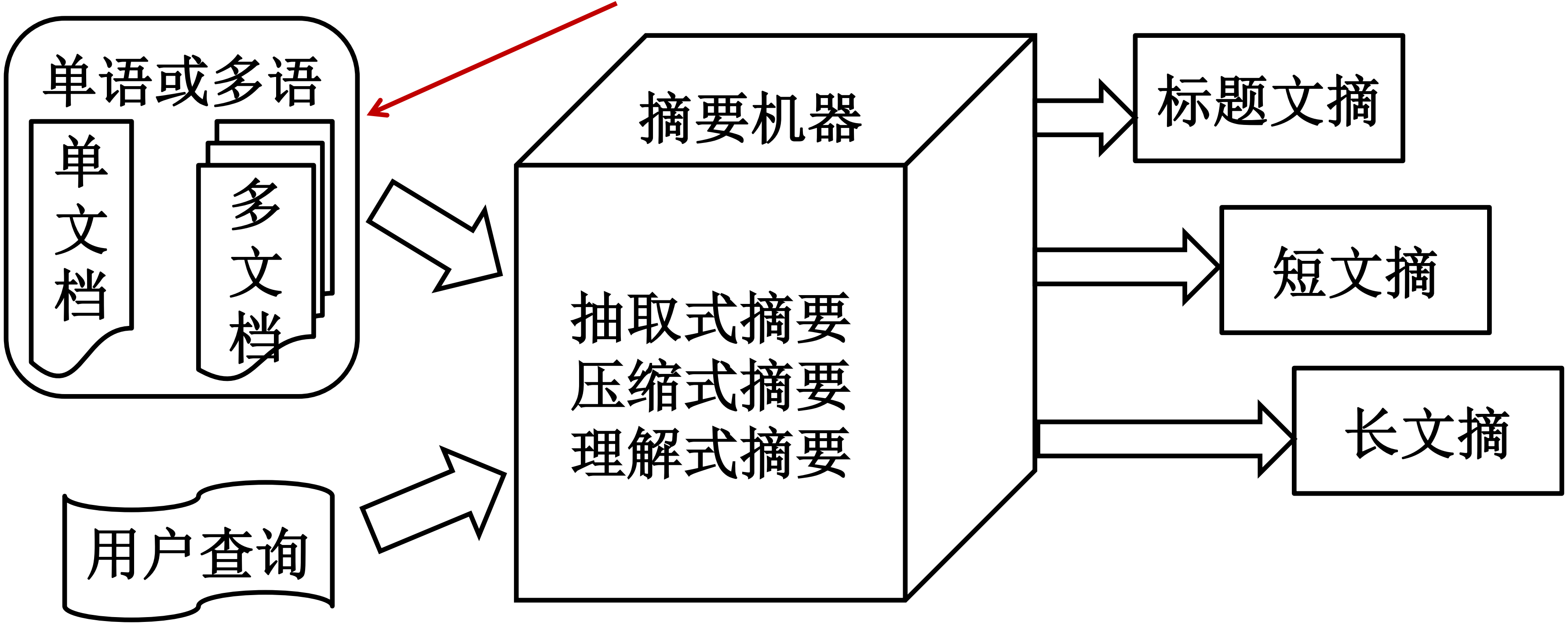
多文档摘要

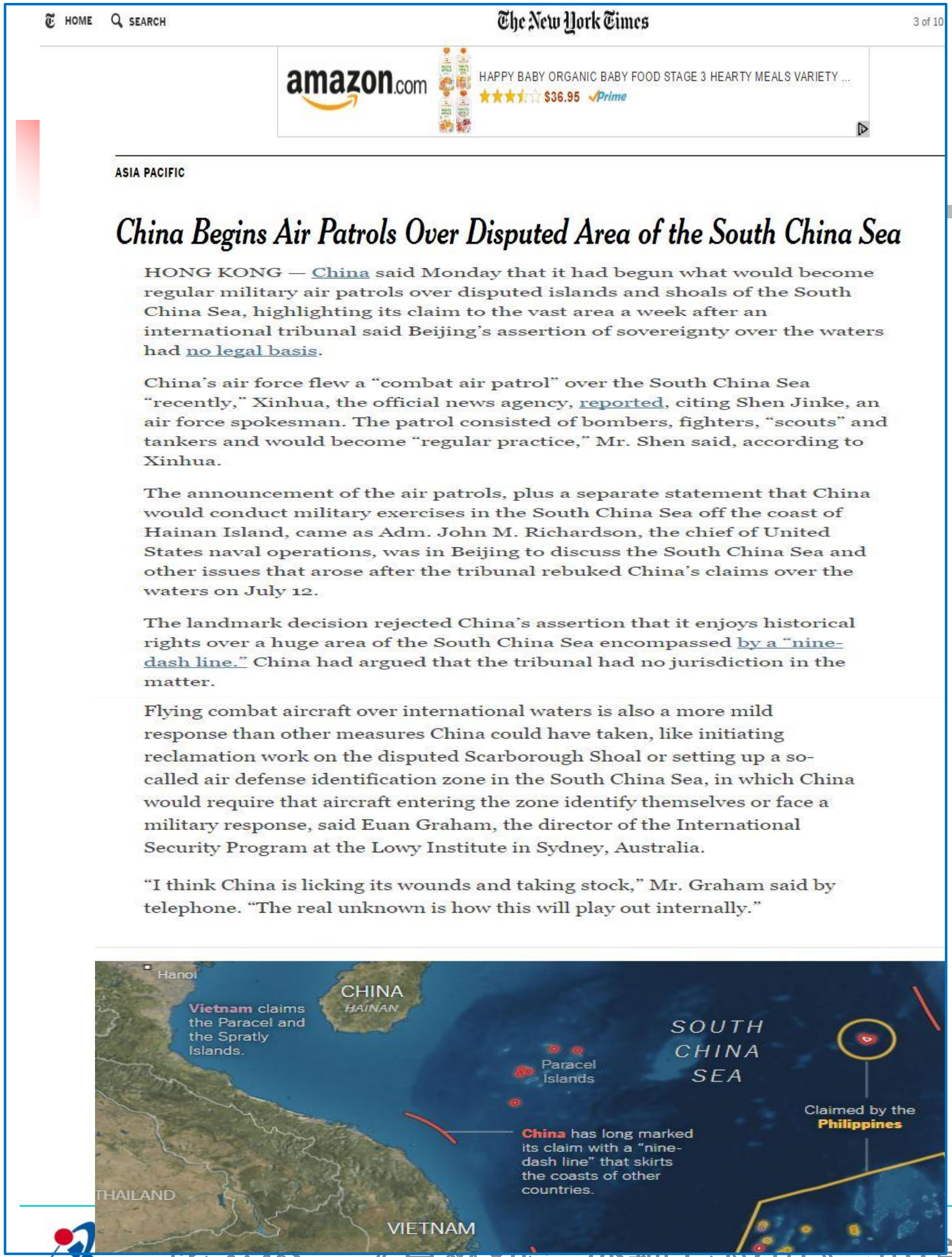


文本摘要分类

②输入语言与输出语言的关系：

单语摘要、跨语言摘要、多语言摘要





amazon.com HAPPY BABY ORGANIC BABY FOOD STAGE 3 HEARTY MEALS VARIETY ... $36.95 Prime

ASIA PACIFIC

China Begins Air Patrols Over Disputed Area of the South China Sea

HONG KONG — China said Monday that it had begun what would become regular military air patrols over disputed islands and shoals of the South China Sea, highlighting its claim to the vast area a week after an international tribunal said Beijing's assertion of sovereignty over the waters had no legal basis.

China's air force flew a "combat air patrol" over the South China Sea "recently," Xinhua, the official news agency, reported, citing Shen Jinke, an air force spokesman. The patrol consisted of bombers, fighters, "scouts" and tankers and would become "regular practice," Mr. Shen said, according to Xinhua.

The announcement of the air patrols, plus a separate statement that China would conduct military exercises in the South China Sea off the coast of Hainan Island, came as Adm. John M. Richardson, the chief of United States naval operations, was in Beijing to discuss the South China Sea and other issues that arose after the tribunal rebuked China's claims over the waters on July 12.

The landmark decision rejected China's assertion that it enjoys historical rights over a huge area of the South China Sea encompassed by a "nine-dash line." China had argued that the tribunal had no jurisdiction in the matter.

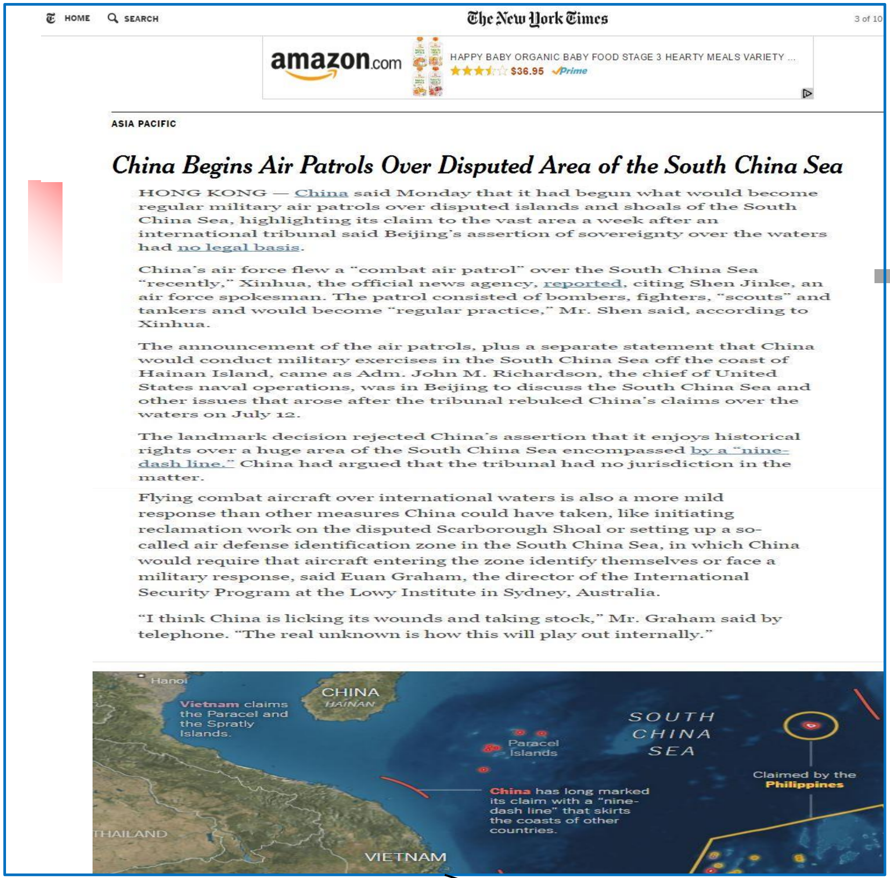
Flying combat aircraft over international waters is also a more mild response than other measures China could have taken, like initiating reclamation work on the disputed Scarborough Shoal or setting up a so-called air defense identification zone in the South China Sea, in which China would require that aircraft entering the zone identify themselves or face a military response, said Euan Graham, the director of the International Security Program at the Lowy Institute in Sydney, Australia.

"I think China is licking its wounds and taking stock," Mr. Graham said by telephone. "The real unknown is how this will play out internally."

Hanoi
CHINA
HAINAN
Vietnam claims the Paracel and the Spratly Islands.
SOUTH CHINA SEA
Paracel Islands
Claimed by the Philippines
China has long marked its claim with a "nine-dash line" that skirts the coasts of other countries.
THAILAND
VIETNAM

跨
语言
摘要

中国空军近期组织了航空兵对南海进行了常规巡逻。专家认为不确定性是该动作如何影响亚太与国际社会舆论。



amazon.com HAPPY BABY ORGANIC BABY FOOD STAGE 3 HEARTY MEALS VARIETY ... $36.95 Prime

ASIA PACIFIC

China Begins Air Patrols Over Disputed Area of the South China Sea

HONG KONG — China said Monday that it had begun what would become regular military air patrols over disputed islands and shoals of the South China Sea, highlighting its claim to the vast area a week after an international tribunal said Beijing's assertion of sovereignty over the waters had no legal basis.

China's air force flew a "combat air patrol" over the South China Sea "recently," Xinhua, the official news agency, reported, citing Shen Jinke, an air force spokesman. The patrol consisted of bombers, fighters, "scouts" and tankers and would become "regular practice," Mr. Shen said, according to Xinhua.

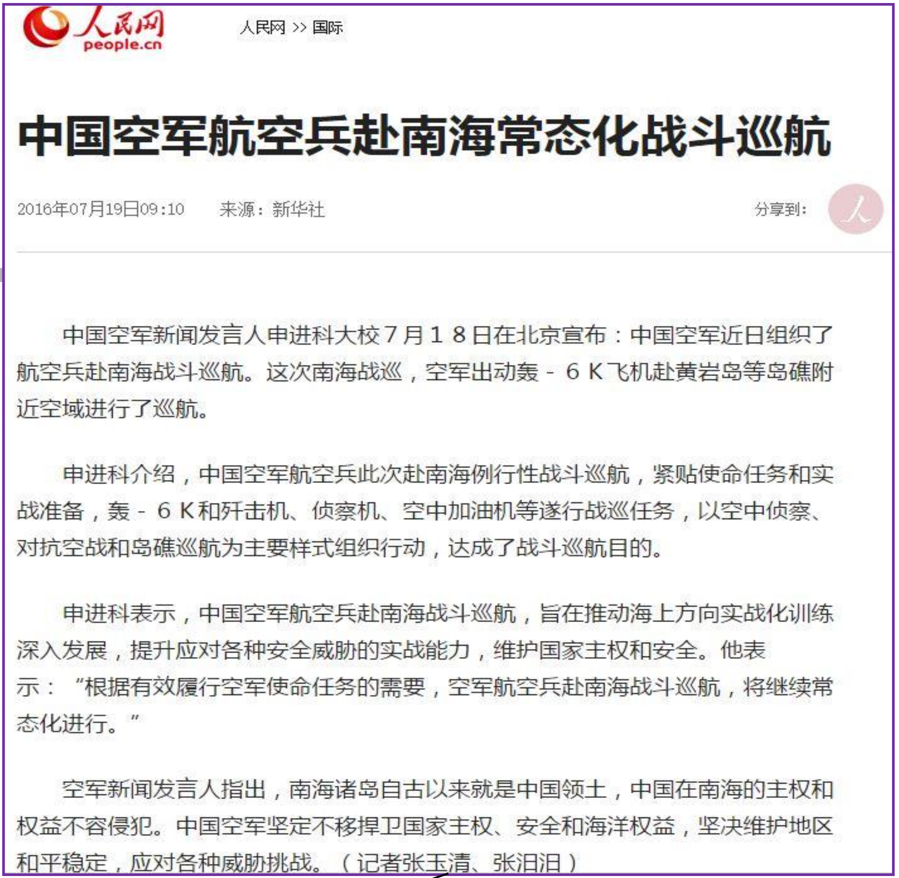
The announcement of the air patrols, plus a separate statement that China would conduct military exercises in the South China Sea off the coast of Hainan Island, came as Adm. John M. Richardson, the chief of United States naval operations, was in Beijing to discuss the South China Sea and other issues that arose after the tribunal rebuked China's claims over the waters on July 12.

The landmark decision rejected China's assertion that it enjoys historical rights over a huge area of the South China Sea encompassed by a "nine-dash line." China had argued that the tribunal had no jurisdiction in the matter.

Flying combat aircraft over international waters is also a more mild response than other measures China could have taken, like initiating reclamation work on the disputed Scarborough Shoal or setting up a so-called air defense identification zone in the South China Sea, in which China would require that aircraft entering the zone identify themselves or face a military response, said Euan Graham, the director of the International Security Program at the Lowy Institute in Sydney, Australia.

"I think China is licking its wounds and taking stock," Mr. Graham said by telephone. "The real unknown is how this will play out internally."

Hanoi
CHINA
HAINAN
Vietnam claims the Paracel and the Spratly Islands.
Paracel Islands
SOUTH CHINA SEA
Claimed by the Philippines
China has long marked its claim with a "nine-dash line" that skirts the coasts of other countries.
THAILAND
VIETNAM



中国空军航空兵赴南海常态化战斗巡航

2016年07月19日09:10　来源：新华社　分享到：

中国空军新闻发言人申进科大校７月１８日在北京宣布：中国空军近日组织了航空兵赴南海战斗巡航。这次南海战巡，空军出动轰－６Ｋ飞机赴黄岩岛等岛礁附近空域进行了巡航。

申进科介绍，中国空军航空兵此次赴南海例行性战斗巡航，紧贴使命任务和实战准备，轰－６Ｋ和歼击机、侦察机、空中加油机等遂行战巡任务，以空中侦察、对抗空战和岛礁巡航为主要样式组织行动，达成了战斗巡航目的。

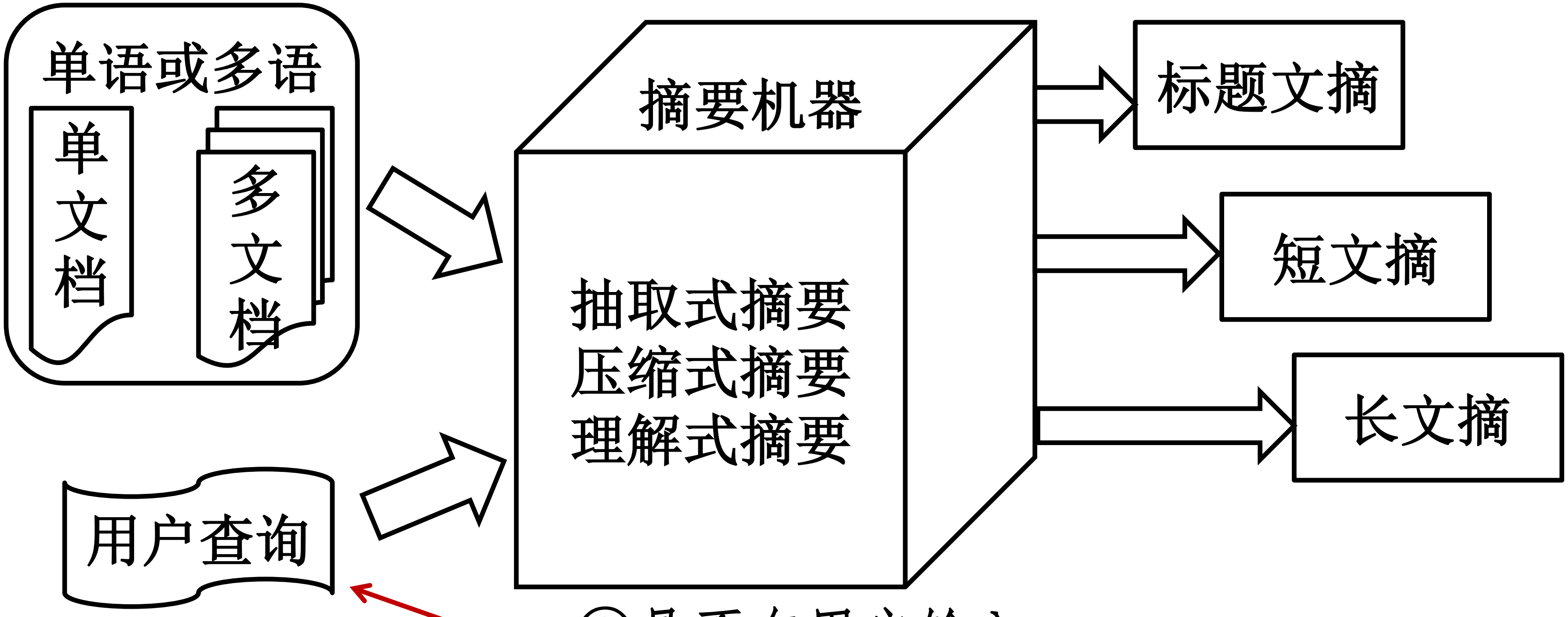
申进科表示，中国空军航空兵赴南海战斗巡航，旨在推动海上方向实战化训练深入发展，提升应对各种安全威胁的实战能力，维护国家主权和安全。他表示："根据有效履行空军使命任务的需要，空军航空兵赴南海战斗巡航，将继续常态化进行。"

空军新闻发言人指出，南海诸岛自古以来就是中国领土，中国在南海的主权和权益不容侵犯。中国空军坚定不移捍卫国家主权、安全和海洋权益，坚决维护地区和平稳定，应对各种威胁挑战。（记者张玉清、张汨汨）

多语言摘要

中国空军近期组织了航空兵对南海进行了常规巡逻。空军新闻发言人指出中国在南海的主权和利益不容侵犯。外国专家认为不确定性是该动作如何影响亚太与国际社会舆论。

文本摘要分类



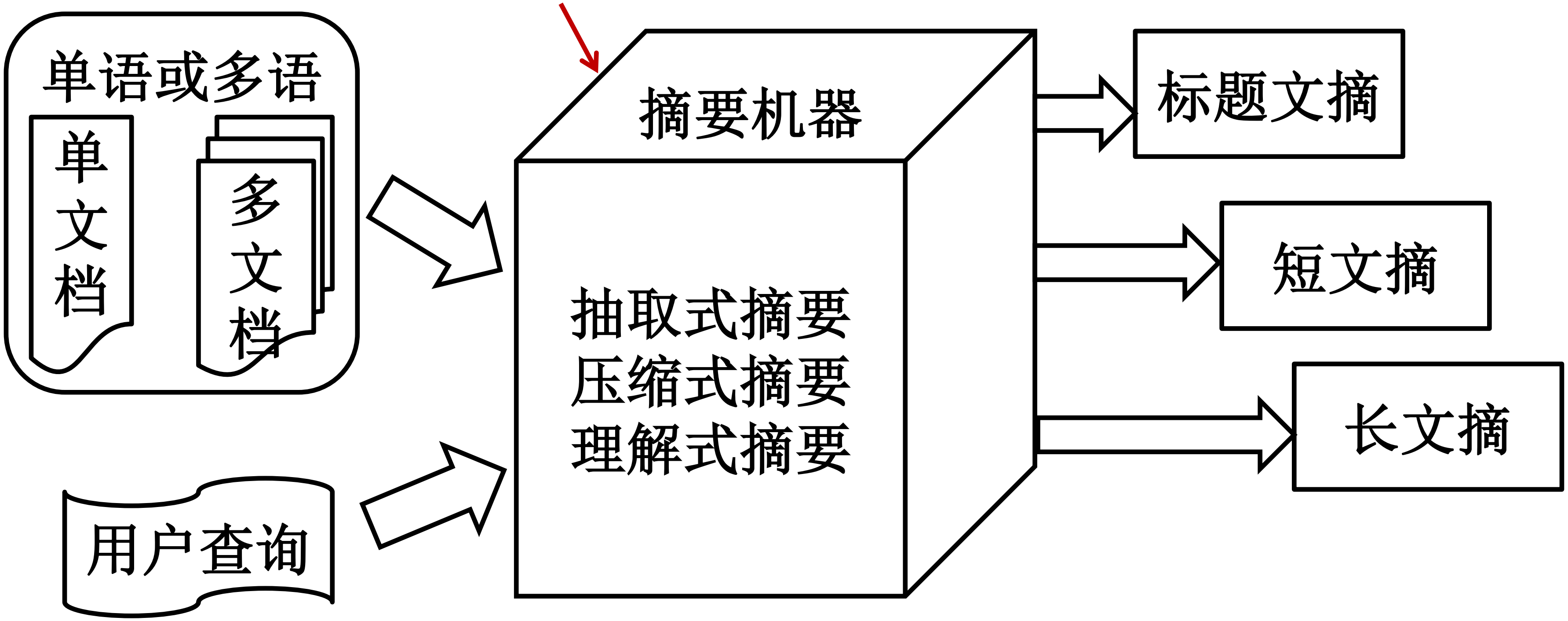
③是否有用户输入：

通用摘要、用户查询摘要

文本摘要分类

④摘要方法：

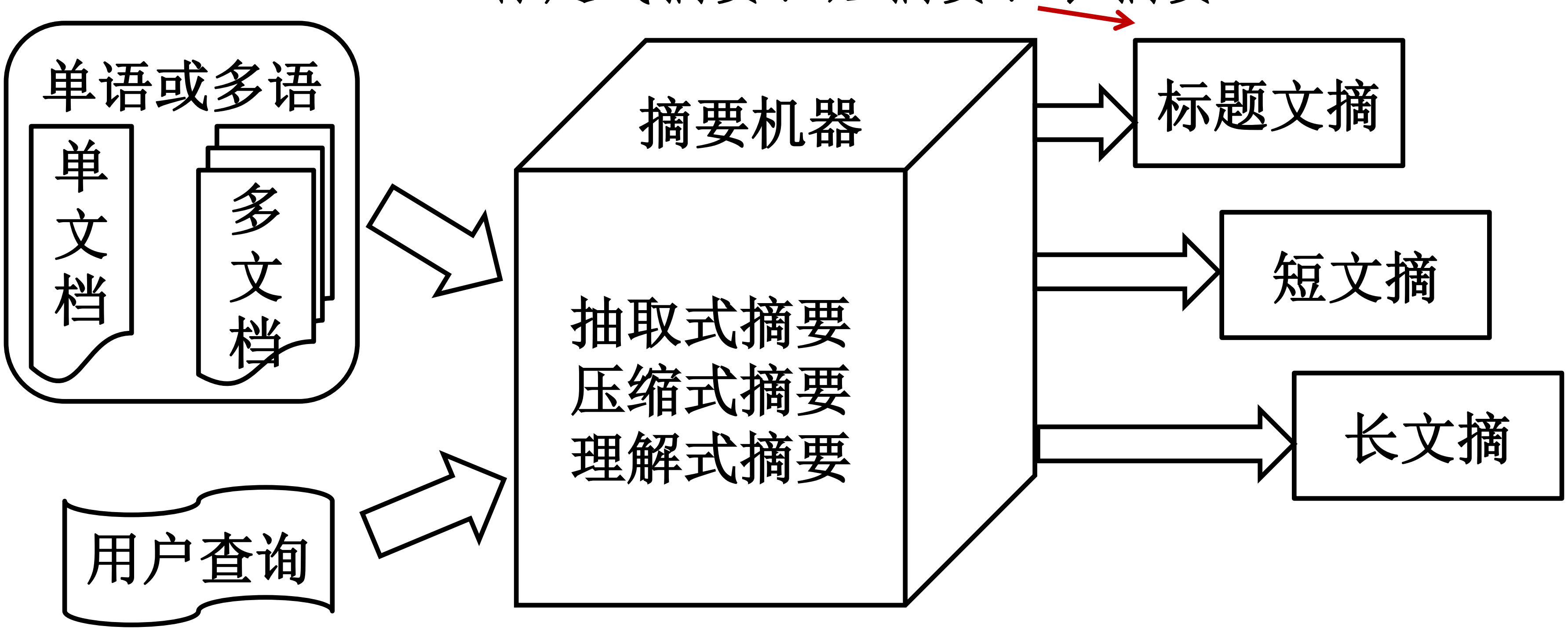
抽取式摘要、压缩式摘要、理解式摘要



文本摘要分类

⑤摘要长度：

标题式摘要、短摘要、长摘要



主要内容

- 文本摘要概述
- 文本摘要分类
- 文本摘要方法
- 文本摘要评价与评测

文本摘要方法

- 抽取式摘要
  - 直接从原文中抽取已有的句子组成摘要
  - 简单易实现，但不符合摘要本质
  - 众多实际系统中，抽取式方法占主导
- 压缩式摘要
  - 抽取并简化原文中的重要句子构成文摘
  - ABACDCDFDSGFGDA → ABADFDSDA
- 理解式摘要
  - 改写或重新组织原文内容形成最终文摘

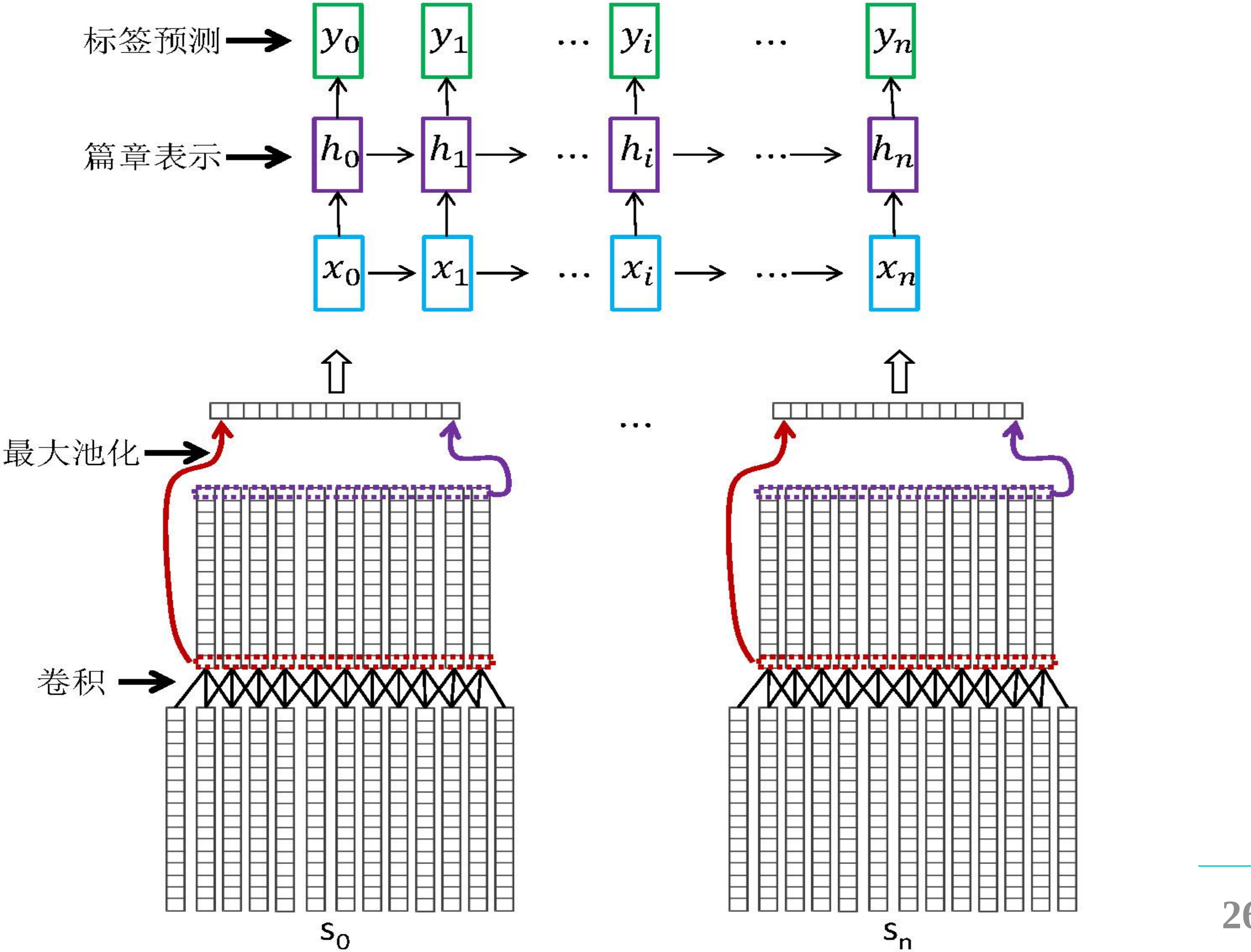
抽取式摘要

- **三个重要模块**
  - 句子重要性评估
  - 信息冗余句子去重复
  - 根据长度、字数等约束生成最终摘要

句子重要性评估

- **启发式规则**
  - 句子位置（越靠段首越重要）、词频、与标题相似度以及线索词（总之、总而言之）等
- **机器学习方法**
  - 句子分类
  - 最优化方法
- **图模型方法**
  - **TextRank**（**PageRank**的无向图模型）
  - **HITS**算法

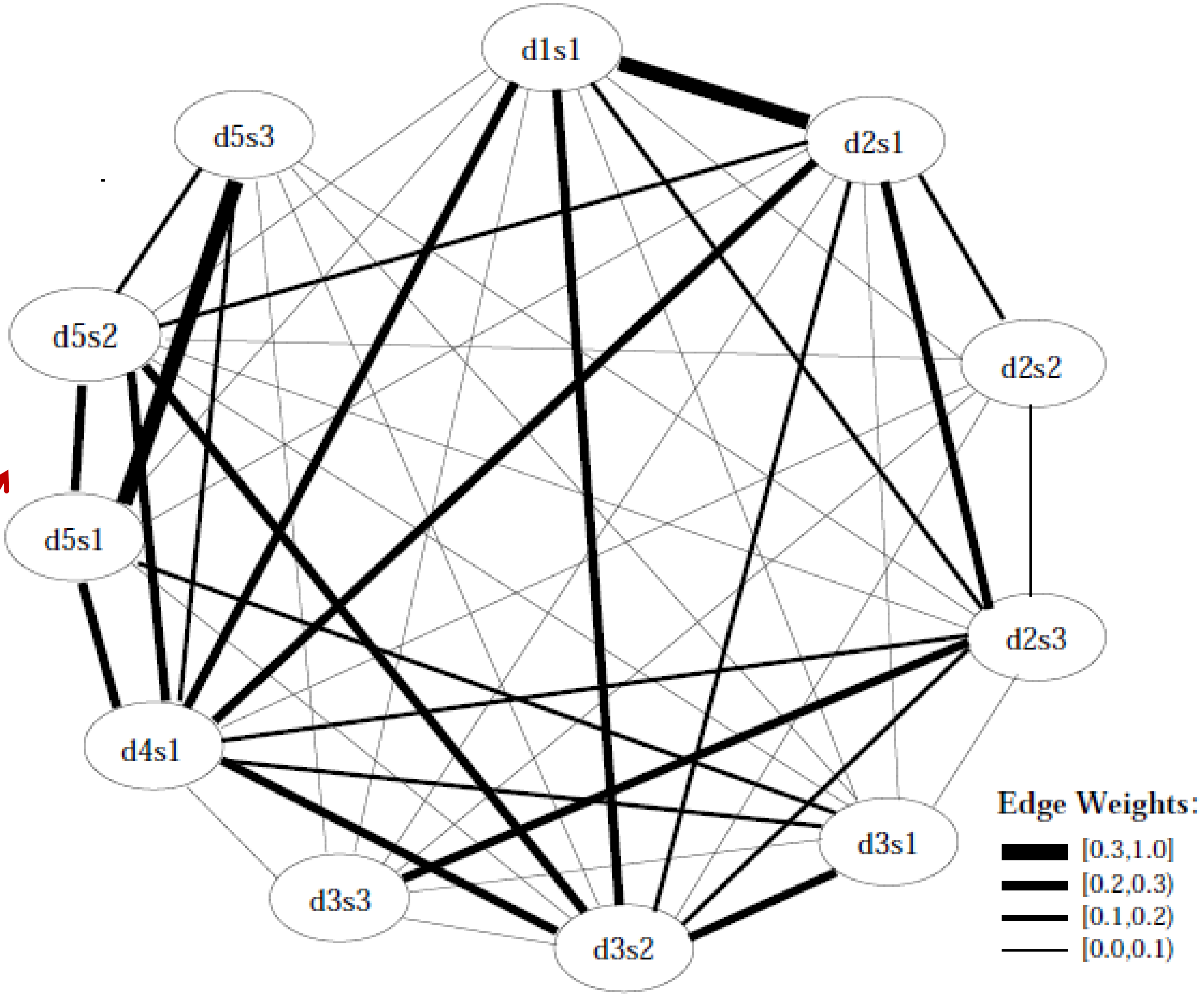
机器学习方法



标签预测
y_0 y_1 … y_i … y_n
篇章表示
h_0 h_1 … h_i … h_n
x_0 x_1 … x_i … x_n
最大池化
…
卷积
s_0
s_n

图模型方法

- G=(V, E)
- V：句子
- E：句间关系



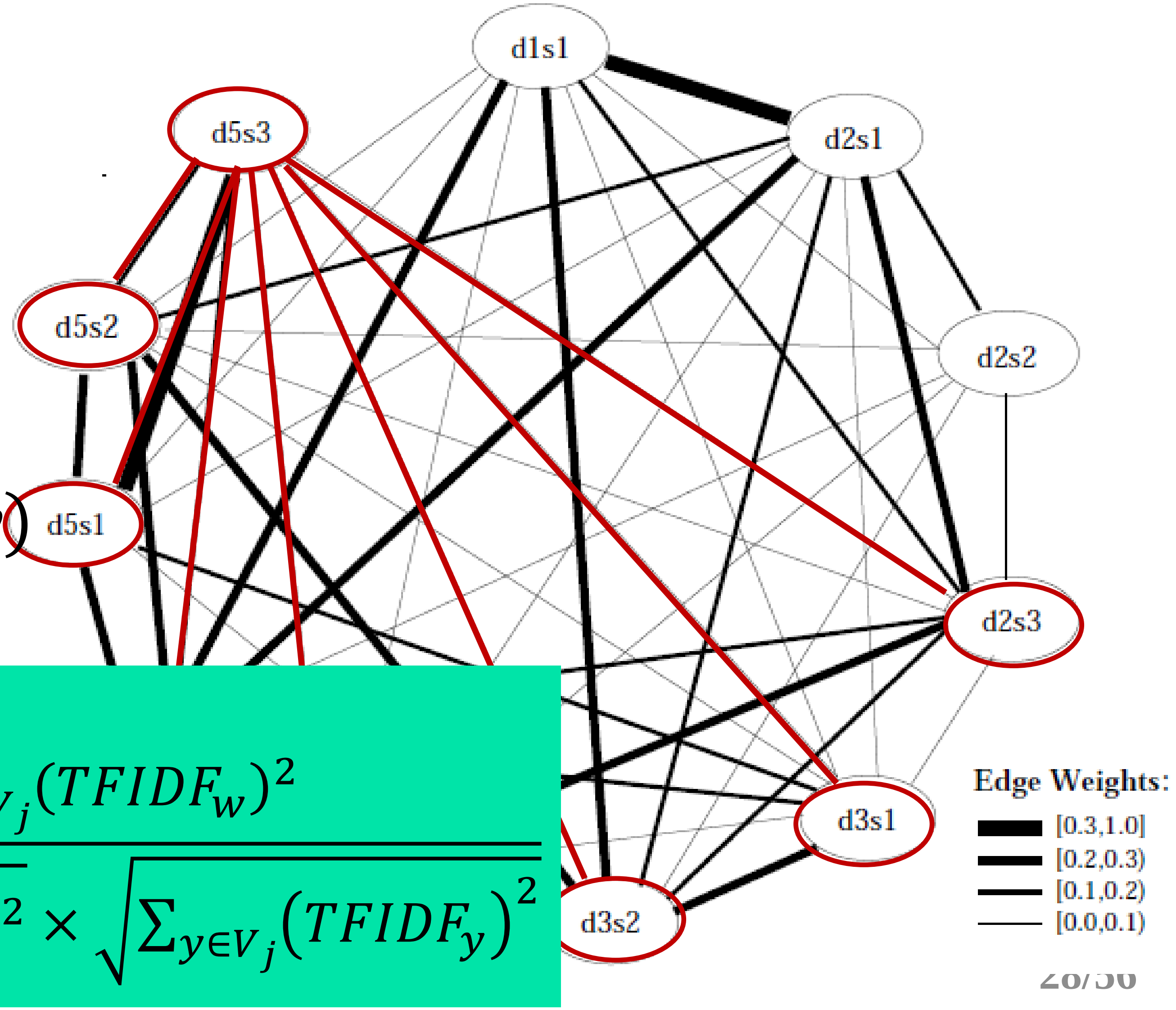
PageRank算法：

计算每个句子的重要性得分

图模型方法–TextRank

- G=(V, E)
- V：句子
- E：句间关系

$$S(u) = \sum_{v\in adj[u]} W_{uv} S(v)$$

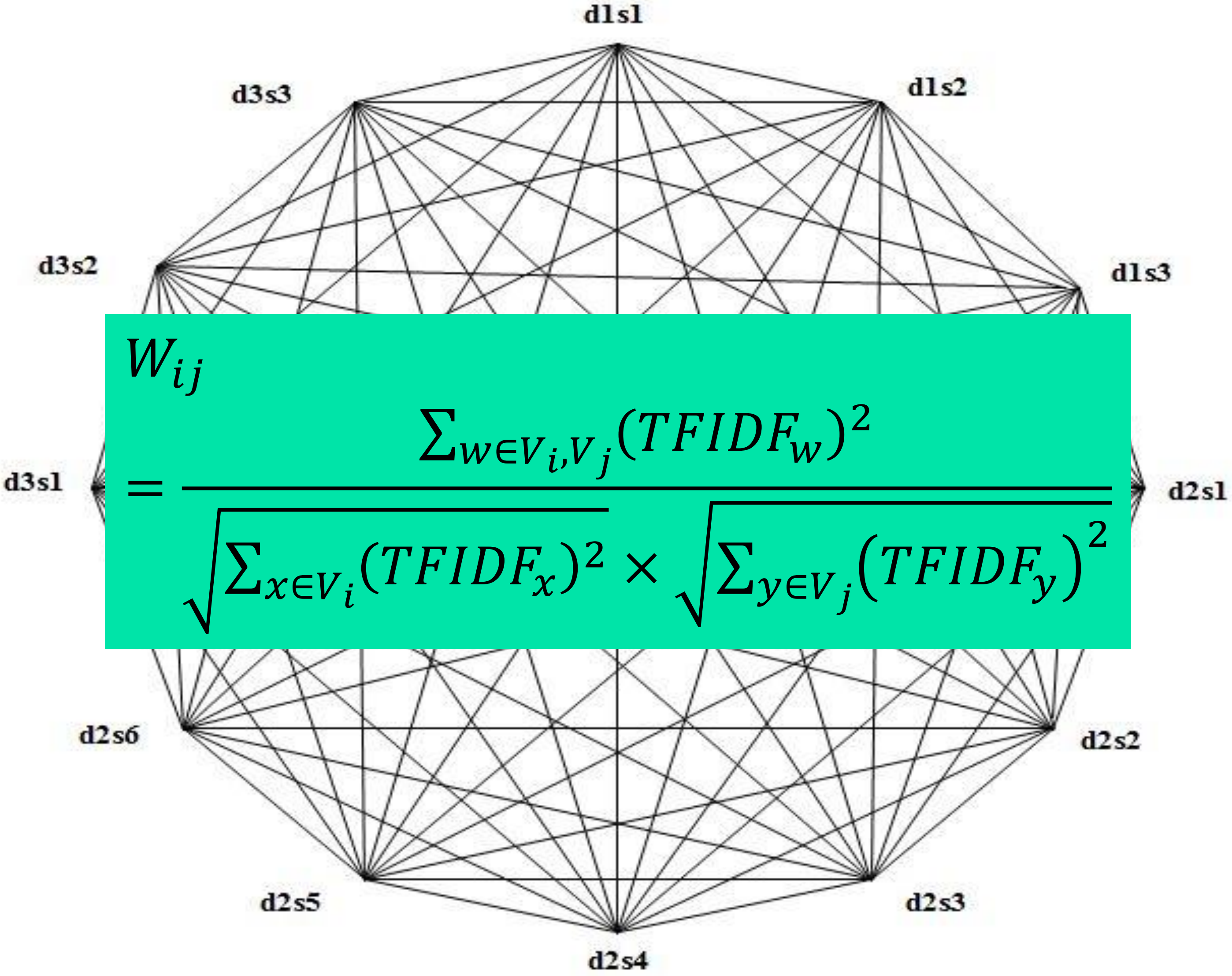


$$W_{ij} = \frac{\sum_{w\in V_i,V_j}(TFIDF_w)^2}{\sqrt{\sum_{x\in V_i}(TFIDF_x)^2} \times \sqrt{\sum_{y\in V_j}(TFIDF_y)^2}}$$

图模型方法—一个例子

序号	ID	句子
1	d1s1	1月10日讯 国际足联周二宣布世界杯扩军至48支球队，这是世界杯自1998年以来首次扩军，而在世界杯87年的历史上，赛制、赛程已经经历了多次改变，参赛球队也从16支扩大到48支。
2	d1s2	因凡蒂诺主政国际足联已接近一年的时间，他上任之初就提出改革的口号，世界杯扩军就是他上任近一年以来最大的改革举措。
3	d1s3	他提出扩军的构想与他此前促成欧洲杯扩军的动机是一样的，他不希望参加世界杯决赛圈的队伍总是一些老面孔，希望有更多的边缘球队能够进入决赛圈，体会足坛盛宴的快乐。
4	d2s1	昨日,国际足联理事会正式对扩军方案进行投票表决,不出意外,扩军至48队分为16个小组的方案获得通过,国际足联官方推特也立即对外宣布了这一消息。
5	d2s2	自2016年2月因凡蒂诺当选国际足联主席后,世界杯扩军便已势在必行,唯一悬念只在于扩军的规模与赛制的改变。
6	d2s3	最开始时,因凡蒂诺提出的是扩军至40支球队参赛,在此前提下又分为两种赛制,一种是分为八个小组,每个小组五支球队,另一种是分为十个小组,每个小组四支球队。
7	d2s4	两个月后,因凡蒂诺再度提出新方案,48支球队参赛,分为16个小组,每个小组3支球队,小组前两名出线,然后进行淘汰赛决出冠军。
8	d2s5	世界杯参赛队伍将从32队扩展到48队,这也意味着未来的世界杯将有接近四分之一的国际足联成员国可以参赛。一些在以前进不了世界杯的足球弱国,终于看到了希望。
9	d2s6	“想在一个地方推广足球,没有比让他们的国家队参与到世界杯更好的方法了。”因凡蒂诺之前就这样表态。
10	d3s1	北京时间1月10日，国际足联宣布，从2026年世界杯开始，世界杯参赛球队将由目前的32支球队扩充至48支。
11	d3s2	最终国际足联官方宣布，自从2026年开始，小组赛将分为16个小组，每个小组3支球队，小组内进行单循环比赛，排名前两位的球队晋级下一轮，然后进行淘汰赛，全部比赛将在32天内完成。
12	d3s3	虽然此前各方意见不尽一致，但扩军符合更多国际足联成员的利益，这也与因凡蒂诺去年当选国际足联主席时的承诺和陈述相符，因此本次扩军乃大势所趋。

图模型方法–一个例子



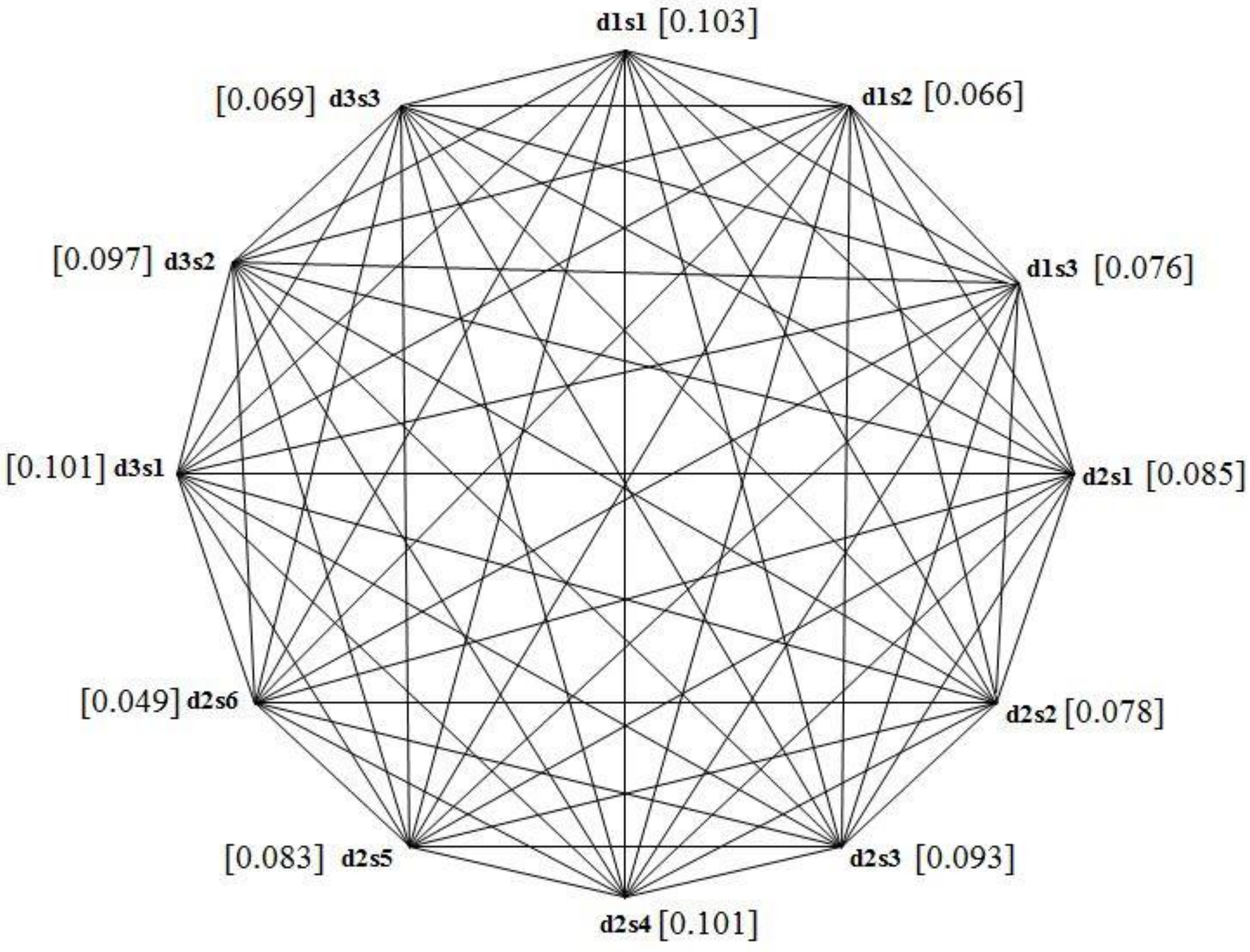
$$W_{ij} = \frac{\sum_{w\in V_{i},V_{j}}(TFIDF_{w})^{2}}{\sqrt{\sum_{x\in V_{i}}(TFIDF_{x})^{2}}\times\sqrt{\sum_{y\in V_{j}}\left(TFIDF_{y}\right)^{2}}}$$

图模型方法—一个例子

1.000	0.129	0.141	0.121	0.187	0.106	0.137	0.173	0.076	0.471	0.266	0.150
0.129	1.000	0.239	0.040	0.114	0.039	0.032	0.086	0.085	0.137	0.109	0.120
0.141	0.239	1.000	0.044	0.101	0.094	0.027	0.167	0.052	0.140	0.144	0.199
0.121	0.040	0.044	1.000	0.152	0.262	0.365	0.197	0.071	0.072	0.138	0.096
0.187	0.114	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.029	0.176
0.106	0.039	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.214	0.051
0.137	0.032	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.282	0.020
0.173	0.086	0.167	0.197	0.109	0.114	0.135	1.000	0.152	0.206	0.091	0.069
0.076	0.085	0.052	0.071	0.088	0.082	0.084	0.152	1.000	0.040	0.021	0.043
0.471	0.137	0.140	0.072	0.091	0.119	0.142	0.206	0.040	1.000	0.369	0.129
0.266	0.109	0.144	0.138	0.029	0.214	0.282	0.091	0.021	0.369	1.000	0.162
0.150	0.120	0.199	0.096	0.176	0.051	0.020	0.069	0.043	0.129	0.162	1.000

$$S(u) = \sum_{v\in adj[u]} W_{uv} S(v)$$

图模型方法--一个例子


d1s1 [0.103]
[0.069] d3s3
d1s2 [0.066]
[0.097] d3s2
d1s3 [0.076]
[0.101] d3s1
d2s1 [0.085]
[0.049] d2s6
d2s2 [0.078]
[0.083] d2s5
d2s3 [0.093]
d2s4 [0.101]

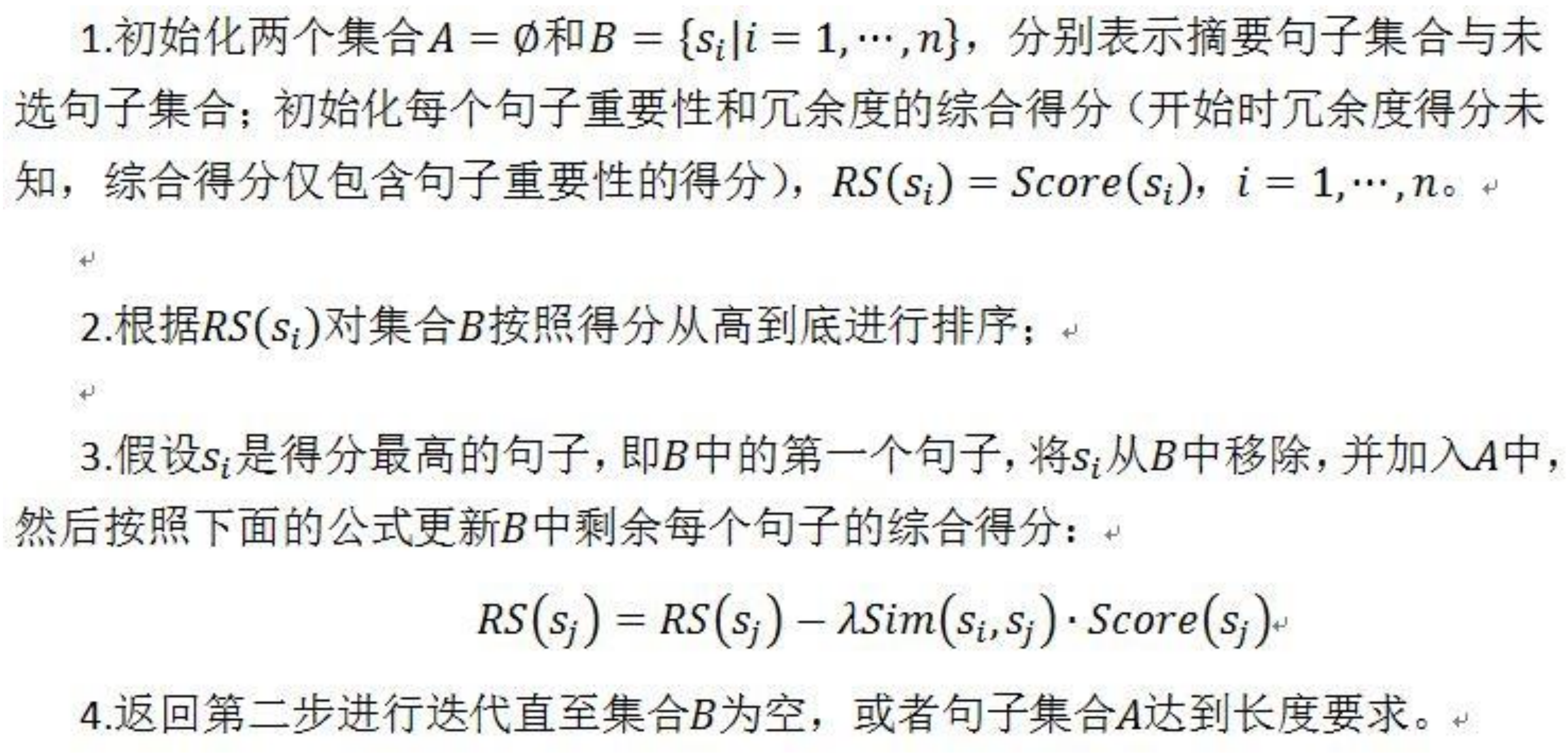
冗余句子消除

- **必要性**
  - 多文档摘要中，不同文档通常包含非常相似的句子
  - 为了得到精简的摘要，需要消除冗余的句子

- **主要方法**
  - **CSIS**
  - **MMR**

$$MMR'(R,A) = \underset{s_i \in R \backslash A}{\operatorname{argmax}} \left\{ \lambda Score(s_i) - (1-\lambda) \max_{s_j \in A} Sim(s_i, s_j) \cdot Score(s_j) \right\}$$

MMR算法



1.初始化两个集合$A=\emptyset$和$B=\{s_i|i=1,\cdots,n\}$，分别表示摘要句子集合与未选句子集合；初始化每个句子重要性和冗余度的综合得分（开始时冗余度得分未知，综合得分仅包含句子重要性的得分），$RS(s_i)=Score(s_i)$，$i=1,\cdots,n$。

2.根据$RS(s_i)$对集合B按照得分从高到底进行排序；

3.假设s_i是得分最高的句子，即B中的第一个句子，将s_i从B中移除，并加入A中，然后按照下面的公式更新B中剩余每个句子的综合得分：

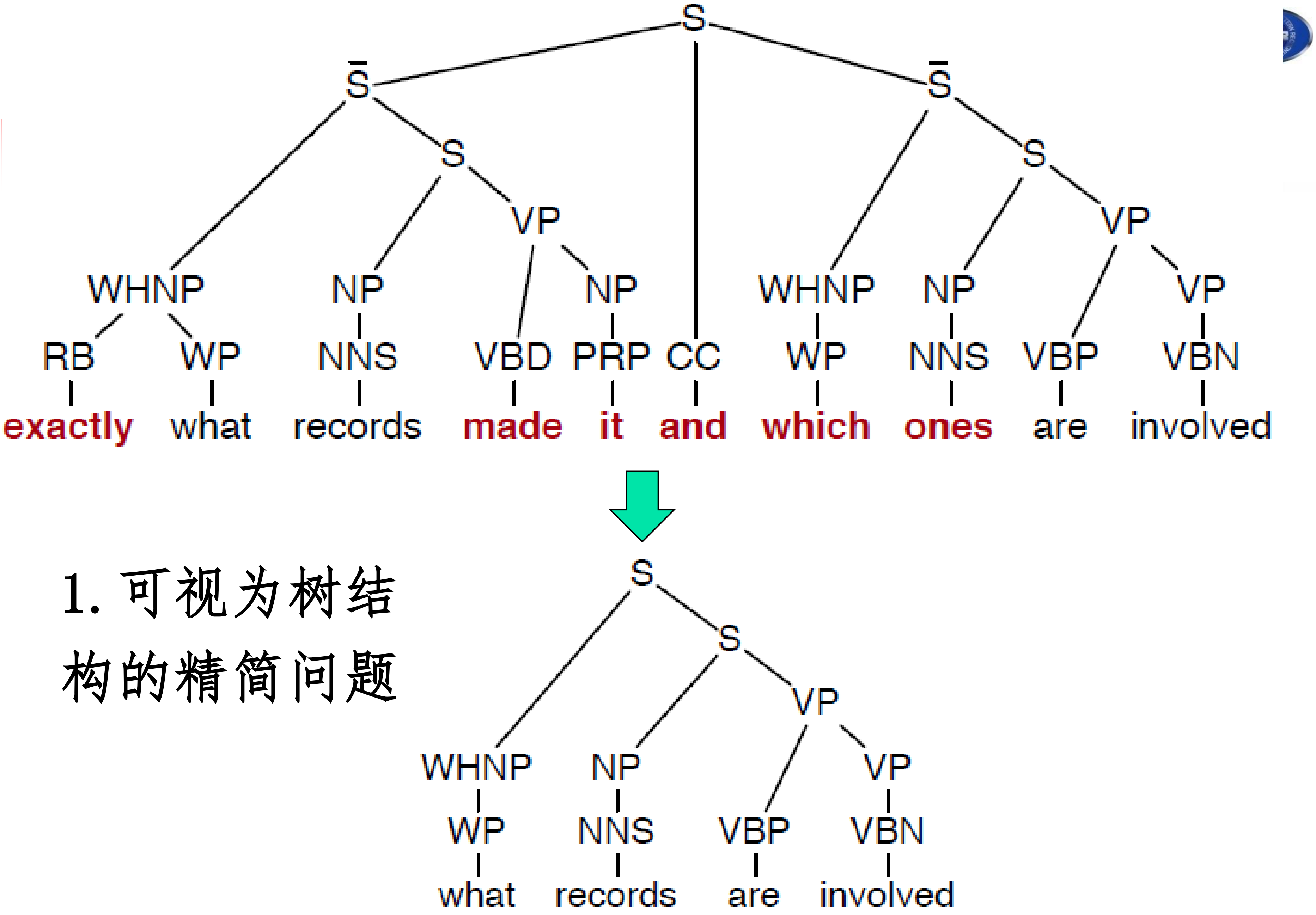
$$RS(s_j)=RS(s_j)-\lambda Sim(s_i,s_j)\cdot Score(s_j)$$

4.返回第二步进行迭代直至集合B为空，或者句子集合A达到长度要求。

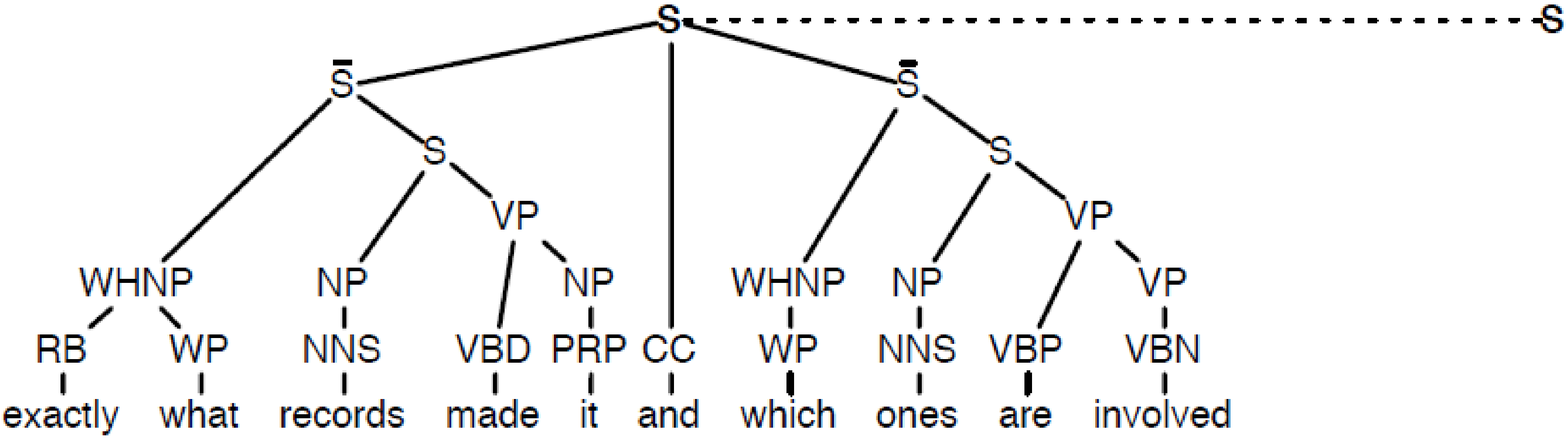
压缩式摘要

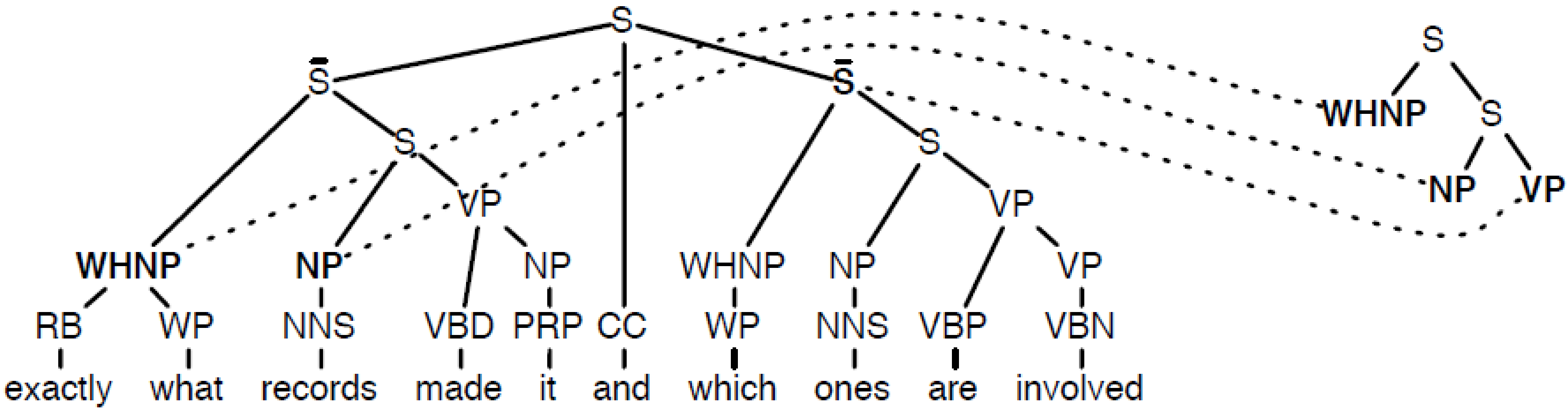
- **核心模块：句子压缩**

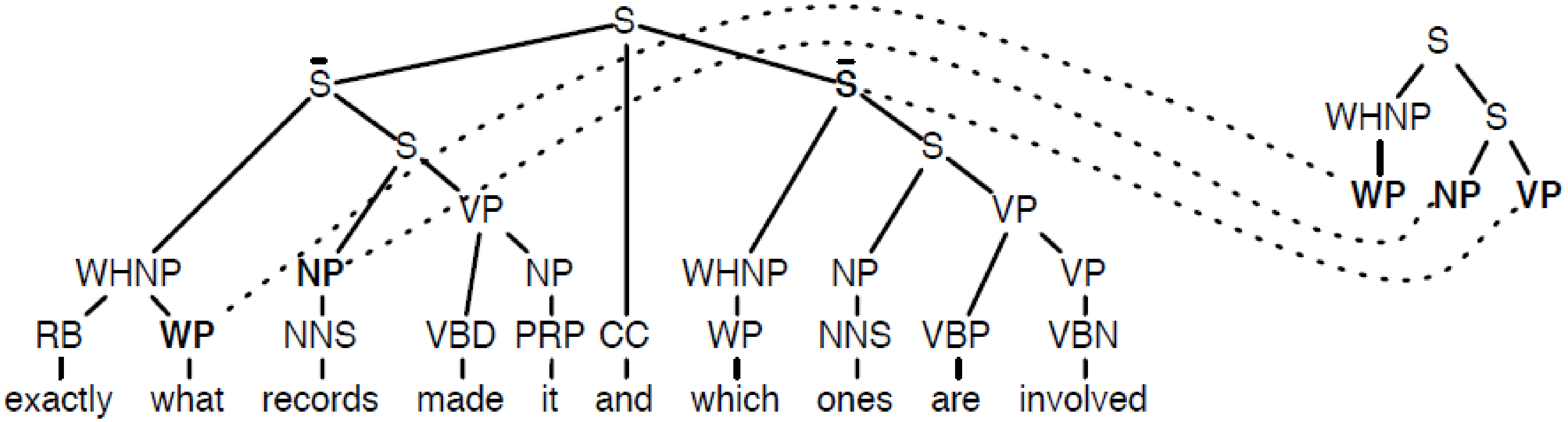
ABACDCDFDSGFGDA → ABADFDSDA



1. 可视为树结构的精简问题


S
S
S̄
S̄
S
S
VP
VP
WHNP
NP
NP
WHNP
NP
VP
RB
WP
NNS
VBD
PRP
CC
WP
NNS
VBP
VBN
exactly
what
records
made
it
and
which
ones
are
involved


S
S̄
S̄
S
S
VP
VP
WHNP
NP
NP
WHNP
NP
VP
RB
WP
NNS
VBD
PRP
CC
WP
NNS
VBP
VBN
exactly
what
records
made
it
and
which
ones
are
involved
S
WHNP
S
NP
VP


S
S̄
S
VP
WHNP
NP
NP
RB
WP
NNS
VBD
PRP
CC
exactly
what
records
made
it
and
S̄
S
VP
WHNP
NP
VP
WP
NNS
VBP
VBN
which
ones
are
involved
S
WHNP
S
WP
NP
VP

压缩式摘要

- 核心模块：句子压缩

2. 可视为01序列标注任务

ABACDCDFDSGFGDA → ABADFDSDA

1 1 1 0 0 0 1 1 1 1 0 0 0 1 1

理解式摘要

- 改写或重新组织原文内容形成文摘

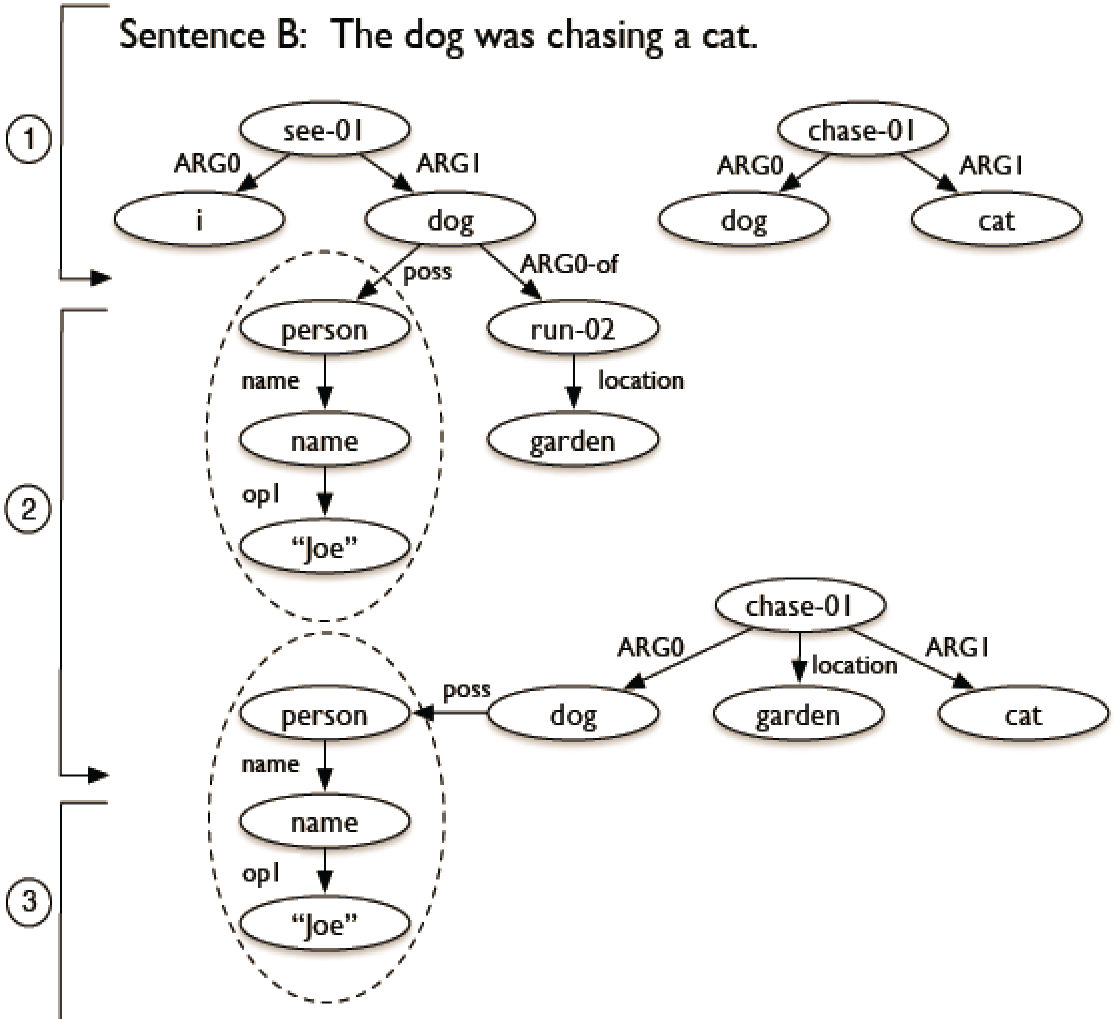
基于AMR的方法

AMR:

Abstractive Meaning Representation

Sentence A: I saw Joe's dog, which was running in the garden.

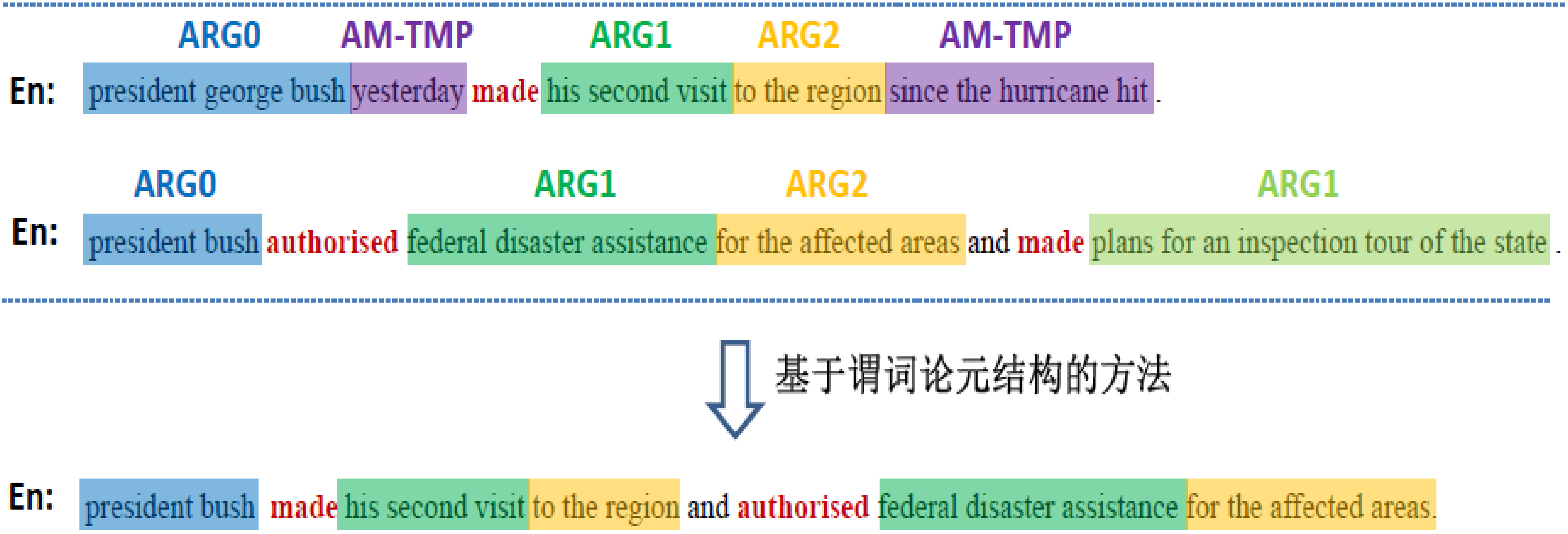
Sentence B: The dog was chasing a cat.



Summary: Joe's dog was chasing a cat in the garden.

理解式摘要

- **基于谓词论元结构的理解式摘要**
  - 核心思想：选择并重组概念与行为
  - 选择：基于图的重要性打分+基于约束的整数线性规划



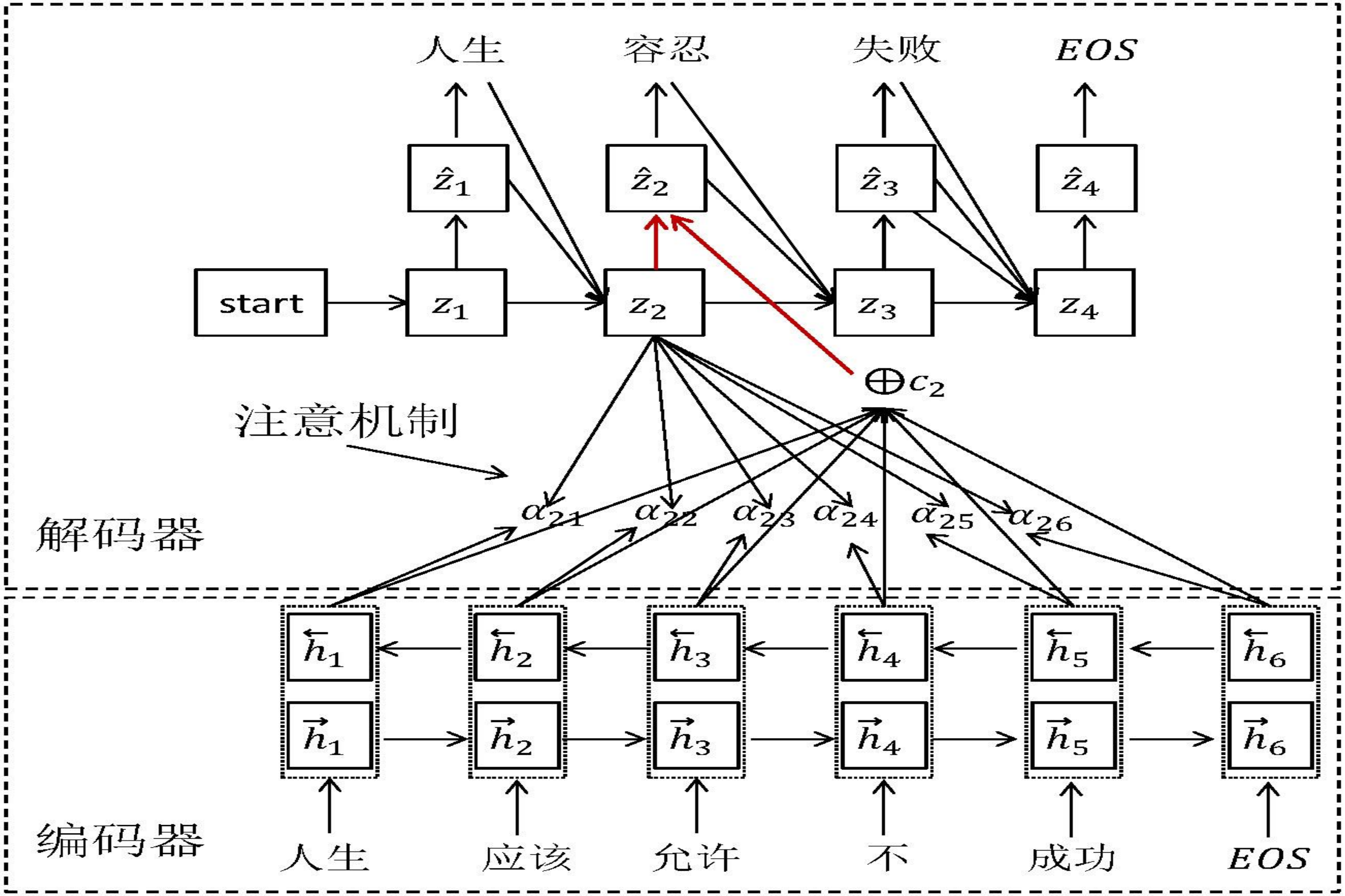
En: president george bush yesterday made his second visit to the region since the hurricane hit .

En: president bush authorised federal disaster assistance for the affected areas and made plans for an inspection tour of the state .

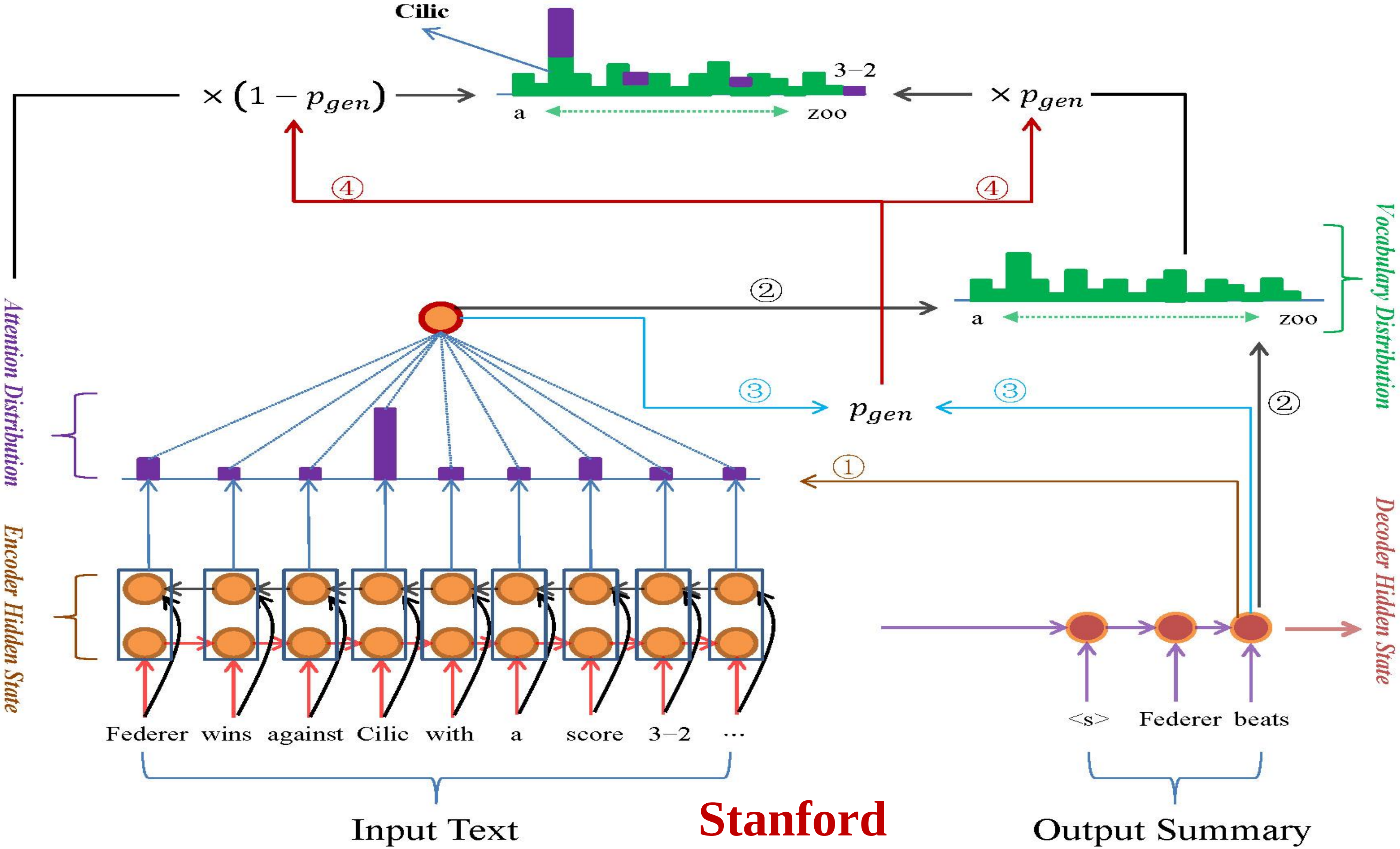
基于谓词论元结构的方法

En: president bush made his second visit to the region and authorised federal disaster assistance for the affected areas.

理解式摘要


人生
容忍
失败
EOS
$\hat{z}_1$
$\hat{z}_2$
$\hat{z}_3$
$\hat{z}_4$
start
z_1
z_2
z_3
z_4
$\oplus c_2$
注意机制
α_{21}
α_{22}
α_{23}
α_{24}
α_{25}
α_{26}
解码器
$\overleftarrow{h}_1$
$\overleftarrow{h}_2$
$\overleftarrow{h}_3$
$\overleftarrow{h}_4$
$\overleftarrow{h}_5$
$\overleftarrow{h}_6$
$\overrightarrow{h}_1$
$\overrightarrow{h}_2$
$\overrightarrow{h}_3$
$\overrightarrow{h}_4$
$\overrightarrow{h}_5$
$\overrightarrow{h}_6$
编码器
人生
应该
允许
不
成功
EOS

理解式摘要



Stanford

端到端摘要方法

- 哈佛大学（Harvard University）
  - Alexander M. Rush提出Seq2Seq摘要的思想
  - 牵头实现并开源了Seq2Seq代码OpenNMT
- 斯坦福大学 （Stanford University）
  - 实现了目前最优的Seq2Seq摘要模型
  - 包括copy机制和coverage机制
  - …

主要内容

- 文本摘要概述
- 文本摘要分类
- 文本摘要方法
- 文本摘要评价与评测

文本摘要评价

- 自动评价
  - 给定人工参考摘要，评价自动摘要结果的质量，综合考虑内容的忠实度与行文的流畅度
  - 省时省力、一致性高、加速方法迭代更新
  - ROUGE：基于N-元组计算自动摘要与人工摘要的匹配率
  - BE：基于语义单元的ROUGE，语义单元由句法分析得到

文本摘要评价

$$ROUGE-N(sum) = \frac{\sum_{r\in R}\sum_{n-gram\in r} count_{match}(n-gram, sum)}{\sum_{r\in R}\sum_{n-gram\in r} count(n-gram)}$$

文本摘要评价

- **人工评价**
  - 人工评价自动摘要结果的质量
  - 可靠性高、主观性强
  - 内容的忠实度：金字塔方法
  - 行文的流畅度（可读性）：1-5

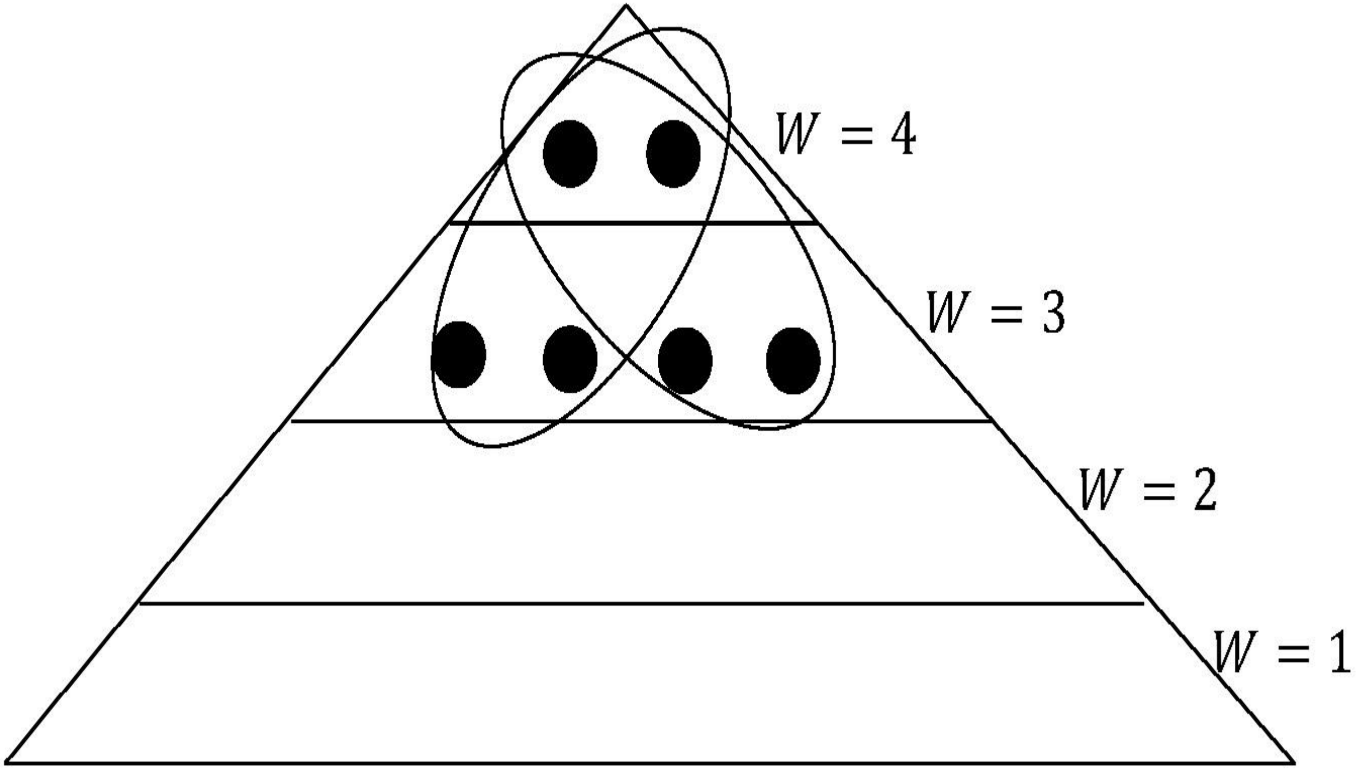
金字塔方法

A2:	**2016年美国大选，特朗普击败希拉里，当选第45任美国总统。**
B4:	他赢得了第45任美国总统大选。
C3:	特朗普成为第45任美国总统。
D1:	2016年的美国大选悬念迭起，最终特朗普有惊无险，赢得胜利。

金字塔方法

SCU1:	**特朗普当选第45届美国总统**
	A2: 特朗普击败希拉里，当选第45任美国总统
	B4: 他赢得了第45届美国总统大选
	C3: 特朗普成为第45任美国总统
	D1: 特朗普有惊无险，赢得胜利
SCU2:	2016年美国举行总统大选
	A2: 2016年美国大选
	D1: 2016年的美国大选

金字塔方法



文本摘要评测

- **DUC（Document Understanding Conference）2000-2007**
  - 由NIST组织
  - 评测任务：单文档、多文档、查询相关
- **TAC （Text Analysis Conference）2008-2011**
  - 仍由NIST组织
  - 评测任务：更新式摘要、观点摘要、指导性摘要、自动摘要评价

主要内容

- 文本摘要概述
- 文本摘要分类
- 文本摘要方法
- 文本摘要评价与评测

参考文献

1. Lexrank graph-based lexical centrality as salience in text summarization, 2004

2. A survey on automatic text summarization, 2007

3. Automatic summarizaiton, 2011

4. Improving multi-documents summarization by sentence compression based on expanded constituent parse trees, 2014

5. Extractive summarization based on keyword profile and language model, 2015

6. Toward abstractive summarization using semantic representations, 2015

扩展阅读

7. A new approache to improving multilingual summarization using a genetic algorithm, 2010

8. Using bilingual information for cross language document summarization, 2011

9. Abstractive Multi-Document Summarization via Phrase Selection and Merging, 2015

10. Rouge: A package for automatic evaluation of summaries, 2004

11. Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In EMNLP-2015.

12. Abigail See, Peter J. Liu, and Christopher D. Manning. Get To The Point: Summarization with Pointer-Generator Networks. In ACL-2017.

Thanks

谢谢!