

# **《语义计算与知识检索》研究生课程**

## **文本推理与复述**

**万小军**

**北京大学语言计算与互联网挖掘组**

<http://www.icst.pku.edu.cn/lcwm/course/sckr2018>

**2018年4月11日**

# 内容

- 文本推理
- 文本复述

# 文本推理技术

# 语义表达的多样性

The Dow Jones Industrial Average closed up  
255

Dow ends up

Dow gains 255  
points

Dow climbs 255

Stock market hits a  
record high



## 不同文本表达之间的关系:

- Equivalence:  $text1 \Leftrightarrow text2$  (paraphrasing)
- Entailment:  $text1 \Rightarrow text2$  the general case

# 文本推理(Textual Entailment)

## 定义

- 两个文本片段之间的有向关系:  
Text(t) and Hypothesis(h):

t entails h ( $t \Rightarrow h$ ) if  
humans reading t will infer that h is most likely true

- 说明:
  - 又称文本蕴涵
  - 基于人工标准
  - 假定具有通用的背景知识

# 背景知识的角色

- 文本推理实际上是说:
  - *text* AND *knowledge*  $\Rightarrow h$
- 但是
  - *knowledge* should not entail *h* alone
- 推理系统不能忽略text而直接验证h的真假

# 文本关系例子

- 蕴涵关系

T: Women form half the population in the country, yet are very poorly represented in parliament.

妇女占国家人口的一半，但在议会中的代表数却很少。

H: Women are poorly represented in parliament.

妇女在议会中的代表数很少。

- 矛盾关系

T: Santer succeeded Delors as president of the European Commission.  
Santer继承Delors成为欧盟委员会的主席。

H: Delors succeeded Santer in the presidency of the European Commission.

Delors继承欧盟委员会主席Santer。

- 无任何关系

T: 632 Air Canada flight attendants will lose their jobs in November.

632加拿大航空空中乘务员将在11月失去工作。

H: European Airlines are cutting jobs.

欧洲航空公司在裁员。

# 典型应用

- 问答系统
  - 答案提取与验证

*Question*  
Who bought Overture?

*Expected answer form*  
>> X bought Overture

Overture's acquisition  
by Yahoo

*text*

⇒  
*entails*

Yahoo bought Overture

*hypothesized answer*



# 典型应用

- 信息抽取

$X \text{ acquire } Y = \rangle$  抽取 “buy” 关系

- 多文档摘要

识别冗余信息

# 研究难点

- **大量背景知识的支持**
  - Beijing [located in] China
  - Barack [president of] America
  - EU=European Union
- **句式结构、语义表达的多样化**
  - X acquire Y vs. Y is bought by X
- **自然语言处理工具的局限性**
  - 词性标注
  - 命名实体识别
  - 实体消解
  - ...

# 文本推理数据

TAC RTE: 2005年以来已经连续举办了7届

## ✓ 任务形式

- RTE-1~RTE-3 : 2-way (推理/非推理关系)
- RTE-4~RTE-5 : 2-way + 3-way Task (推理/反义/无任何关系)
- RTE-6: Main Task + Novelty Detection Subtask
- RTE-7: similar with RTE-6

## ✓ 本文推理和反义识别均基于RTE评测语料

# 文本推理数据

- 对于给定输入文本对 $\langle T, H \rangle$ ，输出它们的推理关系，两个子任务：

## 2-way

- T能推出H
- T不能推出H

## 3-way

- T能推出H
- T与H相互矛盾
- T和H没有任何关系

- 评价指标：

准确率(Accuracy)

平均精度(Average Precision)

# 文本推理数据

- T的长度不断增加，难度增加

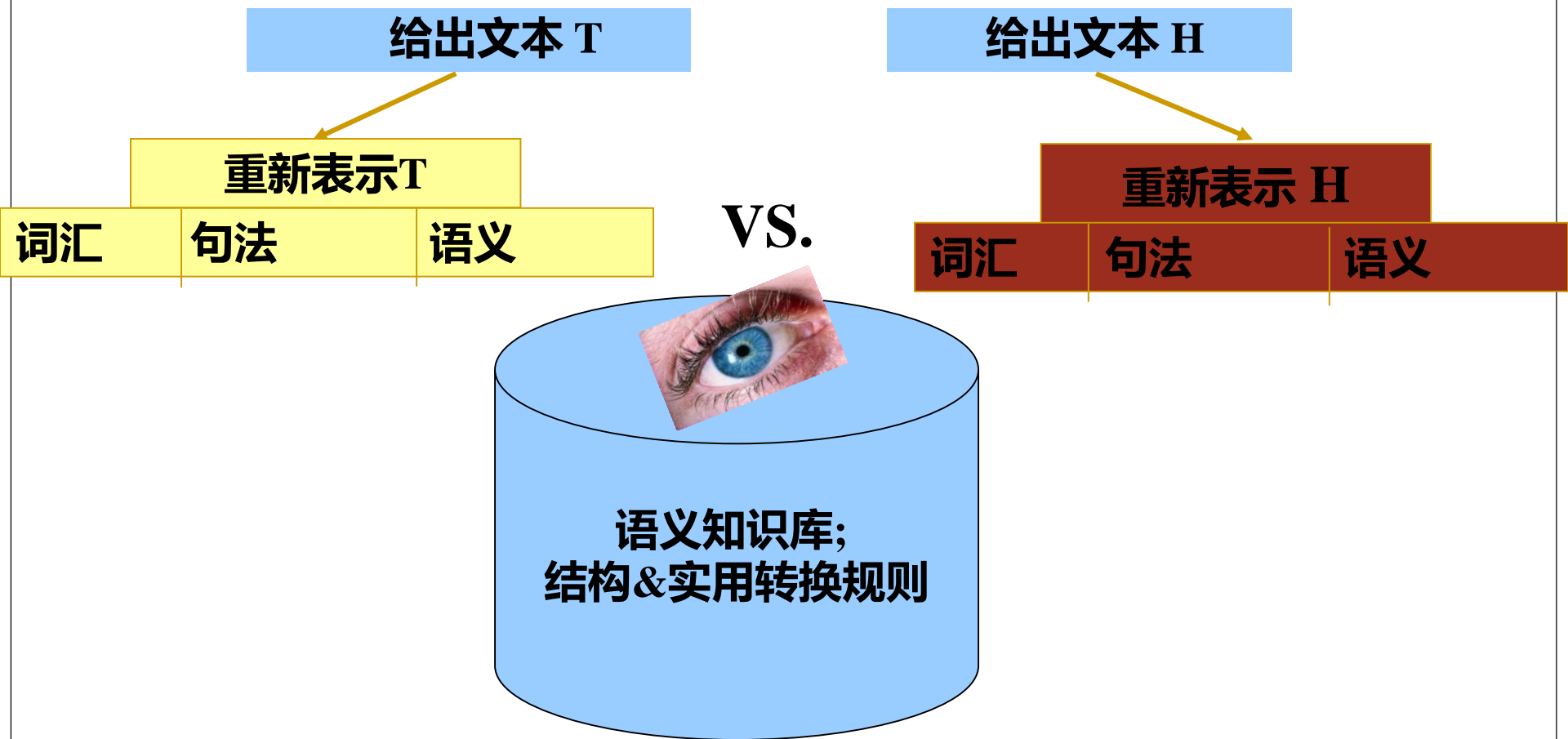
Challenge	Data Set	# of Pairs	H length (# of words)	T length (# of words)
RTE-1	DEV	567	10,08	24,78
	TEST	800	10,8	26,04
RTE-2	DEV	800	9,65	27,15
	TEST	800	8,39	28,37
RTE-3	DEV	800	8,46	34,98
	TEST	800	7,87	30,06
RTE-4	TEST	1000	7,7	40,15
RTE-5	DEV	600	7,79	99,49
	TEST	600	7,92	99,41

表1 RTE-1到RTE-5数据集

# 文本推理数据

- **RITE@NTCIR-9~NTCIR-11**
  - 日文、简体中文、繁体中文
  - 两类分类子任务: entailment or not
  - 多类分类子任务
    - Entailment: forward / reverse / bidirection
    - No entailment: contradiction / independence
- **Stanford SNLI**
  - 570k sentence pairs
  - Entailment/contradiction/neutral

# 文本推理通用策略

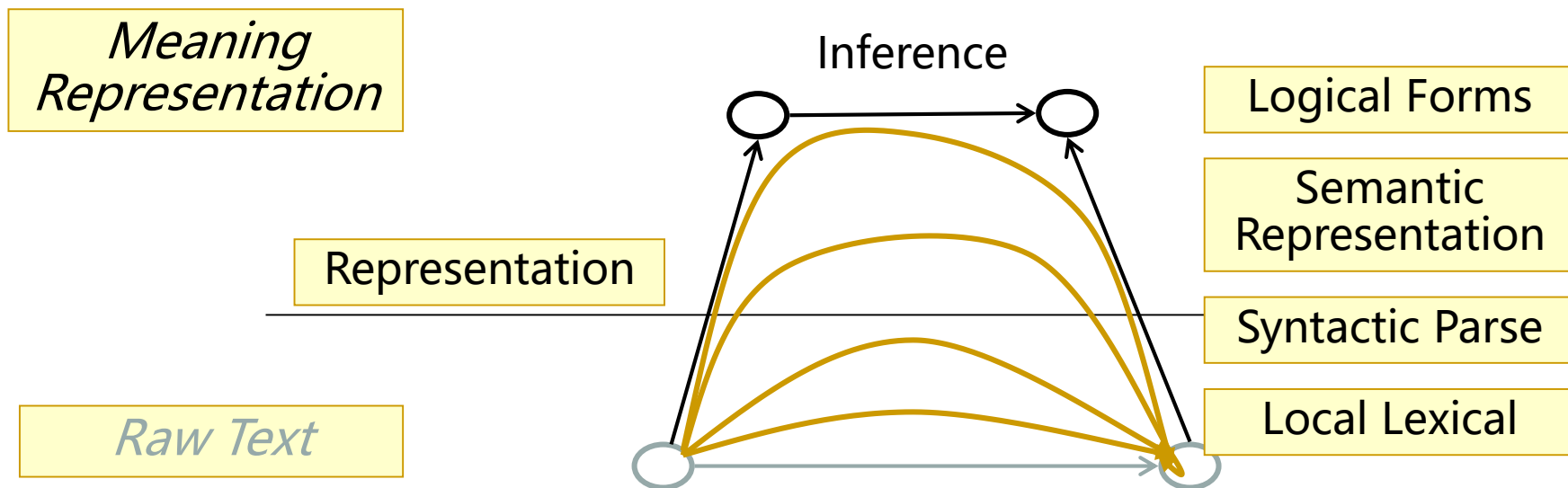


## 主要方法（从语言表示角度）

从自然语言表示角度，在词法层，句法层，语义层，将文本看成：

- 字符串形式
- 成分结构或依存关系结构
- 语义表达/逻辑表达式

## 基于各个表达层次都有一些文本推理方法提出来





# 主要方法（从任务角度）

从文本推理任务角度，将文本推理看成：

- **分类问题**

两文本段之间的推理关系分为两类(推理关系，非推理关系)，利用机器学习方法学习文本段间的特征

- **机器翻译问题**

两文本段，一个为候选译文(自动产生的翻译)，一个为参考译文(人工翻译)，借鉴机器翻译的方法来解决

- **变换问题**

由一个文本段经过一系列的操作变换到另一个文本段，比较变换代价。  
将两文本段看成依存关系树，树编辑距离

- **逻辑推理问题**

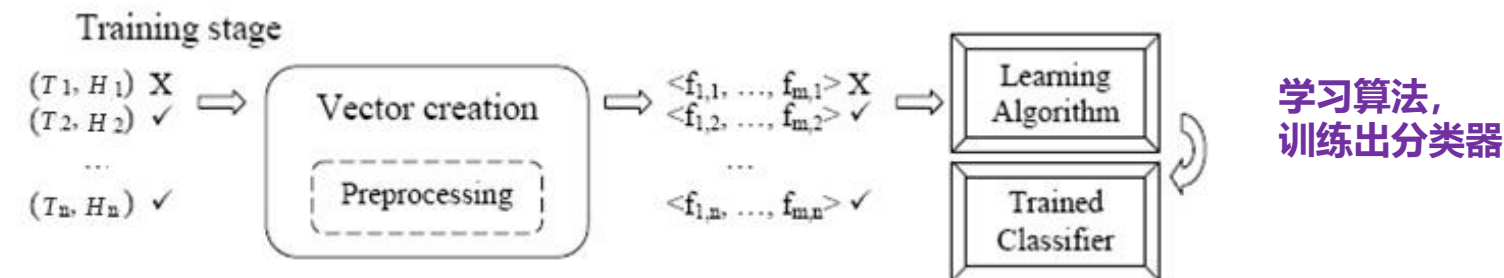
借鉴逻辑推理学的方法来解决

# 基于分类的方法

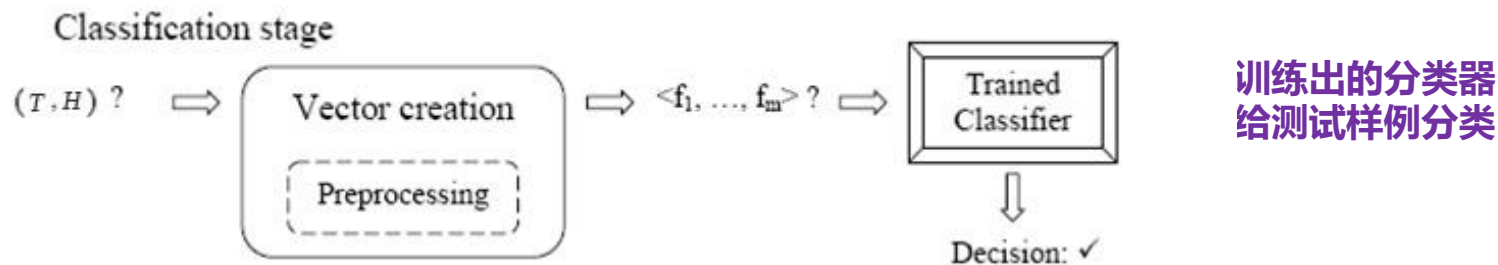
预处理过程

映射为一个特征向量

训练



测试



# 基于分类的方法

- **常用相似度度量**

将文本T和H看成字符串形式，一些作用于字符串的相似度度量可以用来计算T和H之间的相似程度

- Word Overlap
- Cosine similarity
- Longest Common Substring
- Levenshtein distance
- Jaro-Winkler distance
- Manhattan distance
- Euclidean distance
- Jaccard coefficient
- Dice coefficient
- Matching coefficient
- ...

# 基于分类的方法

- 语言学特征

利用一些规则，将T和H中的具有相同或相反意义的词或词组进行匹配或配对，可以为推理判断提取一些有效特征，并且常作为一些较复杂的系统一个必要步骤

- 同义词特征

利用WordNet (或类似的资源)寻找T和H中表达含义相同的词或词组

discovery和reveal同指 “发现”

Oscar和Academy Award 同指 “奥斯卡金像奖”

- Acronym特征(缩略词特征)

UNDP与United Nations Development Programme

联合国开发计划署

# 基于分类的方法

- 语言学特征

- 数字，日期和时间特征

T和H中如果出现数字，日期和时间这些信息，很多时候判断T能否推出H取决于它们能否匹配上，需要对时间表达式进行归一化

T: The opinion poll was conducted on the **sixth** and **seventh** of October, and included a cross section of 861 adults with a margin of error estimated at 4%.

H: The poll was carried out on the **6th** and **7th** of October.

- 反义词特征

主要是指动词之间的，如果T和H中的某两个名词匹配上(利用前述的几个规则)，并且都是某个动词的主语或宾语，那么这时候检查两个动词是很必要的。如果动词对为反义或对立关系，那么T和H表达意义就是对立的。

...

# 基于分类的方法

- Zanzotto(2006)等人率先提出了跨T-H对(cross-pair)特征

不是计算T和H两个文本段之间的相似度或其它语言学特征  
而是计算不同的T-H对之间的相似度

(a)T: Yahoo bought Overture.

(b)H: Yahoo owns Overture.

X bought Y  $\rightarrow$  X owns Y  
(X买了Y  $\rightarrow$  X拥有了Y)

(c)T: Wanadoo bought Kstones.

(d)H: Wanadoo owns Kstones.

(a)-(b)为训练样本中正确的推理关系对, (c)-(d)为测试样本中样例

# 背景知识库

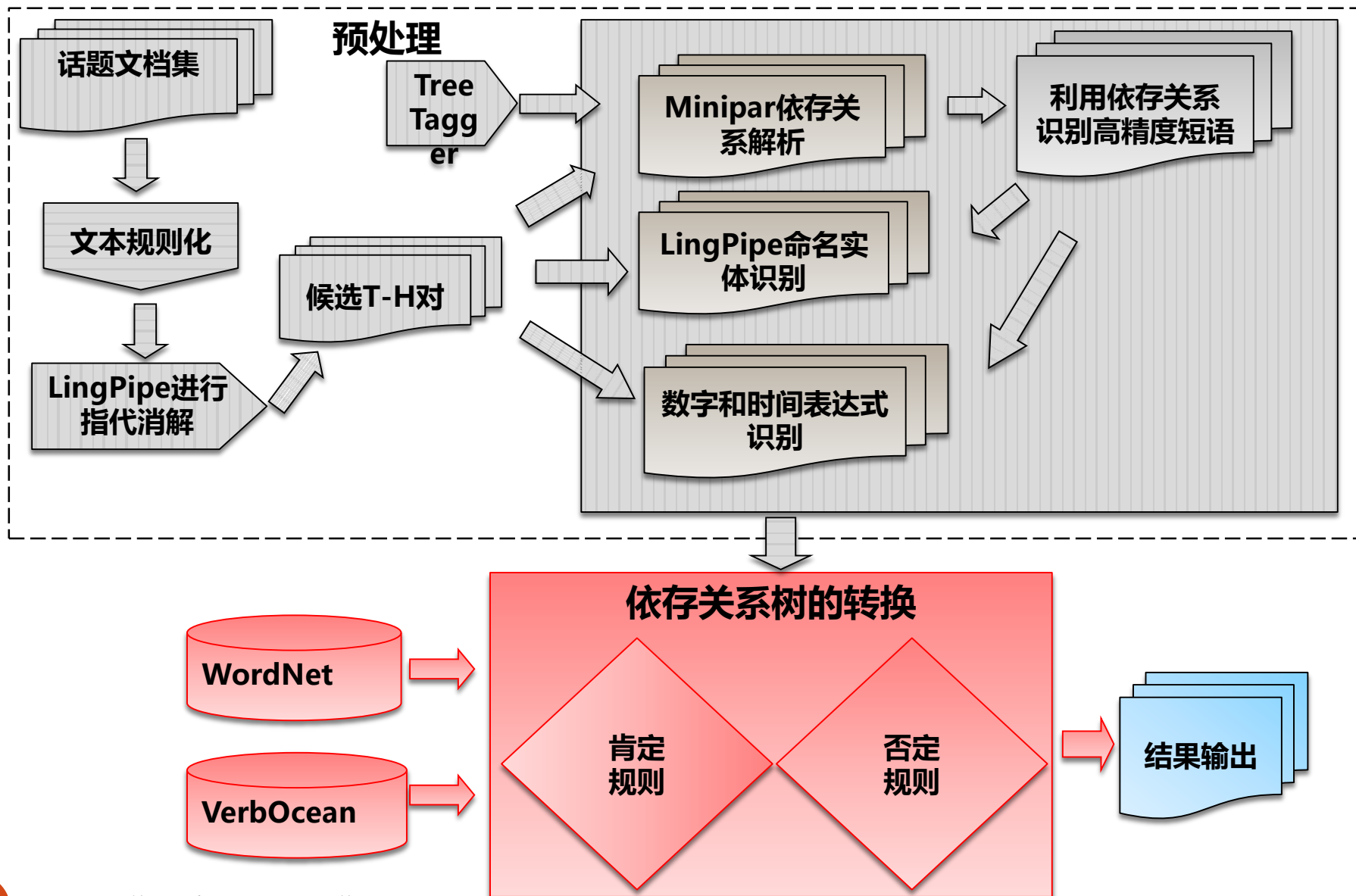
- 人工整理
  - WordNet
  - FrameNet
  - VerbNet
  - .....
  - ✓ 准确率高
  - ✓ 覆盖面不够
- ▶ 自动获取
  - DIRT
  - TEASE
  - VerbOcean
  - .....
  - ✓ 准确率不高
  - ✓ 覆盖面相对广

# 其他知识资源

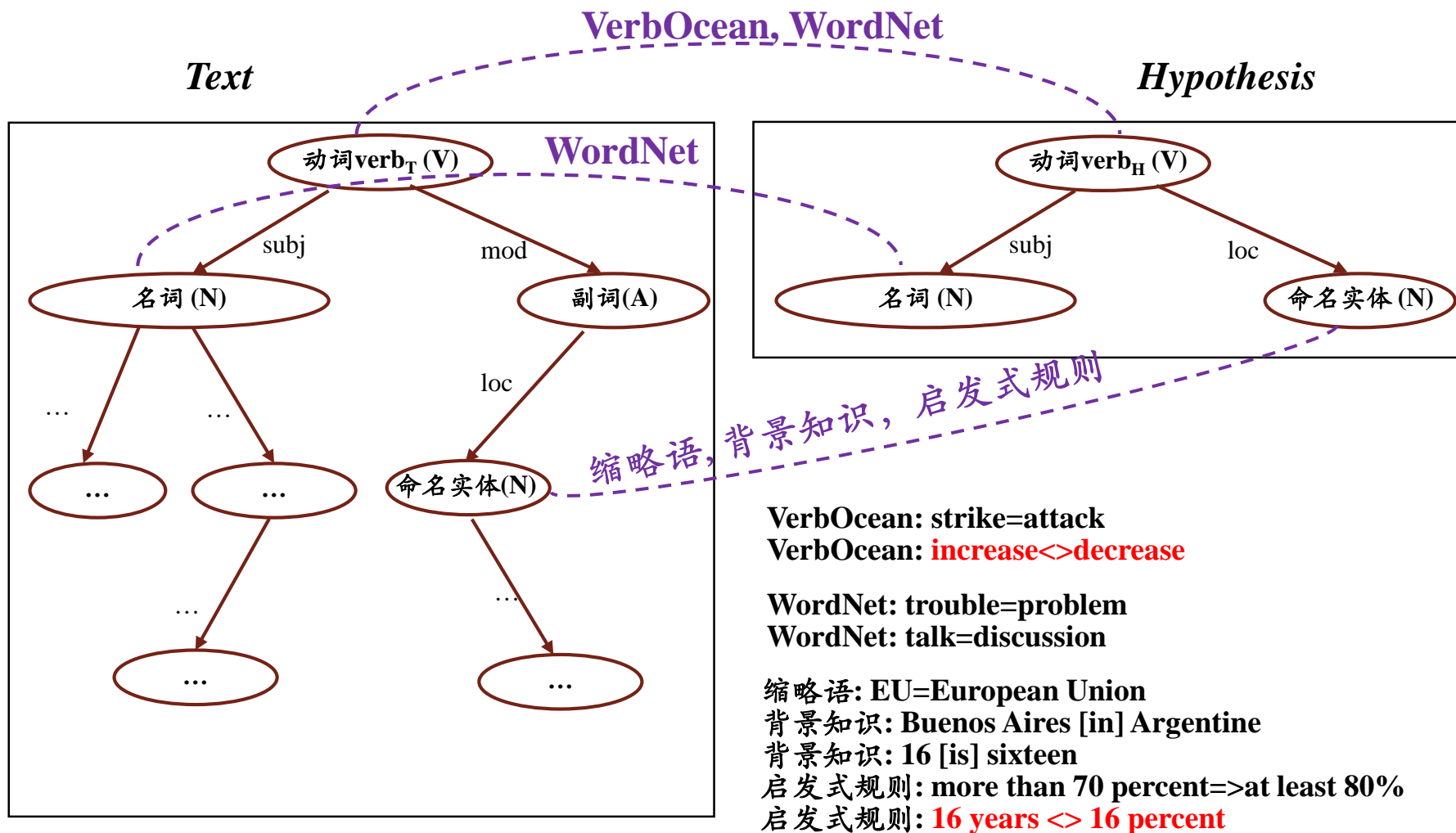
- 复述规则
  - 可从RTE语料中自动获取
    - person killed --> claimed one life
    - hand reins over to --> give starting job to
    - same-sex marriage --> gay nuptials
    - cast ballots in the election -> vote
    - dominant firm --> monopoly power
    - death toll --> kill
    - try to kill --> attack
    - lost their lives --> were killed
    - left people dead --> people were killed



# PKUTM@RTE-6系统(Jia et al. 2010)



# PKUTM@RTE-6系统 依存关系树的转换(H->T)



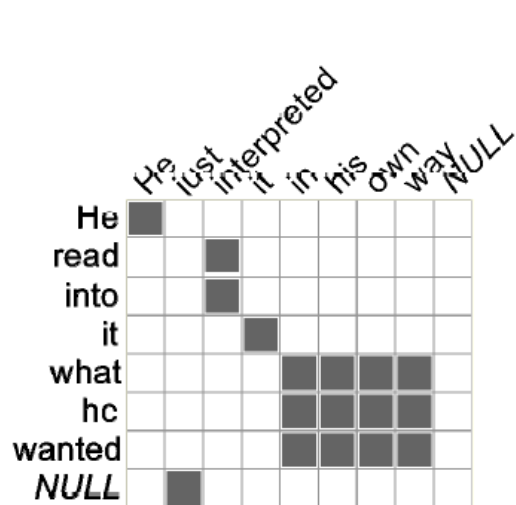
# RTE-6结果 — Main Task

18组参加了Main Task，提交了48组结果，各组队伍的最好结果

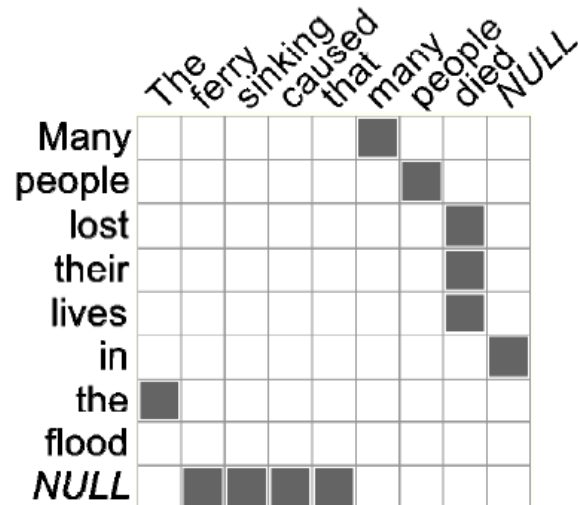
	RUN	Precision	Recall	F-measure	Participation
		n			
1	PKUTM2	68.57	36.93	48.01	ICST, Peking University(北京大学)
2	deb_iitb2	53.43	42.86	47.56	Bombay Powai (印度孟买)
3	IKOMA1	39.71	51.43	44.81	NEC Corporation (日本NEC公司)
4	FBK_irst3	43.46	46.03	44.71	University of Trento (意大利特伦托大学)
5	Boeing1	55.1	36.61	43.99	The Boeing Company(美国波音公司)
6	DirRelCond21	38.99	41.8	40.35	Alpl'ar Perini
7	DFKI2	55.94	30.9	39.81	Saarland University (德国萨尔州大学)
8	SJTU_CIT3	34.35	46.67	39.57	Shanghai Jiao Tong University(上海交通大学)
9	BIU1	37.54	37.46	37.5	Bar-Ilan University (以色列巴伊兰大学)
10	JU_CSE_TAC1	38.63	31.64	34.79	Jadavpur University(印度贾达沃普尔大学)
11	UIUC3	31.53	33.86	32.65	UIUC (美国伊利诺大学)
12	Sangyan1	21.66	46.03	29.46	NEC HCL System Technologies Ltd.(印度NEC公司)
13	SINAI2	23.27	30.69	26.47	University of Ja'en (西班牙哈恩大学)
14	UAIC20101	22.89	27.2	24.85	"Al. I. Cuza" University (罗马尼亚亚历雅西大学)
15	Sagan1	15.98	48.89	24.09	National University of Cordoba(阿根廷科尔多瓦国立大学)
16	UB.dmirg3	11.79	48.68	18.98	University of Ballarat (澳大利亚巴拉瑞特大学)
17	budapestacad1	13.35	31.22	18.71	Hungarian Academy of Sciences(匈牙利科学院)
18	saicnlp1	7.92	21.69	11.6	Science Applications International Corporation(美国科学应用国际公司)

# BCMI-NLP@NTCIR-10系统 (Wang et al. 2010)

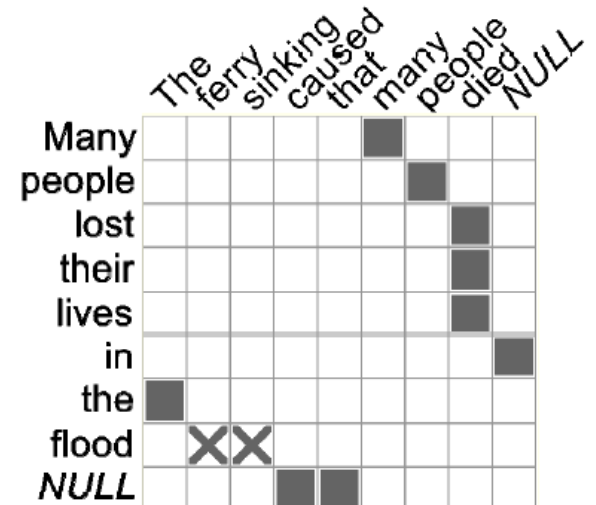
基于有监督分类方法，采用了基于词对齐的很多特征；  
利用人工标注的词汇对齐语料训练了词对齐器；



(a) Alignment on entailment pair

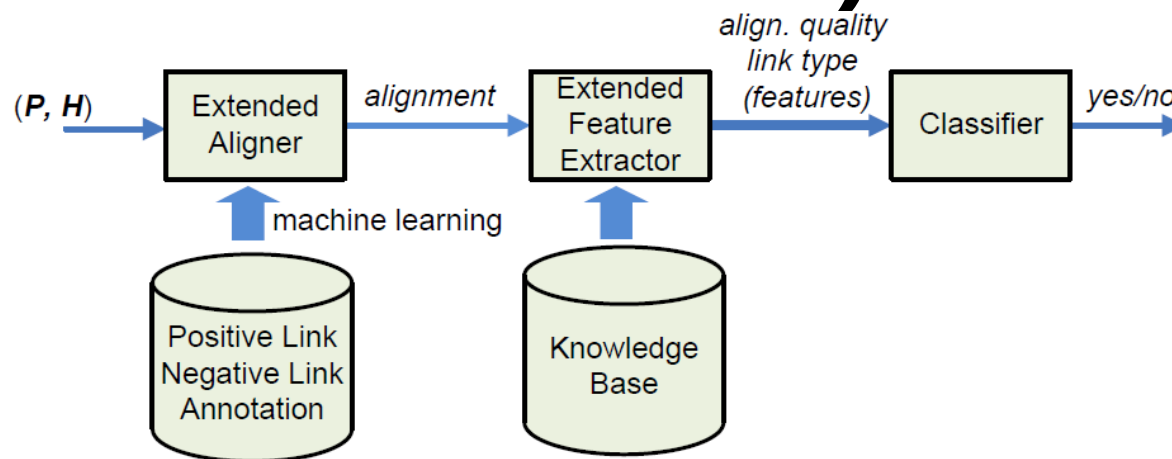


(b) Weakness on non-entailment pair



(c) Labeled alignment on non-entailment pair

# BCMI-NLP@NTCIR-10系统 (Wang et al. 2010)



Category	Feature
Align.	Confidence score of the aligner
Quality	Ratio of linked words in $t_1$
Link Type	Whether $e_1$ and $e_2$ are in an antonym list <sup>a</sup> Whether $e_1$ and $e_2$ are in an synonym list Whether $e_1$ and $e_2$ are unequal numbers Whether $e_1$ and $e_2$ are different named entities Relation of $e_1$ and $e_2$ in an ontology (hyponym, sibling, etc.) Ontology-based similarities of $e_1$ and $e_2$ Count of common characters Length of the common prefixes Length of the common suffix Tuple <sup>b</sup> of the syntactic tags <sup>c</sup> Tuple of the ancestors in an ontology Tuple of whether $e_1$ or $e_2$ is in a list of negative expressions Tuple of whether $e_1$ or $e_2$ is the head of a noun phrase

# NTCIR-10结果

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
bcNLP-CS-BC-03	73.84	74.65	78.43	72.58	85.31	69.25	78.25	62.12
MIG-CS-BC-02	68.09	68.50	71.72	69.64	73.93	64.45	66.97	62.12
CYUT-CS-BC-03	67.86	68.12	70.74	70.16	71.33	64.98	65.63	64.35
bcNLP-CS-BC-01	67.04	69.65	76.32	65.98	90.52	57.75	80.20	45.13
bcNLP-CS-BC-02	66.89	69.91	76.89	65.71	92.65	56.88	83.33	43.18
MIG-CS-BC-01	65.71	65.81	67.56	69.33	65.88	63.87	62.11	65.74
CYUT-CS-BC-02	63.11	63.12	62.50	69.36	56.87	63.73	58.16	70.47
WHUTE-CS-BC-02	61.65	66.58	75.40	62.60	94.79	47.90	84.51	33.43
CYUT-CS-BC-01	61.17	61.59	57.14	71.94	47.39	65.20	55.86	78.27
*IASL-CS-BC-02	60.45	63.25	70.98	61.90	83.18	49.91	66.82	39.83
WHUTE-CS-BC-01	58.20	64.79	74.79	60.99	96.68	41.61	87.50	27.30
MIG-CS-BC-03	57.19	63.64	73.80	60.42	94.79	40.59	81.51	27.02
IMTKU-CS-BC-03	54.28	62.74	73.95	59.42	97.87	34.61	89.53	21.45
Yuntech-CS-BC-03	53.52	59.54	70.24	58.28	88.39	36.80	65.25	25.63
Yuntech-CS-BC-02	52.10	59.03	70.32	57.77	89.81	33.88	65.60	22.84
Yuntech-CS-BC-01	50.91	58.64	70.39	57.40	91.00	31.42	66.07	20.61
IMTKU-CS-BC-01	50.82	60.31	72.42	57.98	96.45	29.22	81.01	17.83
*IASL-CS-BC-01	50.60	54.03	63.63	55.58	74.41	37.57	50.00	30.08
WUST-CS-BC-02	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
WUST-CS-BC-01	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
*WUST-CS-BC-01	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
WUST-CS-BC-03	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
IMTKU-CS-BC-02	50.12	60.31	72.66	57.87	97.63	27.57	85.51	16.43
JUNLP-CS-BC-01	48.49	48.66	51.39	52.61	50.24	45.59	44.44	46.80

Table 10: Results on BC subtask (CS).



# NTCIR-10结果

Team	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
bcNLP-CS-MC-03	56.82	61.08	66.67	77.27	58.62	67.30	53.91	89.53	38.41	64.44	27.36	54.89	69.28	45.45
*IASL-CS-MC-02	50.94	53.91	55.30	61.34	50.34	64.44	57.51	73.29	38.42	40.21	36.79	45.59	50.00	41.90
WHUTE-CS-MC-01	46.79	54.80	61.54	62.41	60.69	64.36	49.90	90.61	18.71	39.39	12.26	42.58	73.08	30.04
WHUTE-CS-MC-02	46.53	56.59	62.25	59.87	64.83	65.09	51.24	89.17	8.26	33.33	4.72	50.53	75.59	37.94
bcNLP-CS-MC-02	44.88	57.62	59.68	71.84	51.03	67.86	52.58	95.67	0.00	0.00	0.00	51.99	63.79	43.87
MIG-CS-MC-02	44.74	51.60	58.50	49.07	72.41	52.84	57.69	48.74	11.35	22.86	7.55	56.26	52.01	61.26
CYUT-CS-MC-02	42.52	48.78	53.64	51.59	55.86	56.13	48.80	66.06	12.42	18.18	9.43	47.87	55.15	42.29
MIG-CS-MC-01	41.82	49.17	56.44	50.83	63.45	50.70	57.27	45.49	5.48	10.00	3.77	54.64	47.65	64.03
Yuntech-CS-MC-02	40.91	51.22	55.02	51.83	58.62	64.81	49.90	92.42	13.43	32.14	8.49	30.40	65.79	19.76
Yuntech-CS-MC-03	40.89	51.22	53.95	51.57	56.55	65.15	50.20	92.78	13.43	32.14	8.49	31.04	63.41	20.55
WUST-CS-MC-02	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
WUST-CS-MC-03	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
*WUST-CS-MC-01	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
CYUT-CS-MC-01	40.37	47.63	60.34	59.33	61.38	56.72	44.94	76.90	12.31	33.33	7.55	32.12	46.62	24.51
WUST-CS-MC-01	40.33	51.73	59.31	54.65	64.83	62.20	47.86	88.81	3.57	33.33	1.89	36.26	69.66	24.51
Yuntech-CS-MC-01	40.33	50.70	53.42	50.62	56.55	64.56	49.61	92.42	13.64	34.62	8.49	29.70	63.64	19.37
CYUT-CS-MC-03	40.10	51.09	48.73	51.54	46.21	57.07	44.31	80.14	0.00	0.00	0.00	54.59	73.33	43.48
bcNLP-CS-MC-01	39.95	53.91	43.43	81.13	29.66	64.70	49.07	94.95	0.00	0.00	0.00	51.69	59.90	45.45
*IASL-CS-MC-01	34.95	41.74	37.29	48.35	30.34	59.39	47.05	80.51	25.24	26.00	24.53	17.89	28.45	13.04
MIG-CS-MC-03	34.42	43.15	53.90	39.80	83.45	58.58	51.96	67.15	12.94	13.68	12.26	12.27	70.83	6.72
IMTKU-CS-MC-03	27.26	40.20	9.81	10.83	8.97	67.10	52.67	92.42	32.14	25.86	42.45	0.00	0.00	0.00
JUNLP-CS-MC-01	24.38	24.71	22.42	19.59	26.21	27.00	32.49	23.10	22.02	16.29	33.96	26.07	32.54	21.74
IMTKU-CS-MC-01	23.89	37.64	5.85	10.00	4.14	63.20	48.73	89.89	22.82	17.71	32.08	3.69	27.78	1.98
IMTKU-CS-MC-02	19.67	36.11	7.73	12.90	5.52	57.78	41.73	93.86	8.74	10.39	7.55	4.41	31.58	2.37

Table 17: Results on MC subtask (CS).

# Stanford SNLI 结果

## Three-way classification

Publication	Model	Parameters	Train (% acc)	Test (% acc)
<b>Feature-based models</b>				
Bowman et al. '15	Unlexicalized features		49.4	50.4
Bowman et al. '15	+ Unigram and bigram features		99.7	78.2
<b>Sentence encoding-based models</b>				
Bowman et al. '15	100D LSTM encoders	220k	84.8	77.6
Bowman et al. '16	300D LSTM encoders	3.0m	83.9	80.6
Vendrov et al. '15	1024D GRU encoders w/ unsupervised 'skip-thoughts' pre-training	15m	98.8	81.4
Mou et al. '15	300D Tree-based CNN encoders	3.5m	83.3	82.1
Bowman et al. '16	300D SPINN-PI encoders	3.7m	89.2	83.2
Yang Liu et al. '16	600D (300+300) BiLSTM encoders	2.0m	86.4	83.3
Munkhdalai & Yu '16b	300D NTI-SLSTM-LSTM encoders	4.0m	82.5	83.4
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention	2.8m	84.5	84.2
Conneau et al. '17	4096D BiLSTM with max-pooling	40m	85.6	84.5
Munkhdalai & Yu '16a	300D NSE encoders	3.0m	86.2	84.6
Qian Chen et al. '17	600D (300+300) Deep Gated Attn. BiLSTM encoders (code)	12m	90.5	85.5
Tao Shen et al. '17	300D Directional self-attention network encoders (code)	2.4m	91.1	85.6
Jihun Choi et al. '17	300D Gumbel TreeLSTM encoders	2.9m	91.2	85.6
Nie and Bansal '17	300D Residual stacked encoders	9.7m	89.8	85.7
Yi Tay et al. '18	300D CAFE (no cross-sentence attention)	3.7m	87.3	85.9
Jihun Choi et al. '17	600D Gumbel TreeLSTM encoders	10m	93.1	86.0
Nie and Bansal '17	600D Residual stacked encoders	29m	91.0	86.0
Tao Shen et al. '18	300D Reinforced Self-Attention Network	3.1m	92.6	86.3
Im and Cho '17	Distance-based Self-Attention Network	4.7m	89.6	86.3



# Stanford SNLI 结果

## Other neural network models

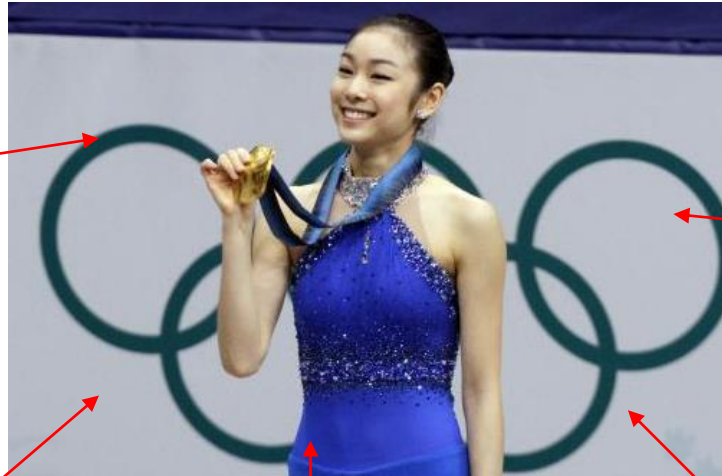
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Jianpeng Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Jianpeng Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Zhiguo Wang et al. '17	BiMPM	1.6m	90.9	87.5
Lei Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, <a href="#">code</a> )	4.4m	91.2	88.0
McCann et al. '17	Biattentive Classification Network + CoVe + Char	22m	88.5	88.1
Ghaeini et al. '18	450D DR-BiLSTM	7.5m	94.1	88.5
Yi Tay et al. '18	300D CAFE	4.7m	89.8	88.5
Qian Chen et al. '17	KIM	4.3m	94.1	88.6
Qian Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM ( <a href="#">code</a> )	7.7m	93.5	88.6
Peters et al. '18	ESIM + ELMo	8.0m	91.6	88.7
Zhiguo Wang et al. '17	BiMPM <b>Ensemble</b>	6.4m	93.2	<b>88.8</b>
Yichen Gong et al. '17	448D Densely Interactive Inference Network (DIIN, <a href="#">code</a> ) <b>Ensemble</b>	17m	92.3	<b>88.9</b>
Qian Chen et al. '17	KIM <b>Ensemble</b>	43m	93.6	<b>89.1</b>
Ghaeini et al. '18	450D DR-BiLSTM <b>Ensemble</b>	45m	94.8	<b>89.3</b>
Peters et al. '18	ESIM + ELMo <b>Ensemble</b>	40m	92.1	<b>89.3</b>
Yi Tay et al. '18	300D CAFE <b>Ensemble</b>	17.5m	92.5	<b>89.3</b>

# 文本复述技术

# 定义

- 可看作是双向的文本推理关系:  $T1 \Leftrightarrow T2$
- 复述(Paraphrase)
  - 作为名词
    - 同一意义的不同表达
  - 作为动词
    - 为输入表达生成复述
- “相同意义” ?
  - 比较主观
  - 不同严格程度
  - 依赖于具体应用

Paraphrase (noun): Alternative expressions of the same meaning



金妍儿

Korean **Kim Yuna** won **gold** with a world-record score in women's figure skating at the **Vancouver Olympics** Thursday.

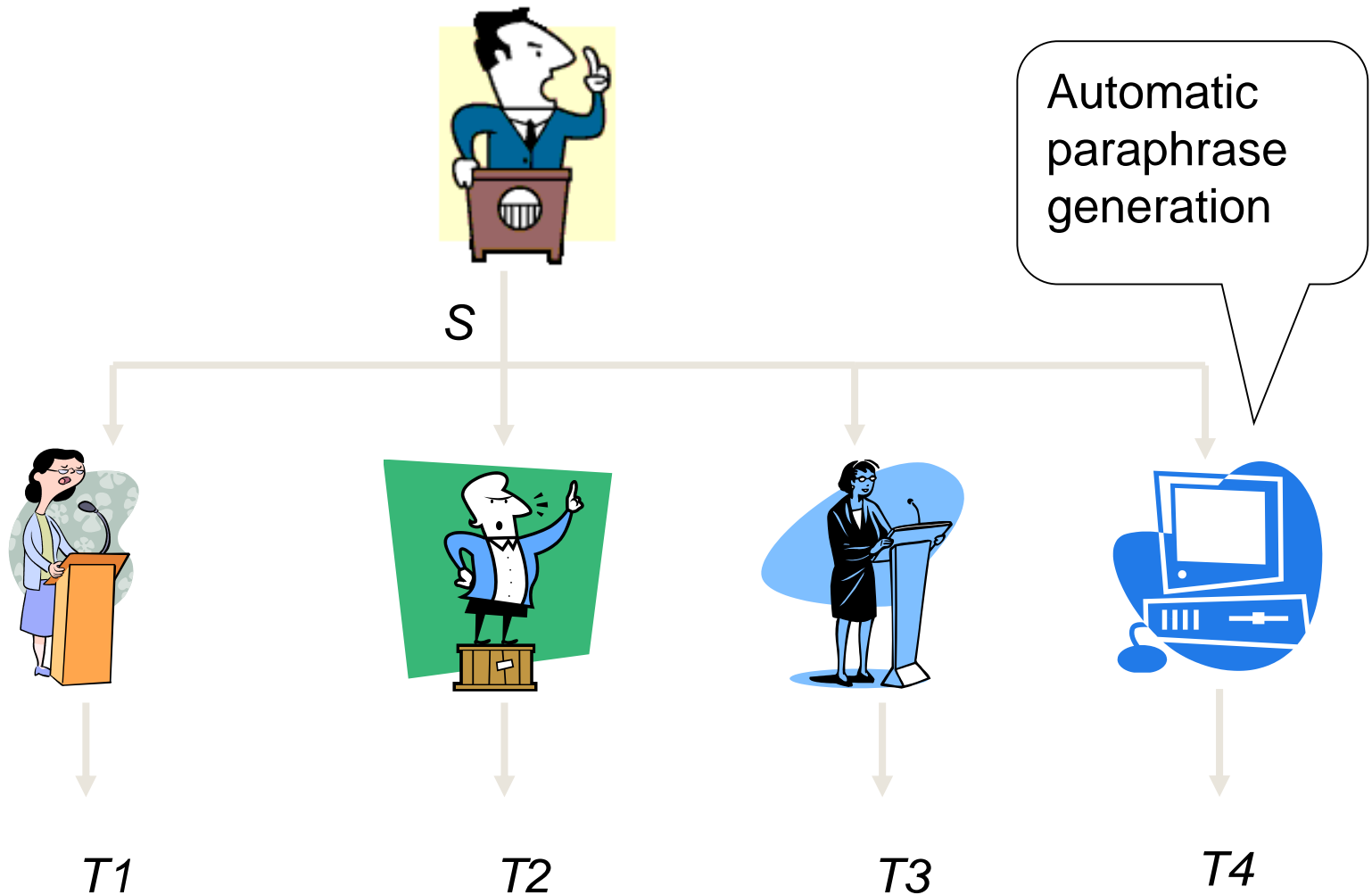
Korean figure skater **Kim Yuna** has **won** the **gold medal** of women's figure skating at the **Winter Olympics in Vancouver**

**Kim Yu-Na** (19) is a South Korean ice skater who **took the gold medal** at the Vancouver Olympics.

**Yuna Kim** of South Korea **won** the women's figure skating **gold medal** at the Vancouver Olympics in record fashion.

**Kim Yuna**, a South Korean figure skater has **won the gold medal** at the on-going **Winter Olympics 2010**.

Paraphrase (verb): Generate paraphrases for an input  $S$ .



# 复述的类别

- **根据不同粒度**
  - **表层复述(Surface paraphrases)**
    - Lexical level
    - Phrase level
    - Sentence level
    - Discourse level
  - **结构复述(Structural paraphrases)**
    - Pattern level
    - Collocation level

# 复述举例

- Lexical paraphrases (generally synonyms)
  - *solve* and *resolve*; he (*works / teaches*) in a school, as a head teacher
- Paraphrase phrases
  - *look after* and *take care of*
  - *The US government* and *the US administration*
- Paraphrase sentences
  - *The table was set up in the carriage shed.*
  - *The table was laid under the cart-shed.*
- Paraphrase patterns
  - *[X] considers [Y]*
  - *[X] takes [Y] into consideration*
- Paraphrase collocations
  - *(turn on, OBJ, light)*
  - *(switch on, OBJ, light)*

# 复述的类别

- **根据复述样式**
  - Trivial change
  - Phrase replacement
  - Phrase reordering
  - Sentence split & merge
  - Complex paraphrases



# 复述举例

- Trivial change
  - *all the members of* and *all members of*
- Phrase replacement
  - *He said there will be major cuts in the salaries of high-level civil servants.*
  - *He said there will be major cuts in the salaries of senior officials.*
- Phrase reordering
  - *Last night, I saw Tom in the shopping mall.*
  - *I saw Tom in the shopping mall last night.*
- Sentence split & merge
  - *He bought a computer, which is very expensive.*
  - *(1) He bought a computer. (2) The computer is very expensive.*
- Complex paraphrases
  - *He said there will be major cuts in the salaries of high-level civil servants.*
  - *He claimed to implement huge salary cut to senior civil servants.*

# 复述的应用

- Machine Translation (MT)
  - Simplify input sentences
  - Alleviate data sparseness
  - Parameter tuning
  - Automatic evaluation
- Question Answering (QA)
  - Question reformulation
- Information Extraction (IE)
  - IE pattern expansion
- Information Retrieval (IR)
  - Query reformulation
- Summarization
  - Sentence clustering
  - Automatic evaluation
- Natural Language Generation (NLG)
  - Sentence rewriting
- Others
  - Changing writing style
  - Text simplification
  - Identifying plagiarism
  - Text steganography
  - .....

# 有关复述的研究工作

- **复述判别(Paraphrase identification)**
  - Identify (sentential) paraphrases
- **复述抽取(Paraphrase extraction)**
  - Extract paraphrase instances (different granularities)
- **复述生成(Paraphrase generation)**
  - Generate (sentential) paraphrases
- **复述应用(Paraphrase applications)**
  - Apply paraphrases in other areas


# 复述判别

- 特别地，针对句子级别的复述判别
  - 给定一对句子，自动判别这对句子是否为复述
- 复述判别并不容易

Susan **often** goes to see movies with her boyfriend.  
Susan **never** goes to see movies with her boyfriend.



He **said** there will be major cuts in the salaries of high-level civil servants.  
He **claimed** to implement huge salary cut to senior civil servants.



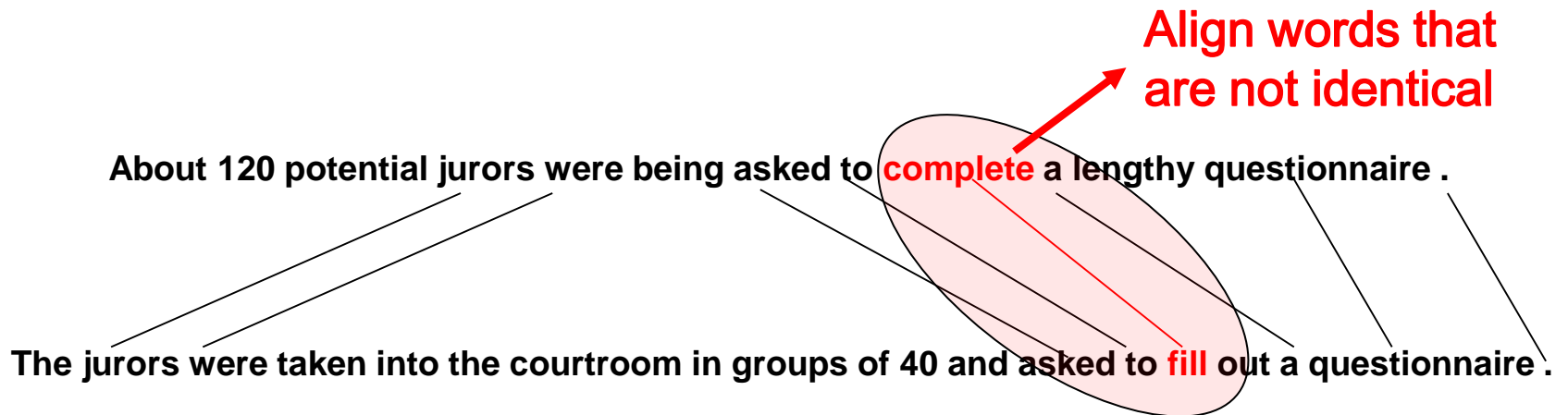
# 复述判别

- 基于分类的方法
  - 二类分类问题：将  $s_1$  and  $s_2$  作为分类器输入，输出0/1
  - 分类特征：不同程度上句子的相似度
    - [Malakasiotis, 2009]
      - String similarity (various levels)
        - Tokens, stems, POS tags, nouns only, verbs only, ...
      - Different measures
        - Edit distance, Jaro-Winkler distance, Manhattan distance...
      - Synonym similarity
        - Treat synonyms in two sentences as identical words
      - Syntax similarity
        - Dependency parsing of two sentences and compute the overlap of dependencies

# 复述判别

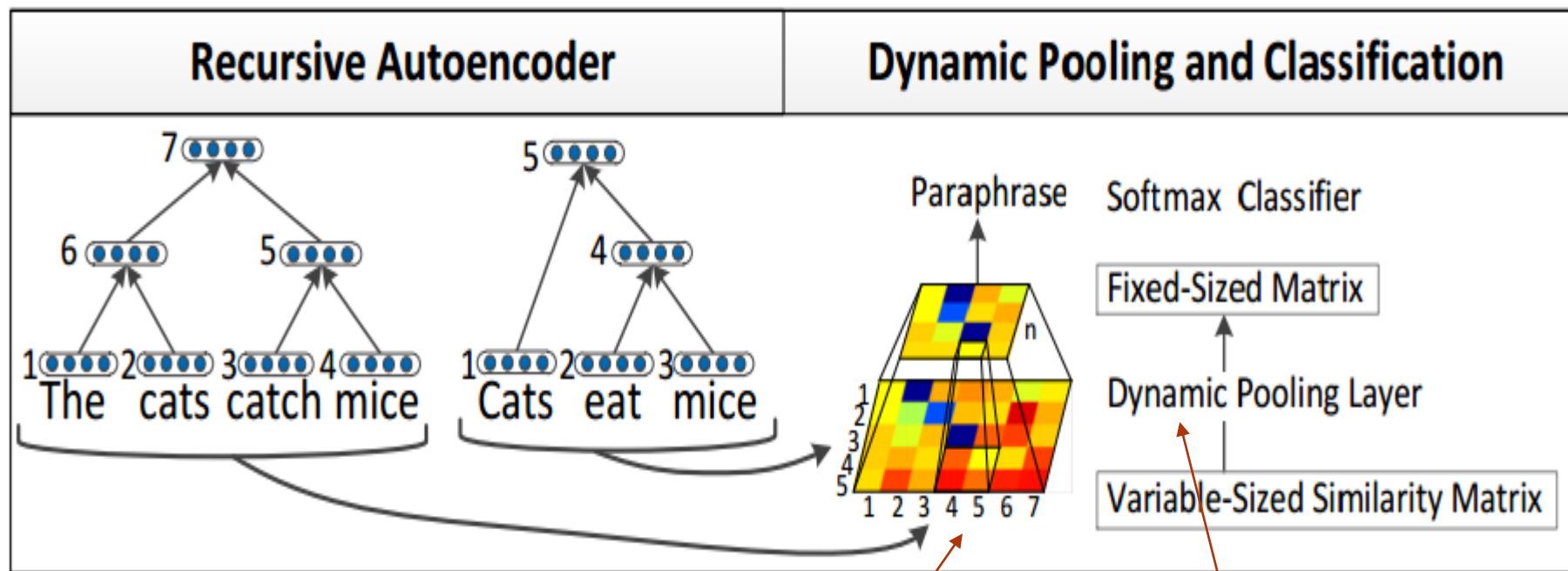
- 基于对齐的方法

- 将 $s_1$ 与 $s_2$ 对齐, 然后基于对齐结果计算得分
- 基于准同步依存文法 (QG)进行对齐 [Das and Smith, 2009]
  - 对两个依存树进行对齐
  - 假设: 两个复述句子的依存树应该能够紧密对齐



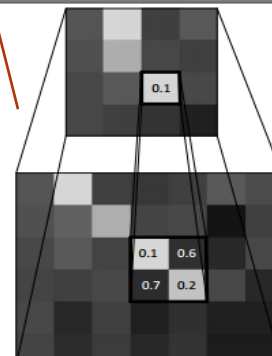
# 复述判别

- 基于深度学习的方法[Socher et al., 2011]



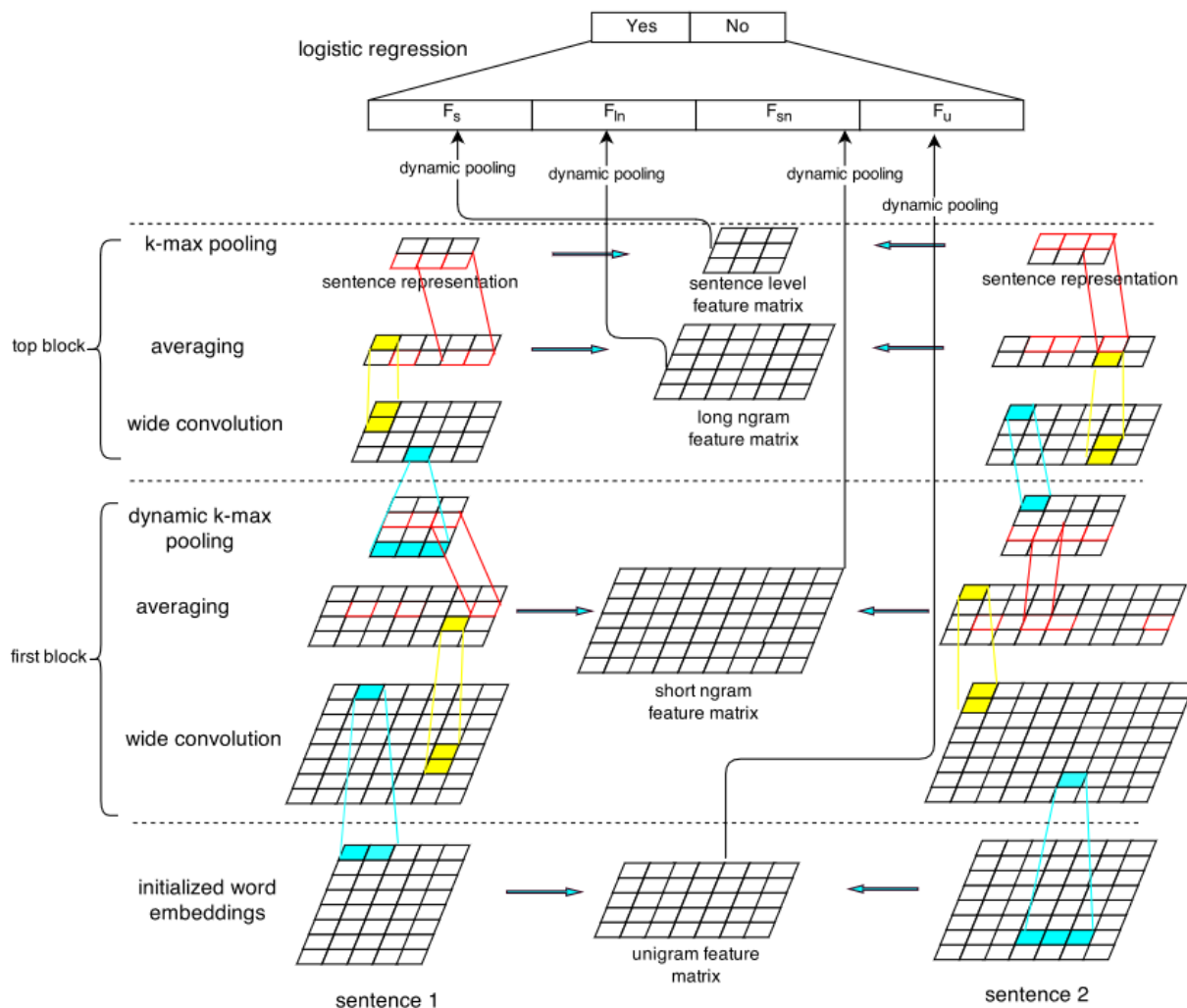
丰富的paraphrase corpus:  
MSRP  
PPDB  
TwitterPPDB

词及短语的欧式距离矩阵



# 复述判别

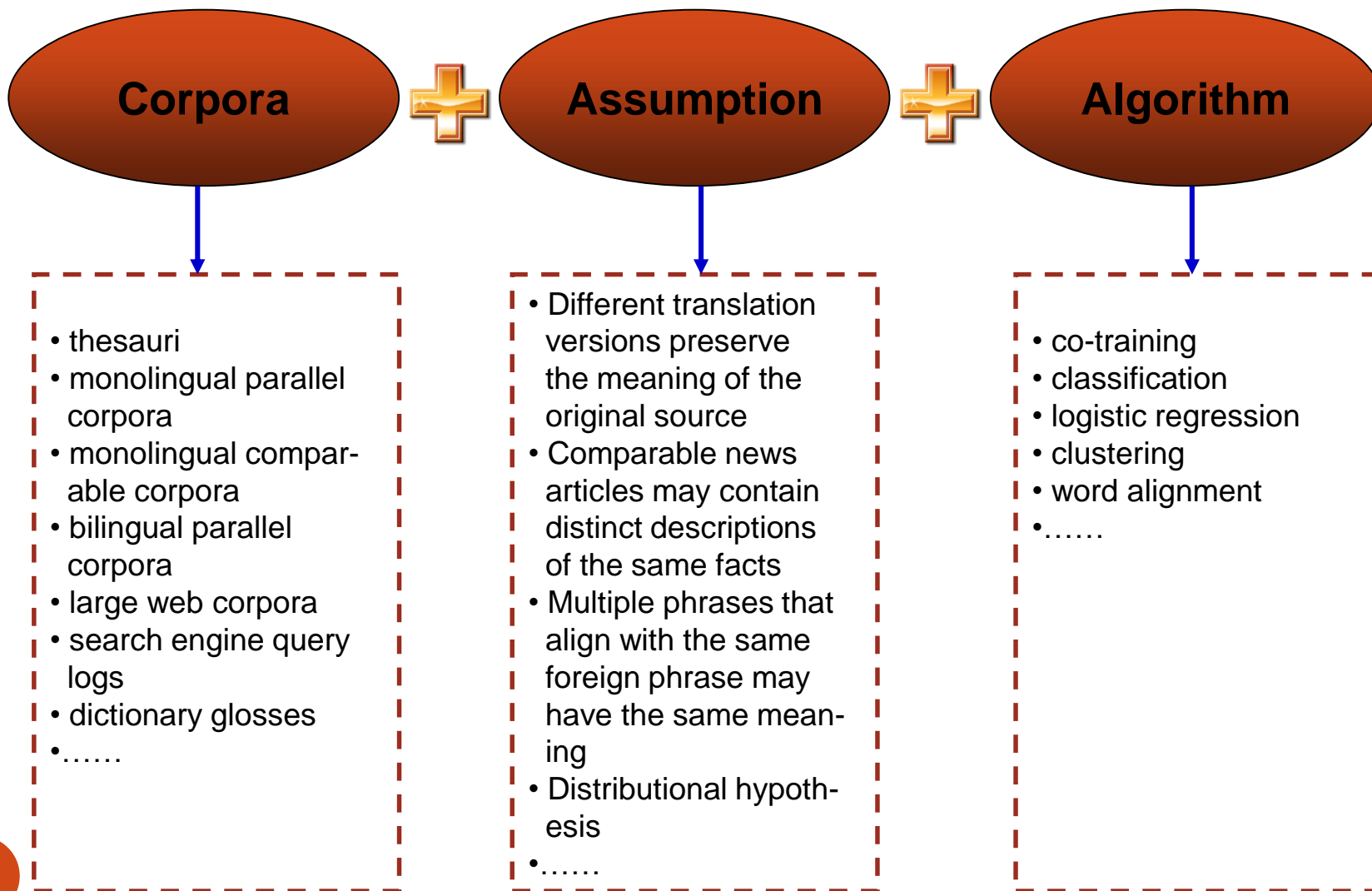
- 基于深度学习的方法[Yin and Schutze, 2015]



丰富的paraphrase  
MSRP  
PPDB  
TwitterPPDB

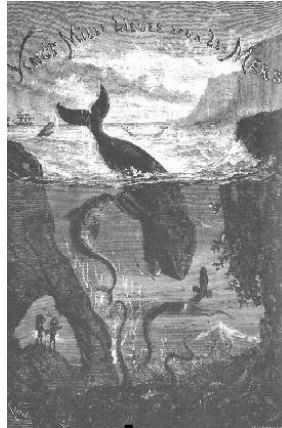


# 复述抽取



# 从单语言平行语料中进行复述抽取（一）

- 语料
  - 同一外文作品的多种翻译版本
- 假设
  - 不同翻译版本保留了原作品的意思，但可能使用不同的语言表达

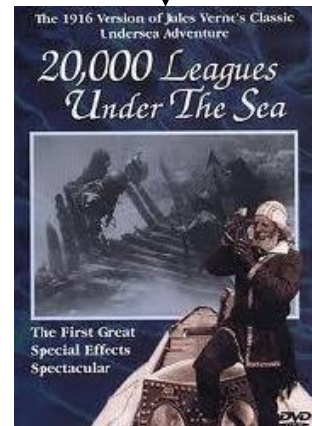
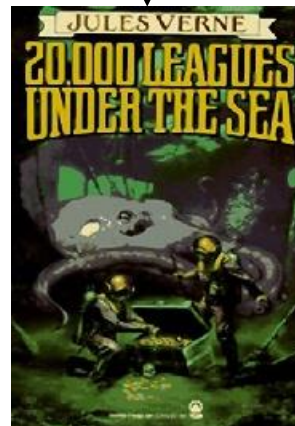
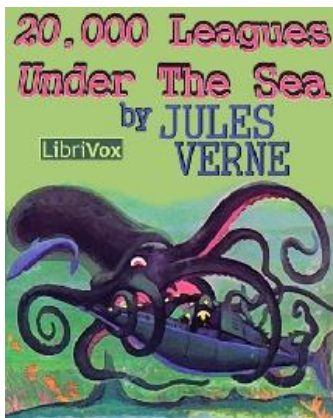


海底两万里

Vingt mille lieues sous les mers  
(in French)

20000 Leagues Under the Sea

(different English translation versions)



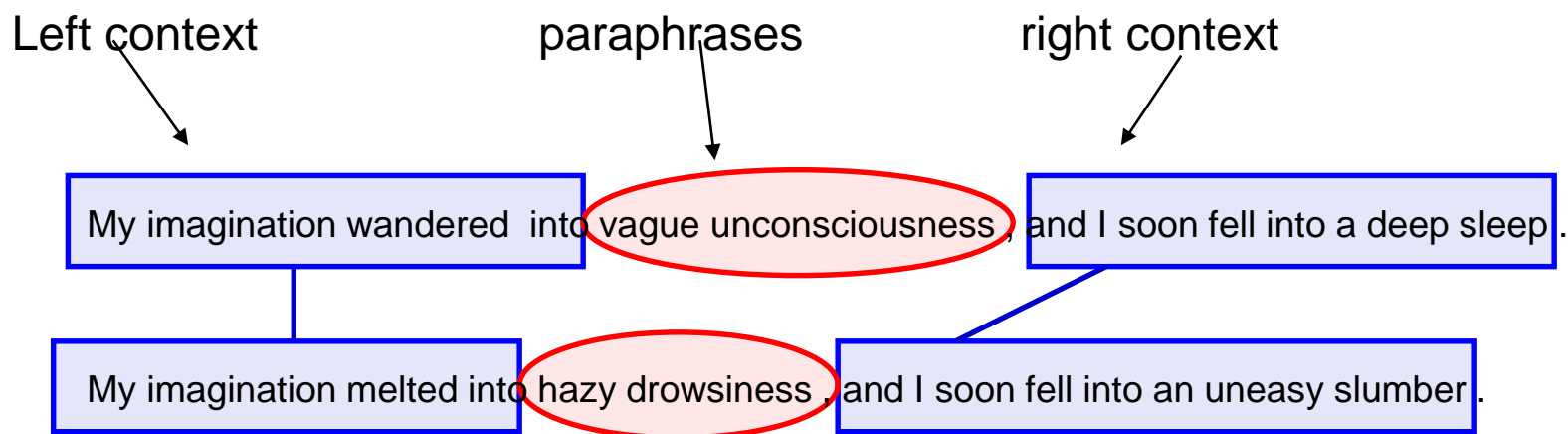
.....

# 从单语言平行语料中进行复述抽取 (一)

- Barzilay and McKeown, 2001
  - 句子对齐
    - 收集了5本外文小说的11中英文翻译本
      - E.g., *Madame Bovary*, *Fairy Tale*, *Twenty Thousand Leagues under the sea...*
    - 句子对齐
      - 得到44,562对平行句子
      - 准确率为94.5%
    - 其他处理
      - 词性标注、短语识别

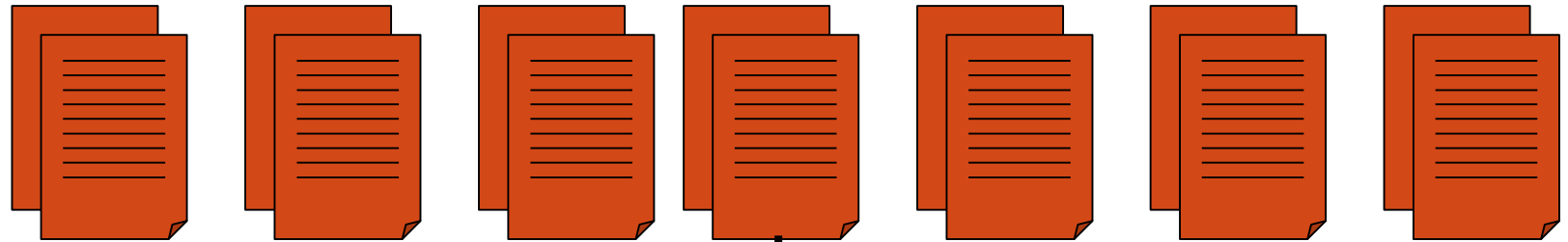
# 从单语言平行语料中进行复述抽取 (一)

- Barzilay and McKeown, 2001
  - 复述短语抽取
    - 假设：在对齐句子中出现在相似上下文中的短语为复述
    - 方法：迭代协同学习上下文与复述短语



# 从单语言平行语料中进行复述抽取 (二)

- 语料
  - 在短时间内报道同一事件的新闻文章
    - 不同新闻社撰写
- 假设
  - 可比新闻文章可能包含同一事实的不同描述



Comparable documents


Home > SPORT > TENNIS > FRENCH OPEN

## French Open 2010: Justine Henin defeats Maria Sharapova

Four-time champion Justine Henin beat Maria Sharapova of Russia 6-2, 3-6, 6-3 in delayed French Open third-round match.

Published: 12:34PM BST 30 May 2010

« Previous 1 of 2 Images Next »



Respect, French Open 2010: Justine Henin (right) acknowledges Maria Sharapova after defeating the Russian at the French Open. Photo: GETTY IMAGES

With the match between two former world number ones held over at a set-all when Sharapova held a 1-0 lead on Saturday, Sharapova was without a set of all the

Share | Digg submit | Email | Text Size +

French Open  
Sport  
Tennis  
Maria Sharapova

Ads by Google

French Open Tennis  
Telegraph  
Tennis Match U

Home / Topics / Story

UPDATED 1 DAY AGO

## 2010 French Open: Justine Henin tops Maria Sharapova in 3 sets

PARIS — One winner take-all set seemed like a final, and Justine Henin emerged the winner. Back on center court Sunday following an overnight suspension of play, Henin outslugged Maria Sharapova in a third-round showdown at the French Open, 6-2, 3-6, 6-3.

FULL ARTICLE AT ESPN

### Related Articles

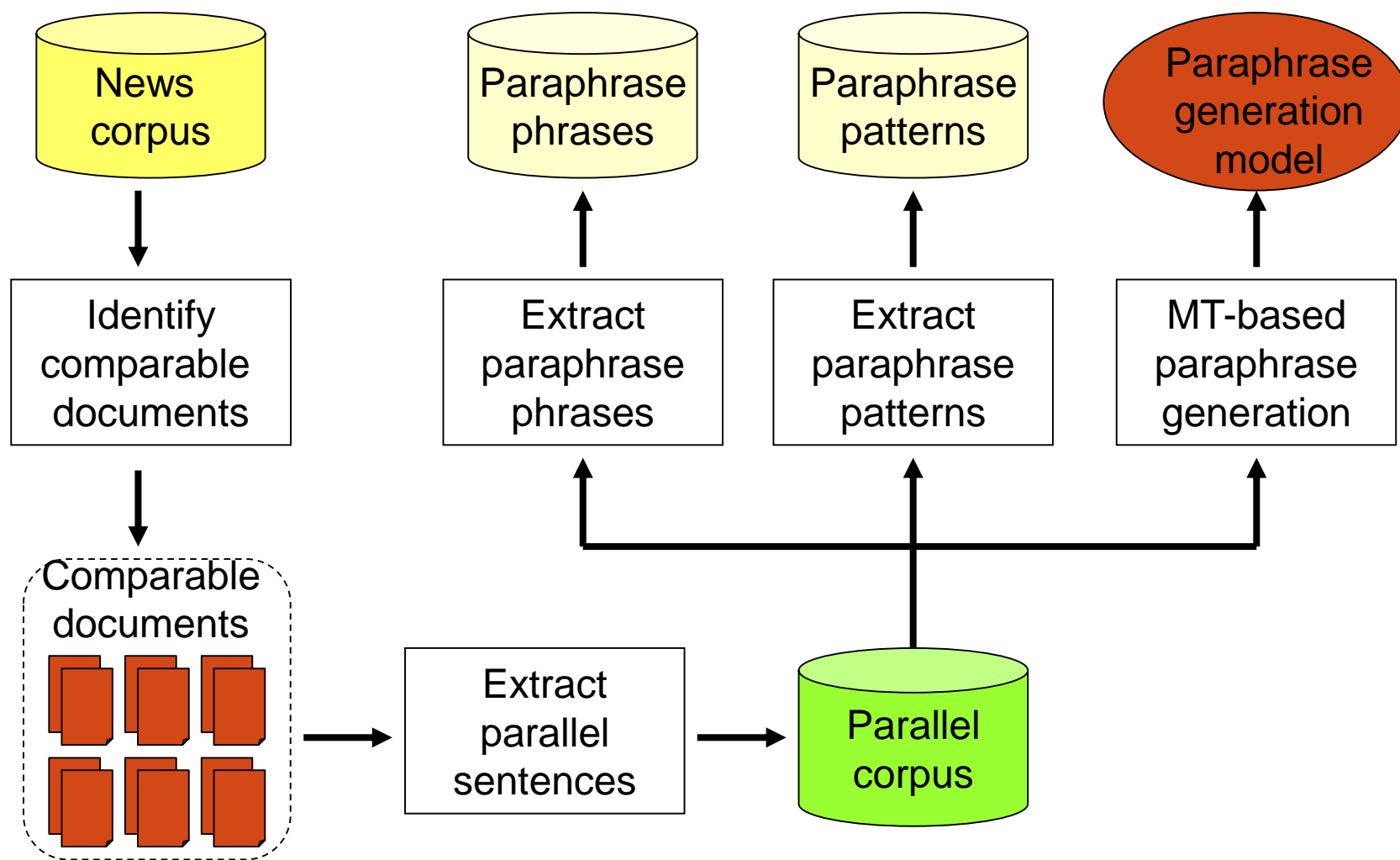
- ☐ Super Stosur proves step too far for shattered Henin 3 HOURS AGO  
Justine Henin revealed she was emotionally exhausted after slipping to her first French Open defeat for six years yesterday. Four-time champion Henin surrendered a one-set lead to lose 2-6 6-1 6-4 in round four to Samantha Stosur on a stunned Suzanne...
- ☐ Match too far for weary Henin as Stosur earns shock win 3 HOURS AGO

d1

d2

# 从单语言平行语料中进行复述抽取

## (二)





# 从单语言平行语料中进行复述抽取 (二)

- 识别可比文档
  - **方法1: 检索给定话题或事件的文档**
    - 需要预定义的话题或事件
  - **方法2: 对文档聚类**
    - 内容相似度、时间因素

# 从单语言平行语料中进行复述抽取 (二)

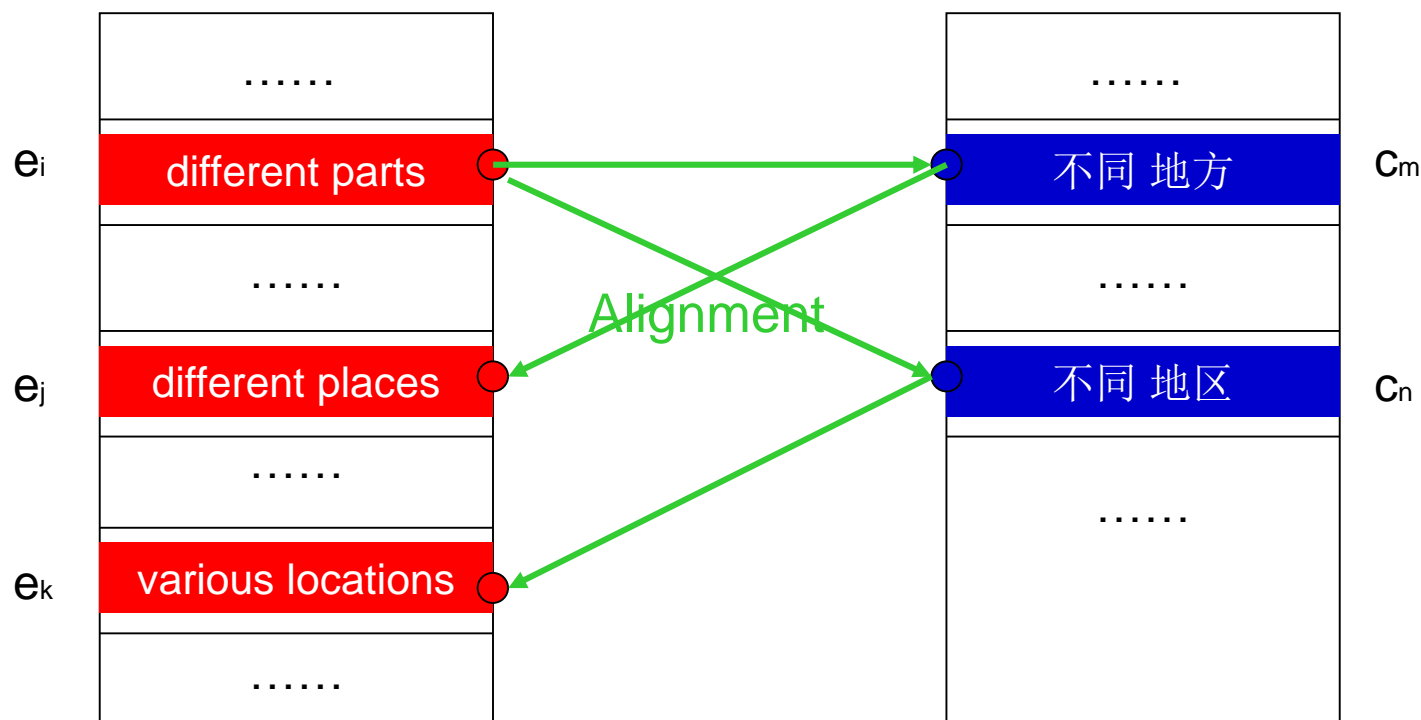
- 抽取平行/复述句子
  - 句子聚类
    - 方法1: 新闻文档的起始句子通常是对内容的摘要, 抽取所有起始句子
    - 方法2: 基于内容相似度

# 从双语平行语料中进行复述抽取

- 语料
  - 源语言和外语的平行语料
- 假设
  - 与同一外文短语对齐的多个短语可能有同样的意义
- “*pivot approach*”

source language

foreign language  
(pivot language)



# 从双语平行语料中进行复述抽取

- Bannard and Callison-Burch, 2005
  - 词对齐、短语抽取
  - 基本假设:
    - 如果两个英文短语  $e_1$  与  $e_2$  与同样的外文短语  $f$  对齐,  $e_1$  与  $e_2$  可能为复述.
  - 复述概率计算:

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2 \neq e_1} p(e_2 | e_1) \\ &= \arg \max_{e_2 \neq e_1} \underbrace{\sum_f p(f | e_1)}_{\text{Translation probability}} \underbrace{p(e_2 | f)}_{\text{Pivot in a foreign language}}\end{aligned}$$

## Bannard & Callison-Burch (2005) 's results:

...should **take the matter into consideration**...

...应当**考虑**这种情况...

...must **take the matter into account**...

...必须**考虑**这种情况...

**The consideration of this matter** will...

**考虑**这种情况会...

He'll **take the matter into consideration**

他将**考虑**这一问题

We need to **consider this matter**

大家需要**考虑**这一问题

take the matter into consideration

take the matter into account

take the matter into consideration

the consideration of this matter

take the matter into account

the consideration of this matter

take the matter into consideration

consider this matter

# 从双语平行语料中进行复述抽取

- 添加句法约束[Callison-Burch, 2008]
  - 思想：两个复述短语应该具有相同的句法类型
  - 复述概率计算：

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} p(e_2 | e_1, s(e_1)) \\ &= \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} \sum_f p(f | e_1, s(e_1)) p(e_2 | f, s(e_1))\end{aligned}$$

given the syntactic type

- 在句子中进行复述短语替换时也要考虑句法约束

## Callison-Burch (2008) 's results:

...should **take the matter into consideration**...

...应当**考虑**这种情况...

...must **take the matter into account**...

...必须**考虑**这种情况...

**The consideration of this matter** will...

**考虑**这种情况会...

He'll **take the matter into consideration**

他将**考虑**这一问题

We need to **consider this matter**

大家需要**考虑**这一问题

take the matter into consideration

take the matter into account

~~take the matter into consideration~~  
~~the consideration of this matter~~

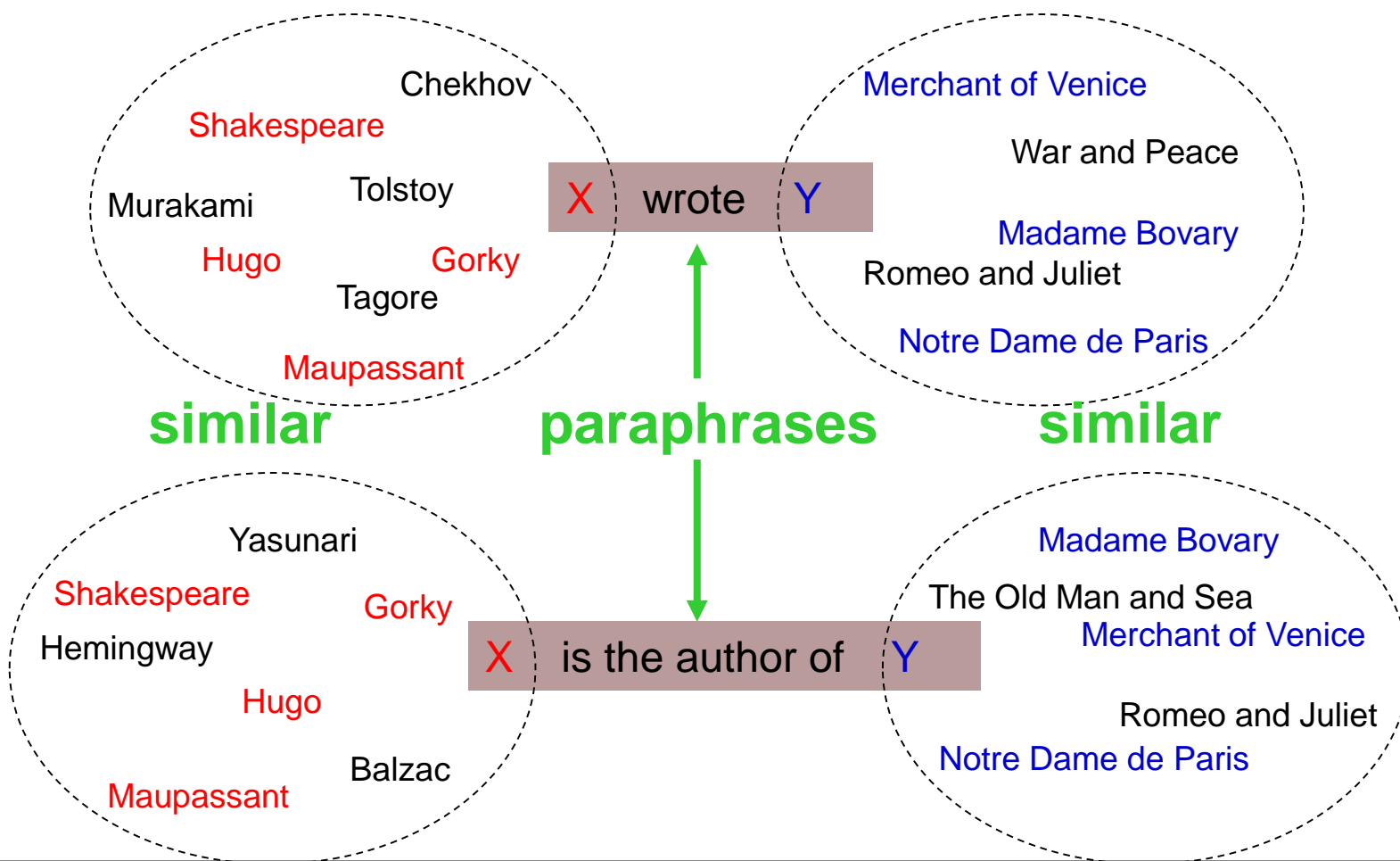
~~take the matter into account~~  
~~the consideration of this matter~~

take the matter into consideration  
consider this matter



# 基于大规模Web语料进行复述抽取

- 假设: Distributional hypothesis
  - 出现在相似上下文中的词语/短语/模式的意义相似



# 基于大规模Web语料进行复述抽取

- 基于Web Mining进行复述模式抽取

[Ravichandran and Hovy, 2002]

- 能够为每种类型抽取复述模式, e.g., “*BIRTHDAY*”
- 提供人工标注种子对, e.g., “*Mozart, 1756*”
- 利用搜索引擎从Web上检索包含种子的句子
- 抽取复述模式, e.g.,
  - *born in* <ANSWER> , <NAME>
  - <NAME> *was born on* <ANSWER> ,
  - .....

# 利用搜索引擎日志进行复述抽取

- Zhao et al., 2010
- 语料
  - 搜索引擎查询日志 (包括查询、结果页面标题)
- 假设
  - 如果查询 $q$ 命中页面标题  $t$ , 那么 $q$ 与  $t$ 可能为复述
  - 如果查询 $q_1$  与 $q_2$  命中同样的页面标题 $t$ , 那么 $q_1$ 与 $q_2$  可能为复述
  - 如果查询 $q$  命中标题 $t_1$ 与 $t_2$ , 那么 $t_1$ 与 $t_2$  可能为复述
- 候选复述均经二分类进一步验证

$q_1$

关于 草原 的 诗词

$q_2$

赞美 大 草原 的 诗

$t_1$

描写 草原 的 诗句

.....

$t_2$

有关 草原 的 诗歌

.....

Paraphrases:

$\langle q_1, t_1 \rangle$

$\langle q_1, t_2 \rangle$

$\langle q_2, t_1 \rangle$

} query-title

$\langle q_1, q_2 \rangle$

→ query-query

$\langle t_1, t_2 \rangle$

→ title-title

# 复述生成

- 基于规则的方法
- 基于词典的方法
- 基于自然语言生成的方法
- 基于机器翻译的方法
- 基于Pivot的方法

# 复述生成

- 基于规则的方法
- 基于词典的方法
- 基于自然语言生成的方法
- 基于机器翻译的方法
- 基于Pivot的方法

# 基于规则的方法

- 两类方法

- 基于人工规则

- 广泛使用于早期复述生成工作

- 基于自动抽取的规则

- 从语料中抽取复述模式，然后应用复述模式进行复述生成

- 复述规则举例

- 改变副词位置

- *He booked a single room in Beijing yesterday.* =>
      - *Yesterday*, he booked a single room in Beijing.

- 将复合句切分成一组简单句

- *He booked a single room in Beijing yesterday* =>
      - *He booked a single room in Beijing.*
      - *He booked a single room yesterday.*
      - *He booked a room.*

- 基于复述模式修改句子

- *Can I have a cup of tea?* =>
      - *May I have a cup of tea?*
      - *I would like a cup of tea, please.*
      - *Give me a cup of tea.*

# 基于词典的方法

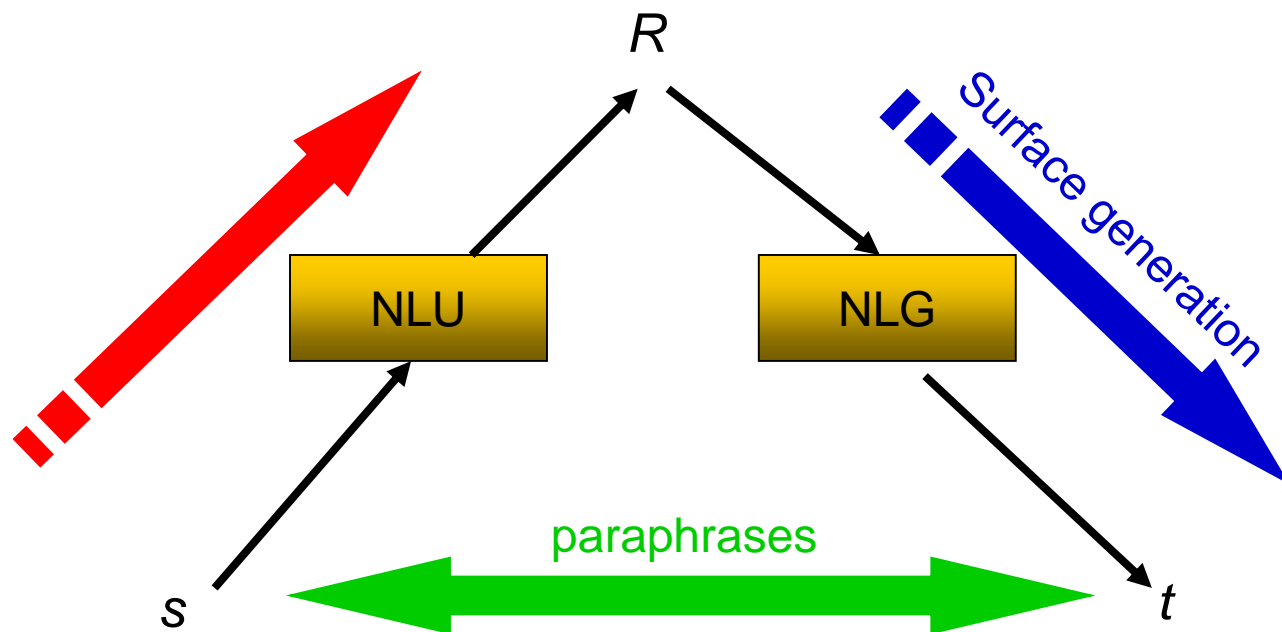
- 本质上进行词汇替换(lexical substitution)
  - 将句子中词语替换为同义词，满足给定上下文
  - SemEval-2007: English lexical substitution task
  - SemEval-2010: Cross-lingual lexical substitution
  - Example:
    - *There will be major cuts in the **salaries** of high-level civil servants.*
    - *There will be major cuts in the **wages** of high-level civil servants.*



# 基于词典的方法

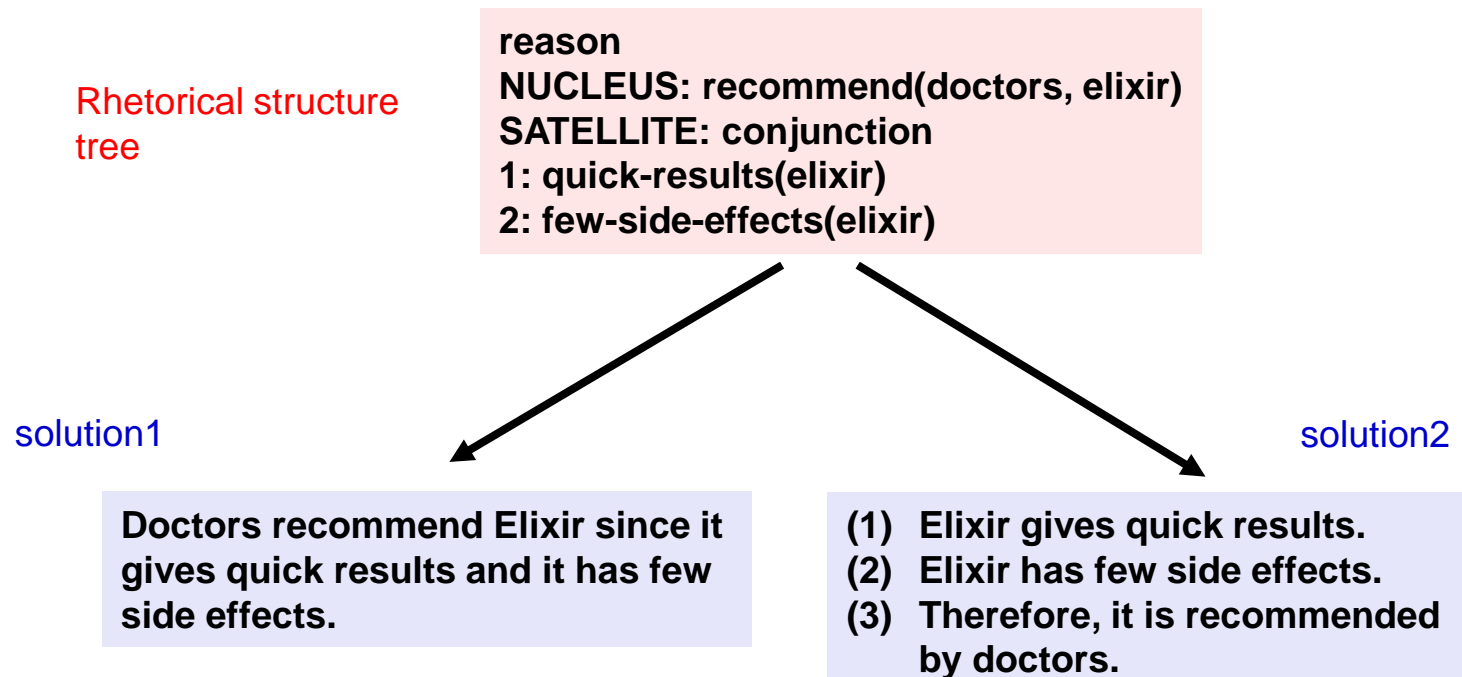
- **两个步骤**
  - **Stage-1: 从预定义词典中抽取候选替代词, E.g., WordNet**
  - **Stage-2: 确定适合给定上下文的替代词**
    - 使用语言模型 (e.g., Google 5-gram)
    - 可以进行词义消歧

# 基于自然语言生成的方法

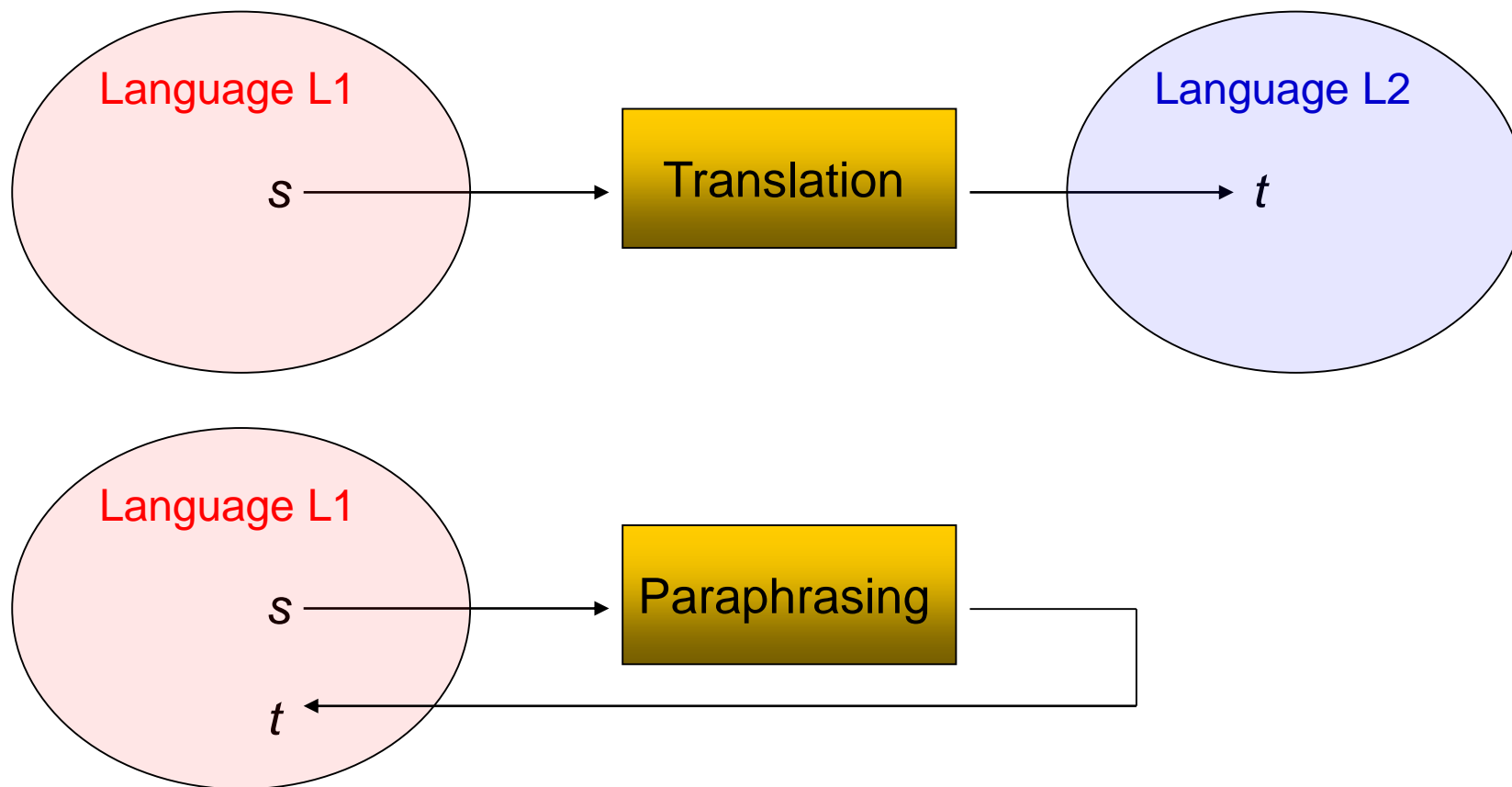


# 基于自然语言生成的方法

- 需要自然语言生成技术（本身是个难题！）
- 理论上可以生成更复杂，更自然的复述文本
- 通常输入为非自然语言语句
  - 例如Power and Scott, 2005输入为修辞结构



# 基于机器翻译的方法

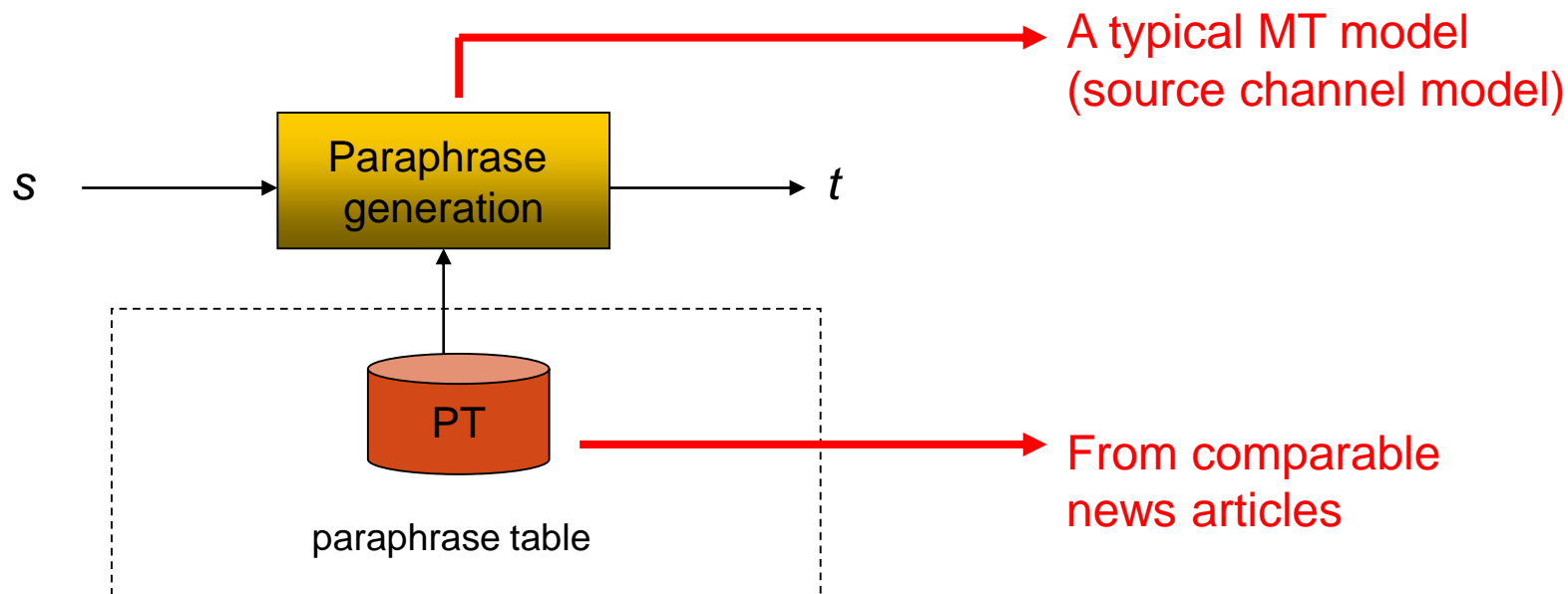


For both machine translation and paraphrase generation:

- (1)  $t$  should preserve the meaning of  $s$
- (2)  $t$  should be a fluent sentence

# 基于机器翻译的方法

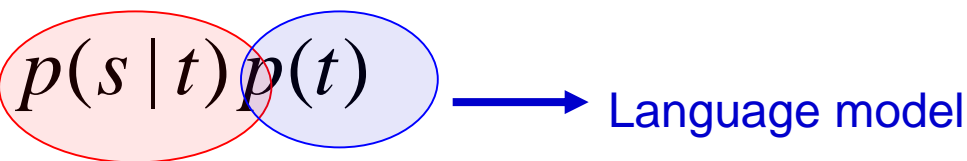
- 将复述生成看做单语言机器翻译过程【Quirk et al., 2004】



# 基于机器翻译的方法

- 将复述生成看做单语言机器翻译过程【Quirk et al., 2004】
  - 模型：Source channel model

$$t^* = \arg \max_t p(t | s)$$

$$= \arg \max_t p(s | t) p(t)$$


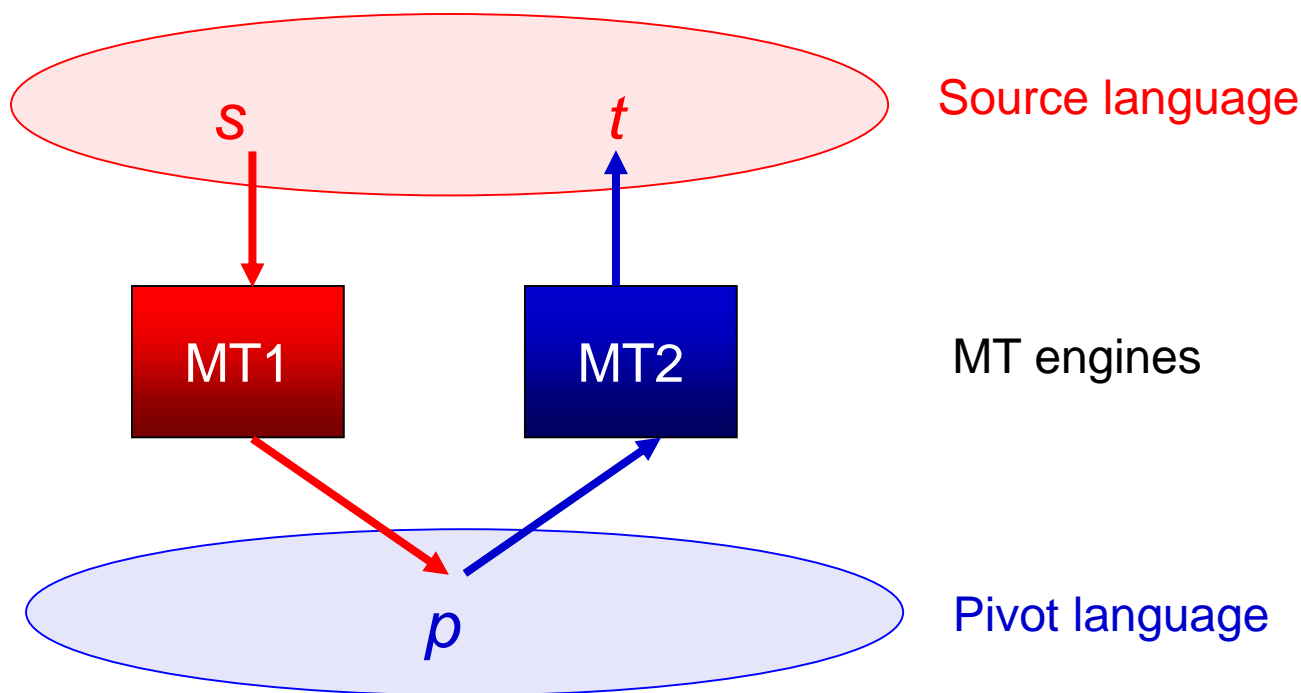
“Translation” model

(based on a phrasal paraphrase table)

- Based on Monolingual parallel sentences extracted from news articles
- Word alignment & phrase pair extraction With Giza++

# 基于Pivot的方法

- 将句子 $s$ 翻译到另一语言，然后再翻译到当前语言



# 基于Pivot的方法

- **Example:**

English      What toxins are most **hazardous** to **expectant mothers**?



Italian      Che tossine sono più pericolose alle donne incinte?



English      What toxins are more **dangerous** to **pregnant women**?

- **Single-pivot**
  - Using a single pivot language
- **Multi-pivot**
  - Using multiple pivot languages



# 深度学习时代

## Sequence-to-Sequence Models

# 文本推理阅读材料

- Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. (2006). Learning to distinguish valid textual entailments. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment.
- Fabio Zanzotto, Alessandro Moschitti. (2006). Automatic learning of textual entailments with cross-pair similarities. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.
- Milen Kouylekov and Bernardo Magnini. (2005). Tree Edit Distance for Textual Entailment. In Proceedings of the RANLP.
- Matteo Negri, Milen Kouylekov, Bernardo Magnini, Yashar Mehdad, and Elena Cabrio. (2009). Towards Extensible Textual Entailment Engines: the EDITS Package. In proceeding of the 11th Conference of the Italian Association for Artificial Intelligence.
- Fabio Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. (2009). A machine-learning approach to textual entailment recognition. In Journal of Natural Language Engineering.
- Diana Pérez and Enrique Alfonseca. (2005). Application of the Bleu algorithm for recognising textual entailments. In Proceedings of the First Challenge Workshop Recognising Textual Entailment.
- Houping Jia, Xiaojiang Huang, Tengfei Ma, Xiaojun Wan, Jianguo Xiao. (2010). PKUTM Participation at TAC 2010 RTE and Summarization Track. In Proceedings of the 2010 Text Analysis Conference (TAC 2010).
- Wang, Xiao-Lin, Hai Zhao, and Bao-Liang Lu. "BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RTE-2 Task." *Proceeding of the 10th NTCIR Conference, Tokyo, Japan*. 2013.

# 文本复述阅读材料

- Brockett and Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction.
- Finch et al. 2005. Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence.
- Wu. 2005. Recognizing Paraphrases and Textual Entailment using Inversion Transduction Grammars.
- Malakasiotis. 2009. Paraphrase Recognition Using Machine Learning to Combine Similarity Measures.
- Das and Smith. 2009. Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition.
- Zhao et al. 2010. Paraphrasing with Search Engine Query Logs.
- Barzilay and McKeown. 2001. Extracting Paraphrases from a Parallel Corpus.
- Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo. 2002. Automatic Paraphrase Acquisition from News Articles.
- Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.
- Takao et al. 2002. Comparing and Extracting Paraphrasing Words with 2-Way Bilingual Dictionaries.
- Bannard and Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora.
- Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora.
- Kok and Brockett. 2010. Hitting the Right Paraphrases in Good Time.
- Zhao et al. 2008. Pivot Approach for Extracting Paraphrase Patterns from bilingual corpora.
- Zhao et al. 2010. Leveraging Multiple MT Engines for Paraphrase Generation.
- Quirk et al. 2004. Monolingual Machine Translation for Paraphrase Generation.
- Power and Scott. 2005. Automatic generation of large-scale paraphrases.
- McCarthy and Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task.
- Ravichandran and Hovy. 2002. Learning Surface Text Patterns for a Question Answering System.

# 文本复述阅读材料

- Brockett and Dolan. 2005. Support Vector Machines for Paraphrase Identification and Corpus Construction.
- Finch et al. 2005. Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence.
- Wu. 2005. Recognizing Paraphrases and Textual Entailment using Inversion Transduction Grammars.
- Malakasiotis. 2009. Paraphrase Recognition Using Machine Learning to Combine Similarity Measures.
- Das and Smith. 2009. Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition.
- Zhao et al. 2010. Paraphrasing with Search Engine Query Logs.
- Barzilay and McKeown. 2001. Extracting Paraphrases from a Parallel Corpus.
- Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo. 2002. Automatic Paraphrase Acquisition from News Articles.
- Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.
- Takao et al. 2002. Comparing and Extracting Paraphrasing Words with 2-Way Bilingual Dictionaries.
- Bannard and Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora.
- Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora.
- Kok and Brockett. 2010. Hitting the Right Paraphrases in Good Time.
- Zhao et al. 2008. Pivot Approach for Extracting Paraphrase Patterns from bilingual corpora.
- Zhao et al. 2010. Leveraging Multiple MT Engines for Paraphrase Generation.
- Quirk et al. 2004. Monolingual Machine Translation for Paraphrase Generation.
- Power and Scott. 2005. Automatic generation of large-scale paraphrases.
- McCarthy and Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task.
- Ravichandran and Hovy. 2002. Learning Surface Text Patterns for a Question Answering System.
- Socher, Richard, et al. "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection." *Advances in Neural Information Processing Systems*. 2011.
- Wenpeng Yin and Hinrich Schutze. Convolutional Neural Network for Paraphrase Identification. NAACL 2015.

- **Some slides were taken or adapted from related slides written by Ido Dagan, Dan Roth, Fabio Massimo Zanzotto, Houping Jia, Shiqi Zhao, Haifeng Wang, etc. Thank them for sharing their slides.**

