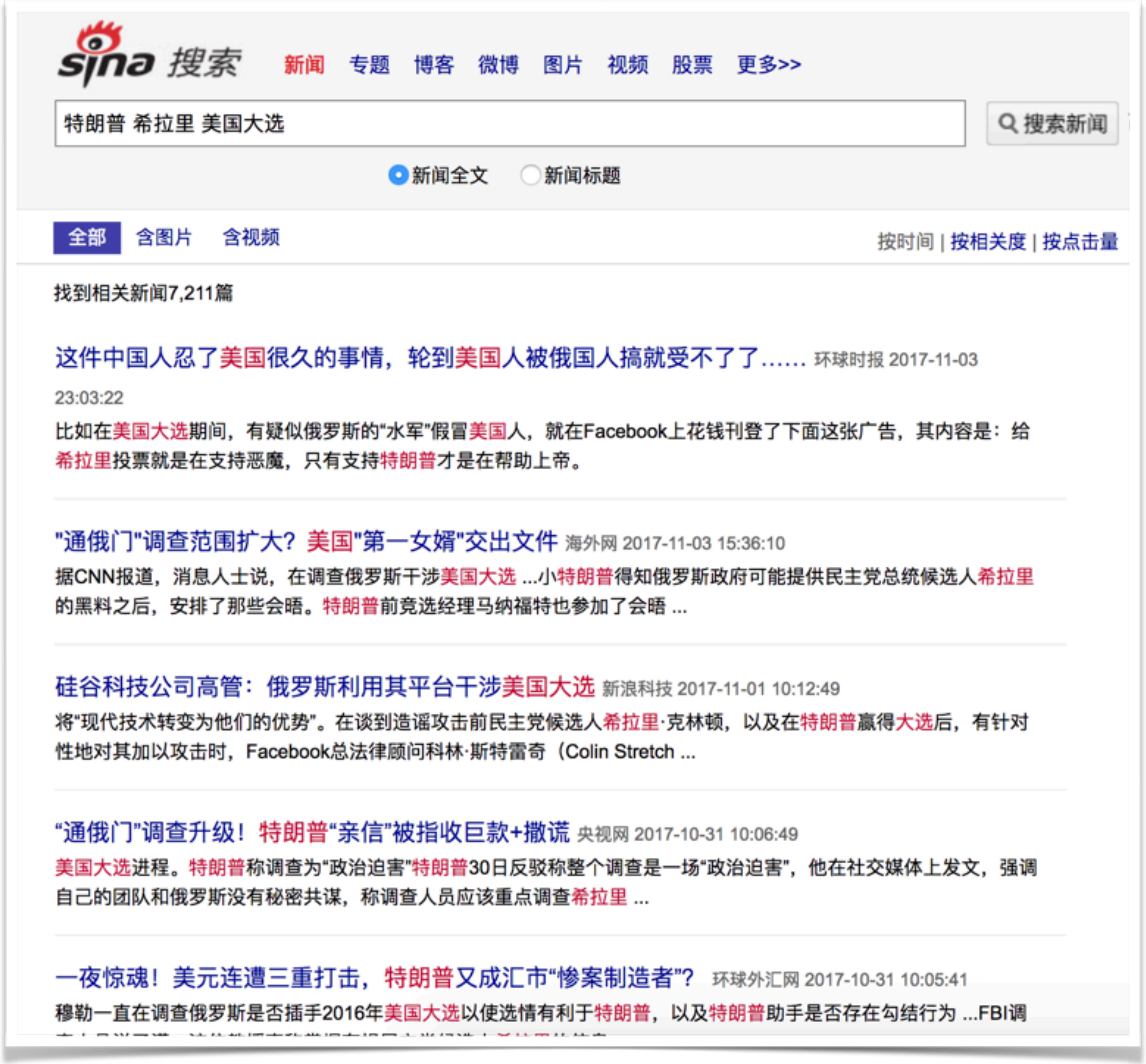
Growing Story Forest Online from Massive Breaking News

Presentation by **Di Niu**
University of Alberta
Computer Lab Paris 6

Information Explosion

Reading vs. Browsing

News Reading: Search Engine



sina 搜索
新闻 专题 博客 微博 图片 视频 股票 更多>>
搜索新闻
全部 含图片 含视频
按时间 | 按相关度 | 按点击量

特朗普 希拉里 美国大选

● 新闻全文 ○ 新闻标题

找到相关新闻7,211篇

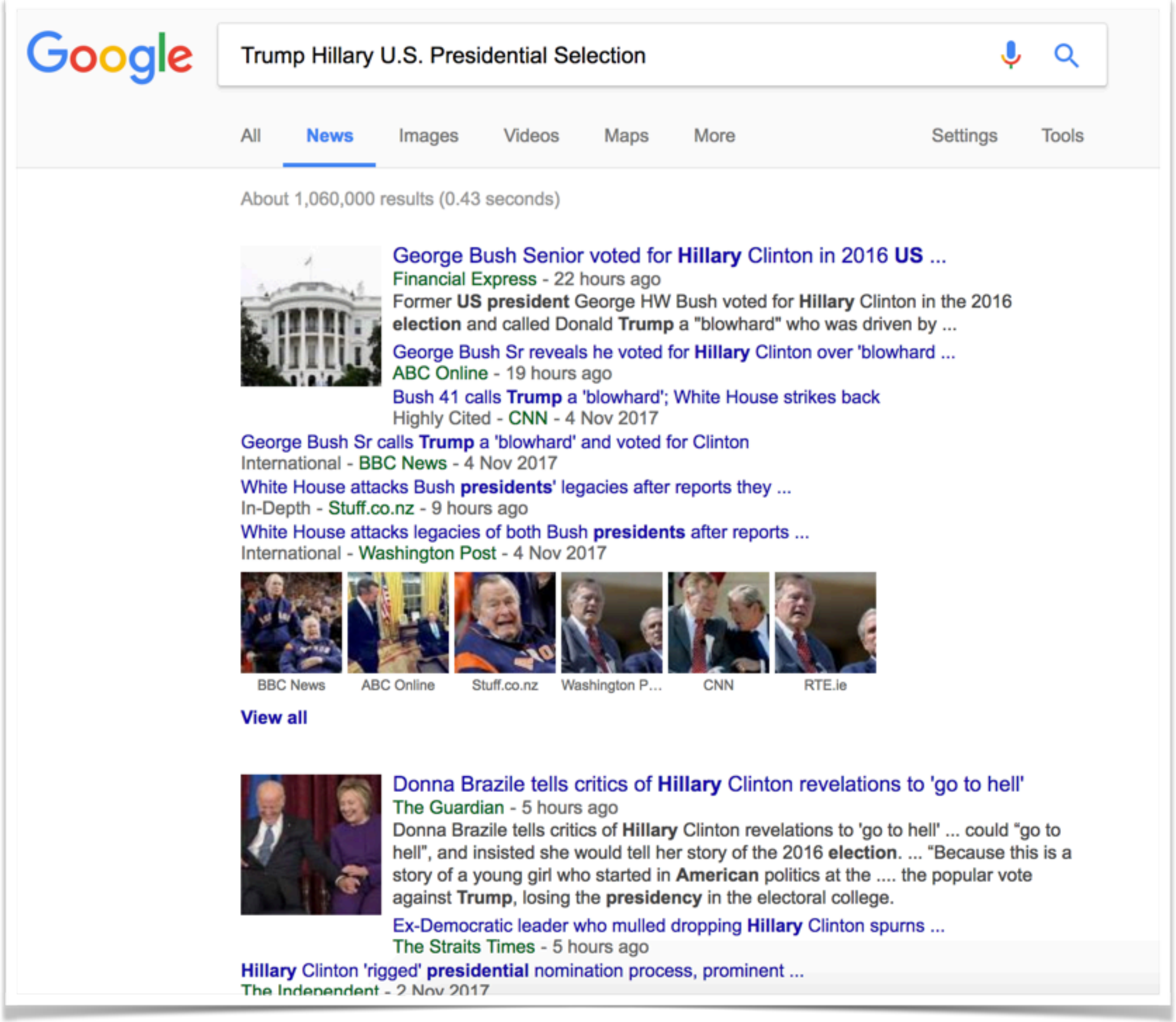
这件中国人忍了美国很久的事情，轮到美国人被俄国人搞就受不了了…… 环球时报 2017-11-03 23:03:22
比如在美国大选期间，有疑似俄罗斯的"水军"假冒美国人，就在Facebook上花钱刊登了下面这张广告，其内容是：给希拉里投票就是在支持恶魔，只有支持特朗普才是在帮助上帝。

"通俄门"调查范围扩大? 美国"第一女婿"交出文件 海外网 2017-11-03 15:36:10
据CNN报道，消息人士说，在调查俄罗斯干涉美国大选 ...小特朗普得知俄罗斯政府可能提供民主党总统候选人希拉里的黑料之后，安排了那些会晤。特朗普前竞选经理马纳福特也参加了会晤 ...

硅谷科技公司高管：俄罗斯利用其平台干涉美国大选 新浪科技 2017-11-01 10:12:49
将"现代技术转变为他们的优势"。在谈到造谣攻击前民主党候选人希拉里·克林顿，以及在特朗普赢得大选后，有针对性地对其加以攻击时，Facebook总法律顾问科林·斯特雷奇（Colin Stretch ...

"通俄门"调查升级！特朗普"亲信"被指收巨款+撒谎 央视网 2017-10-31 10:06:49
美国大选进程。特朗普称调查为"政治迫害"特朗普30日反驳称整个调查是一场"政治迫害"，他在社交媒体上发文，强调自己的团队和俄罗斯没有秘密共谋，称调查人员应该重点调查希拉里 ...

一夜惊魂！美元连遭三重打击，特朗普又成汇市"惨案制造者"? 环球外汇网 2017-10-31 10:05:41
穆勒一直在调查俄罗斯是否插手2016年美国大选以使选情有利于特朗普，以及特朗普助手是否存在勾结行为 ...FBI调



Trump Hillary U.S. Presidential Selection

About 1,060,000 results (0.43 seconds)

George Bush Senior voted for **Hillary** Clinton in 2016 **US** ...
Financial Express - 22 hours ago
Former **US president** George HW Bush voted for **Hillary** Clinton in the 2016 **election** and called Donald **Trump** a "blowhard" who was driven by ...

George Bush Sr reveals he voted for **Hillary** Clinton over 'blowhard ...
ABC Online - 19 hours ago

Bush 41 calls **Trump** a 'blowhard'; White House strikes back
Highly Cited - CNN - 4 Nov 2017

George Bush Sr calls **Trump** a 'blowhard' and voted for Clinton
International - BBC News - 4 Nov 2017

White House attacks Bush **presidents**' legacies after reports they ...
In-Depth - Stuff.co.nz - 9 hours ago

White House attacks legacies of both Bush **presidents** after reports ...
International - Washington Post - 4 Nov 2017

View all

Donna Brazile tells critics of **Hillary** Clinton revelations to 'go to hell'
The Guardian - 5 hours ago
Donna Brazile tells critics of **Hillary** Clinton revelations to 'go to hell' ... could "go to hell", and insisted she would tell her story of the 2016 **election**. ... "Because this is a story of a young girl who started in **American** politics at the .... the popular vote against **Trump**, losing the **presidency** in the electoral college.

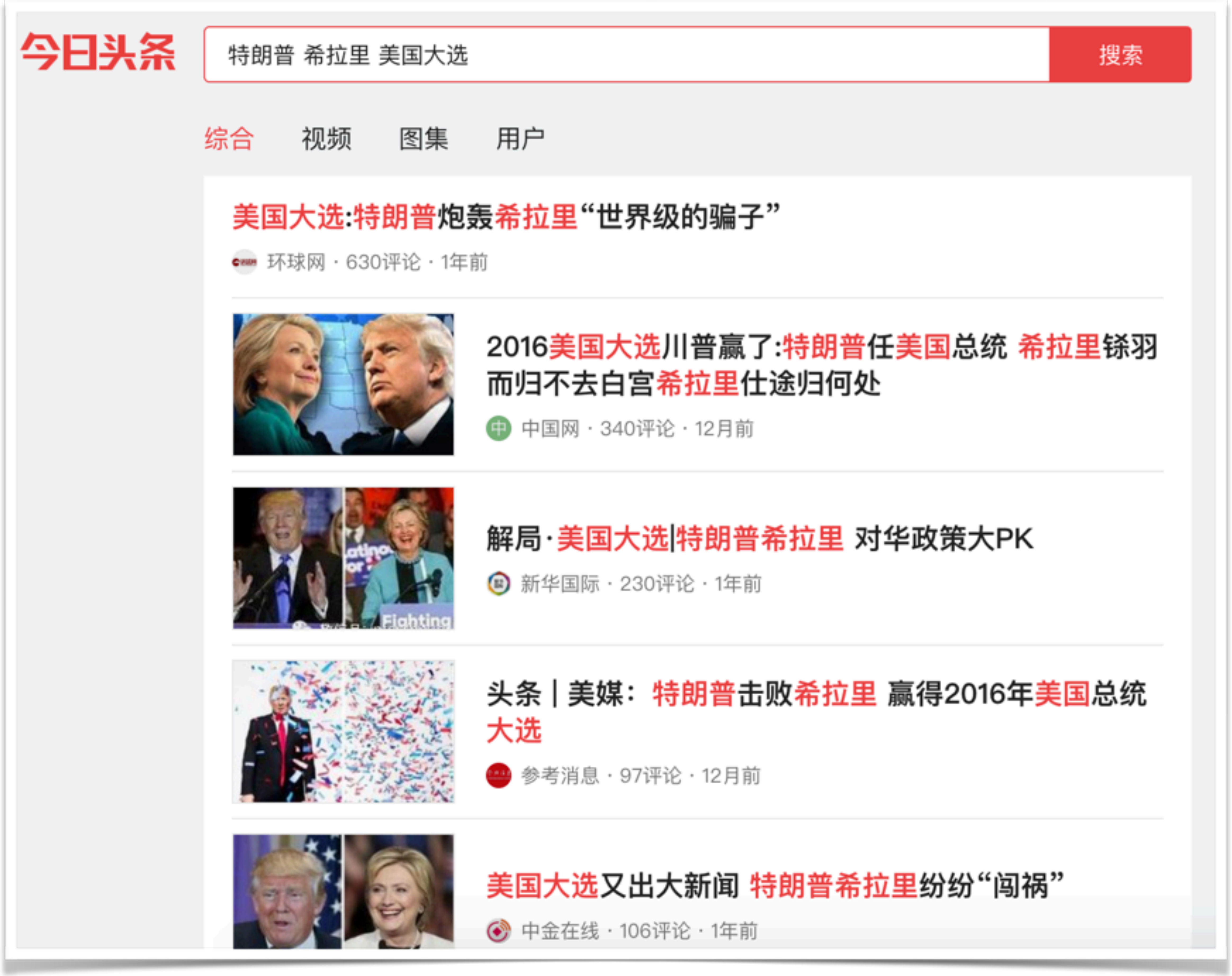
Ex-Democratic leader who mulled dropping **Hillary** Clinton spurns ...
The Straits Times - 5 hours ago

Hillary Clinton 'rigged' **presidential** nomination process, prominent ...
The Independent - 2 Nov 2017

News Reading: Feed Stream

Disadvantages of existing systems

- Messed document lists
- Extremely fine-grained (articles)
- Redundant useless information
- Unstructured information



特朗普 希拉里 美国大选

美国大选:特朗普炮轰希拉里“世界级的骗子”
环球网 · 630评论 · 1年前

2016美国大选川普赢了:特朗普任美国总统 希拉里铩羽而归不去白宫希拉里仕途归何处
中国网 · 340评论 · 12月前

解局·美国大选|特朗普希拉里 对华政策大PK
新华国际 · 230评论 · 1年前

头条 | 美媒：特朗普击败希拉里 赢得2016年美国总统大选
参考消息 · 97评论 · 12月前

美国大选又出大新闻 特朗普希拉里纷纷“闯祸”
中金在线 · 106评论 · 1年前

How We Remember Information

Event: something revolve around one or a group of specific persons (or entities) and happen at certain place during specific time .
Examples: Trump becomes a candidate, The first game between Kejie and AlphaGo

Story: multiple events that interdependent and evolve by time form a story.
Examples: 2016 U.S. Presidential Election, Kejie VS AlphaGo

How We Remember Information

Event: something revolve around one or a group of specific persons (or entities) and happen at certain place during specific time .
Examples: Trump becomes a candidate, The first game between Kejie and AlphaGo

Story: multiple events that interdependent and evolve by time form a story.
Examples: 2016 U.S. Presidential Election, Kejie VS AlphaGo

The smallest granularity of memory: event

Why Event Matters

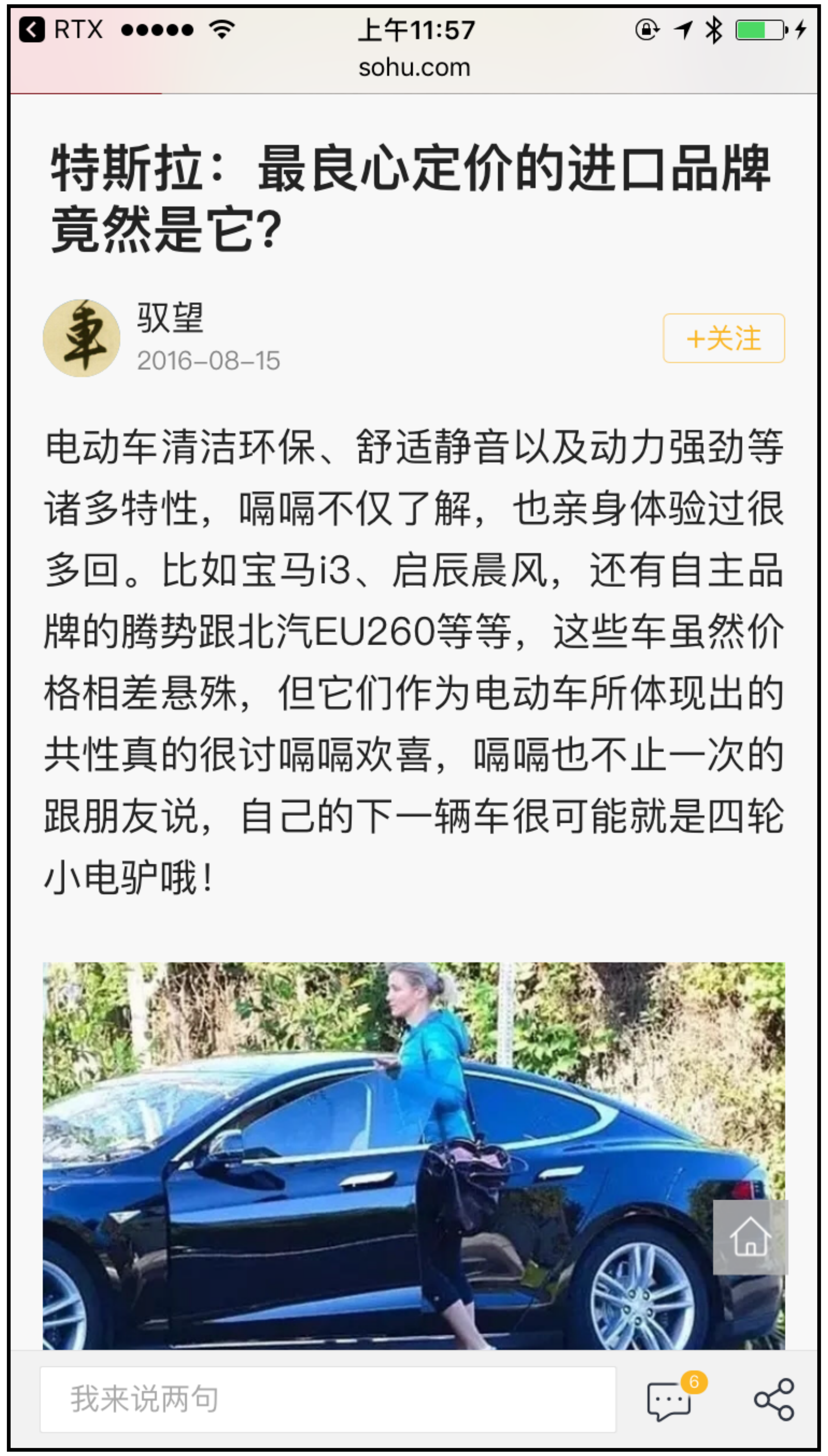
Tags we have

# Category tags
# Automotive Technology

# Entity tags
# Tesla

Tags we don't have

# Event tags
# Tesla launches new model X



特斯拉：最良心定价的进口品牌竟然是它?

驭望
2016–08–15

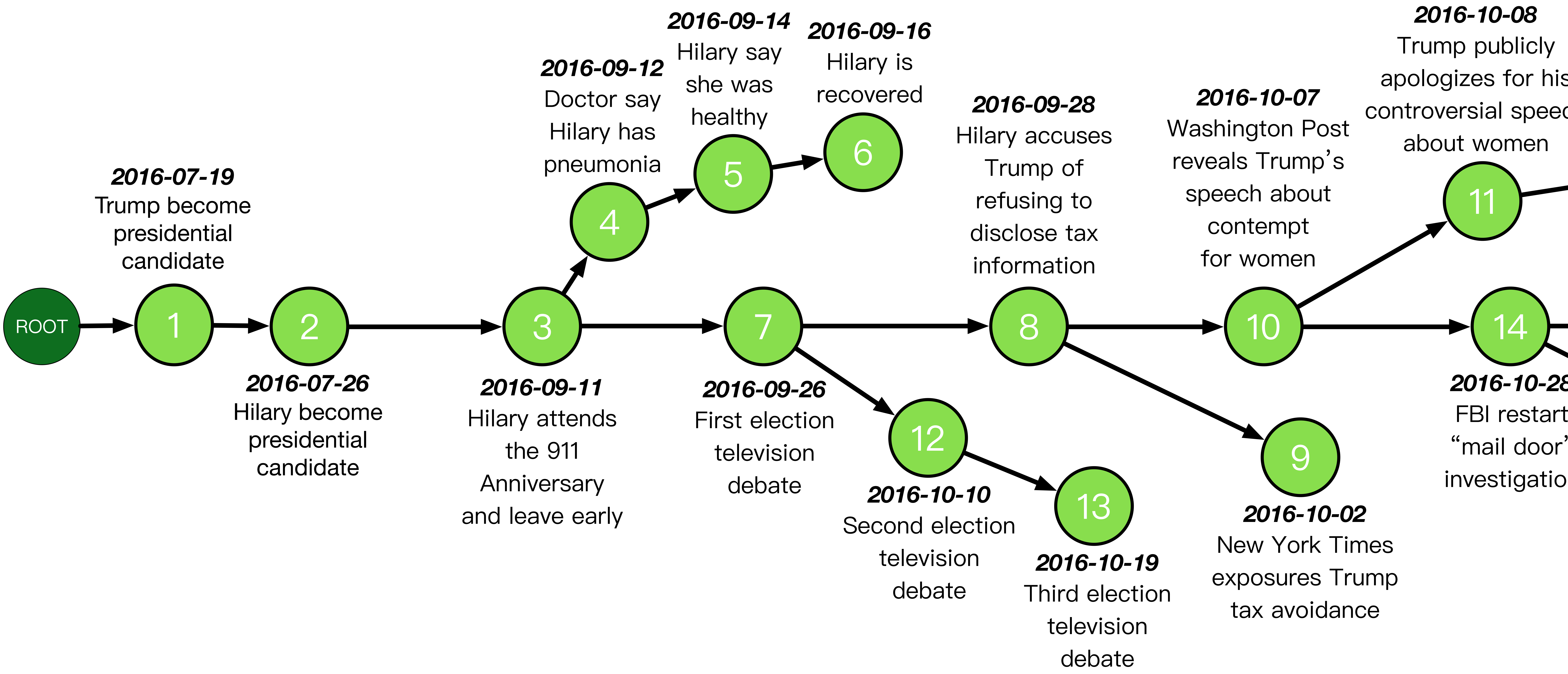
电动车清洁环保、舒适静音以及动力强劲等诸多特性，嗝嗝不仅了解，也亲身体验过很多回。比如宝马i3、启辰晨风，还有自主品牌的腾势跟北汽EU260等等，这些车虽然价格相差悬殊，但它们作为电动车所体现出的共性真的很讨嗝嗝欢喜，嗝嗝也不止一次的跟朋友说，自己的下一辆车很可能就是四轮小电驴哦！

Title translation

Tesla: The most conscientious pricing of imported brands turned out to be it?

7.5% articles with event tags account for **40%** of the user traffic

How Human Brain Organizes Information



2016-07-19
Trump become presidential candidate

2016-07-26
Hilary become presidential candidate

2016-09-11
Hilary attends the 911 Anniversary and leave early

2016-09-12
Doctor say Hilary has pneumonia

2016-09-14
Hilary say she was healthy

2016-09-16
Hilary is recovered

2016-09-26
First election television debate

2016-09-28
Hilary accuses Trump of refusing to disclose tax information

2016-10-02
New York Times exposures Trump tax avoidance

2016-10-07
Washington Post reveals Trump's speech about contempt for women

2016-10-08
Trump publicly apologizes for his controversial speec about women

2016-10-10
Second election television debate

2016-10-19
Third election television debate

2016-10-2
FBI restart “mail door’ investigatio

How Human Brain Organizes Information

12
ay
as
nia
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

2016-09-14
Hilary say
she was
healthy

2016-09-16
Hilary is
recovered

2016-09-26
First election
television
debate

2016-09-28
Hilary accuses
Trump of
refusing to
disclose tax
information

2016-10-02
New York Times
exposures Trump
tax avoidance

2016-10-07
Washington Post
reveals Trump's
speech about
contempt
for women

2016-10-08
Trump publicly
apologizes for his
controversial speech
about women

2016-10-10
Second election
television
debate

2016-10-19
Third election
television
debate

2016-10-28
FBI restart
"mail door"
investigation

2016-10-29
FBI explain
for restarting
"mail door"
investigation

2016-10-30
Hilary questions
FBI's motivation
for restarting
investigation

2016-11-02
Hilary condemns
Trump for
bullying women

2016-11-06
FBI director:
No charges
after new
review of
Hilary emails

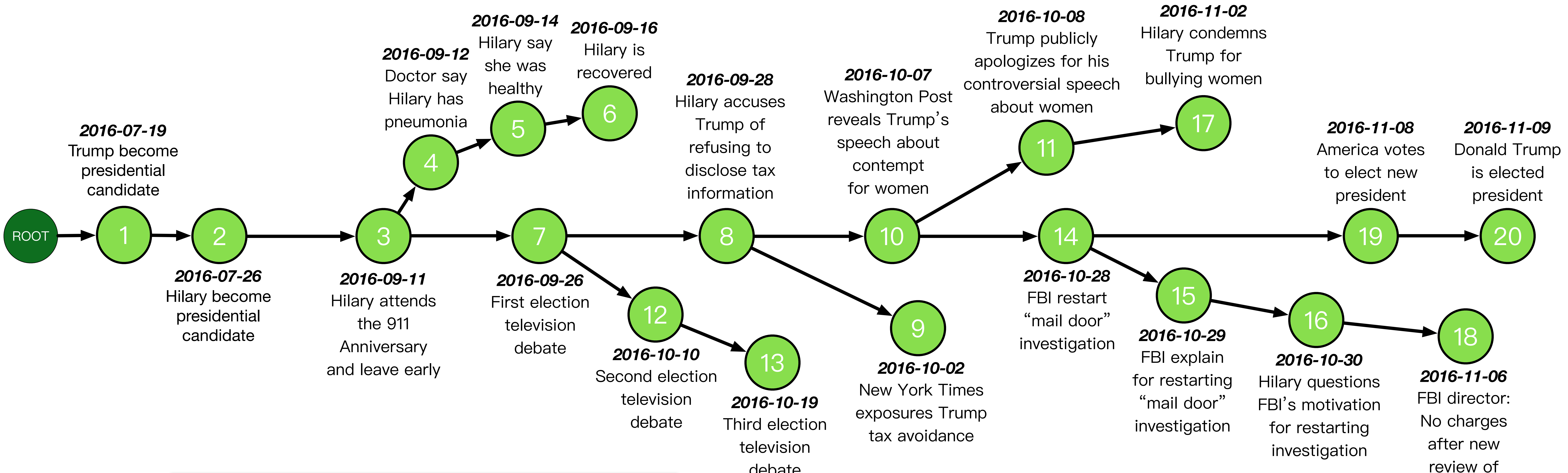
2016-11-08
America votes
to elect new
president

2016-11-09
Donald Trump
is elected
president

Reinvent
information platform
that matches human habits

Story Forest

Detect events automatically from massive news articles



2016-07-19 Trump become presidential candidate

2016-07-26 Hilary become presidential candidate

2016-09-11 Hilary attends the 911 Anniversary and leave early

2016-09-12 Doctor say Hilary has pneumonia

2016-09-14 Hilary say she was healthy

2016-09-16 Hilary is recovered

2016-09-26 First election television debate

2016-09-28 Hilary accuses Trump of refusing to disclose tax information

2016-10-02 New York Times exposures Trump tax avoidance

2016-10-07 Washington Post reveals Trump's speech about contempt for women

2016-10-08 Trump publicly apologizes for his controversial speech about women

2016-10-10 Second election television debate

2016-10-19 Third election television debate

2016-10-28 FBI restart "mail door" investigation

2016-10-29 FBI explain for restarting "mail door" investigation

2016-10-30 Hilary questions FBI's motivation for restarting investigation

2016-11-02 Hilary condemns Trump for bullying women

2016-11-06 FBI director: No charges after new review of Hilary emails

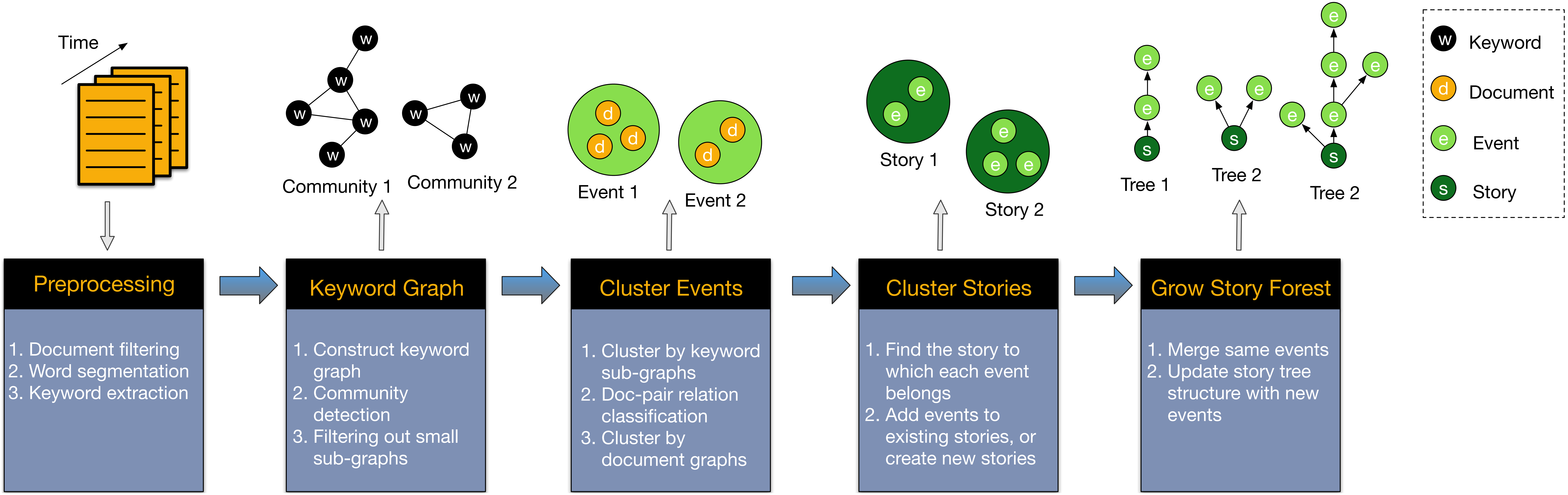
2016-11-08 America votes to elect new president

2016-11-09 Donald Trump is elected president

Trees denotes stories, nodes denotes events

Edges in the tree denotes events evolving relationship

Story Forest System Overview



Preprocessing

1. Document filtering
2. Word segmentation
3. Keyword extraction

Keyword Graph

1. Construct keyword graph
2. Community detection
3. Filtering out small sub-graphs

Cluster Events

1. Cluster by keyword sub-graphs
2. Doc-pair relation classification
3. Cluster by document graphs

Cluster Stories

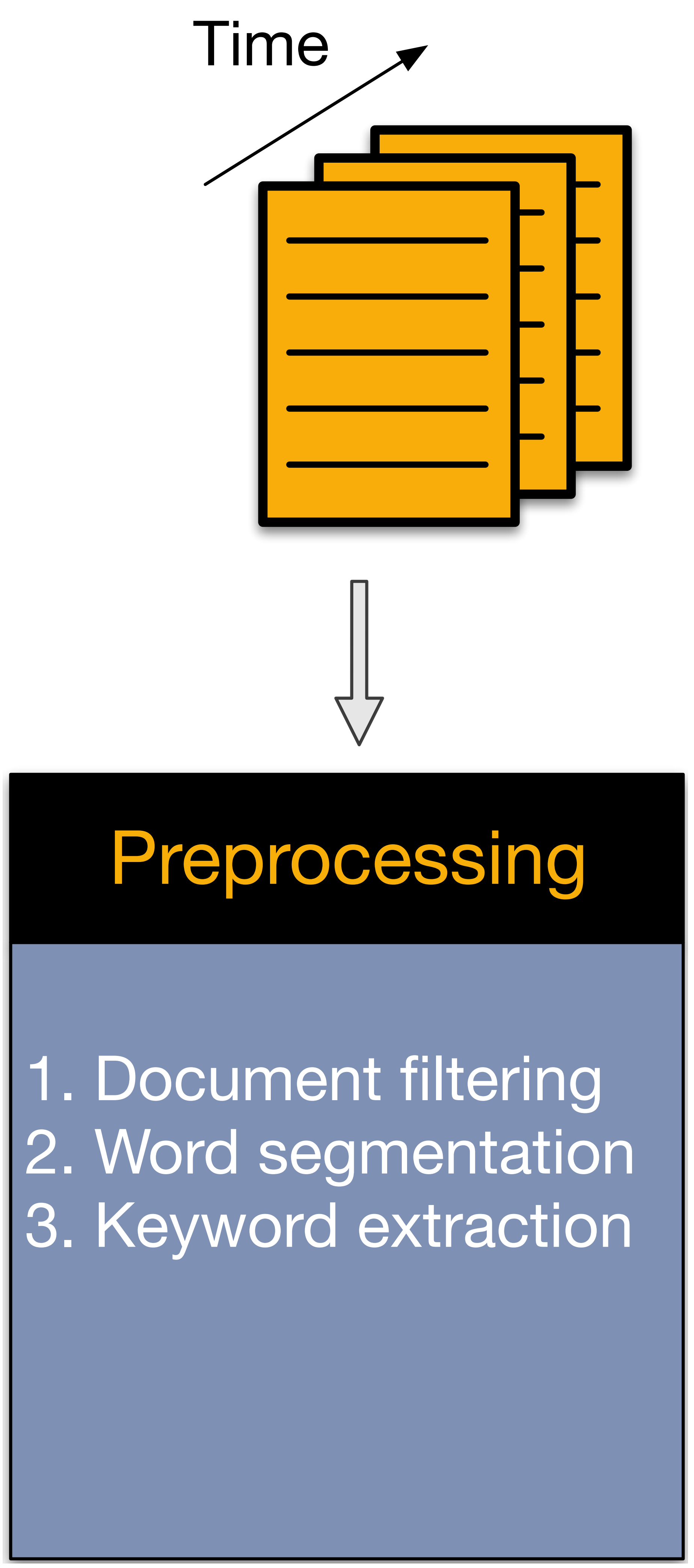
1. Find the story to which each event belongs
2. Add events to existing stories, or create new stories

Grow Story Forest

1. Merge same events
2. Update story tree structure with new events

Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, Yu Xu. “Growing Story Forest Online from Massive Breaking News,” in **CIKM 2017**.

Preprocessing



1. Document filtering
2. Word segmentation
3. Keyword extraction

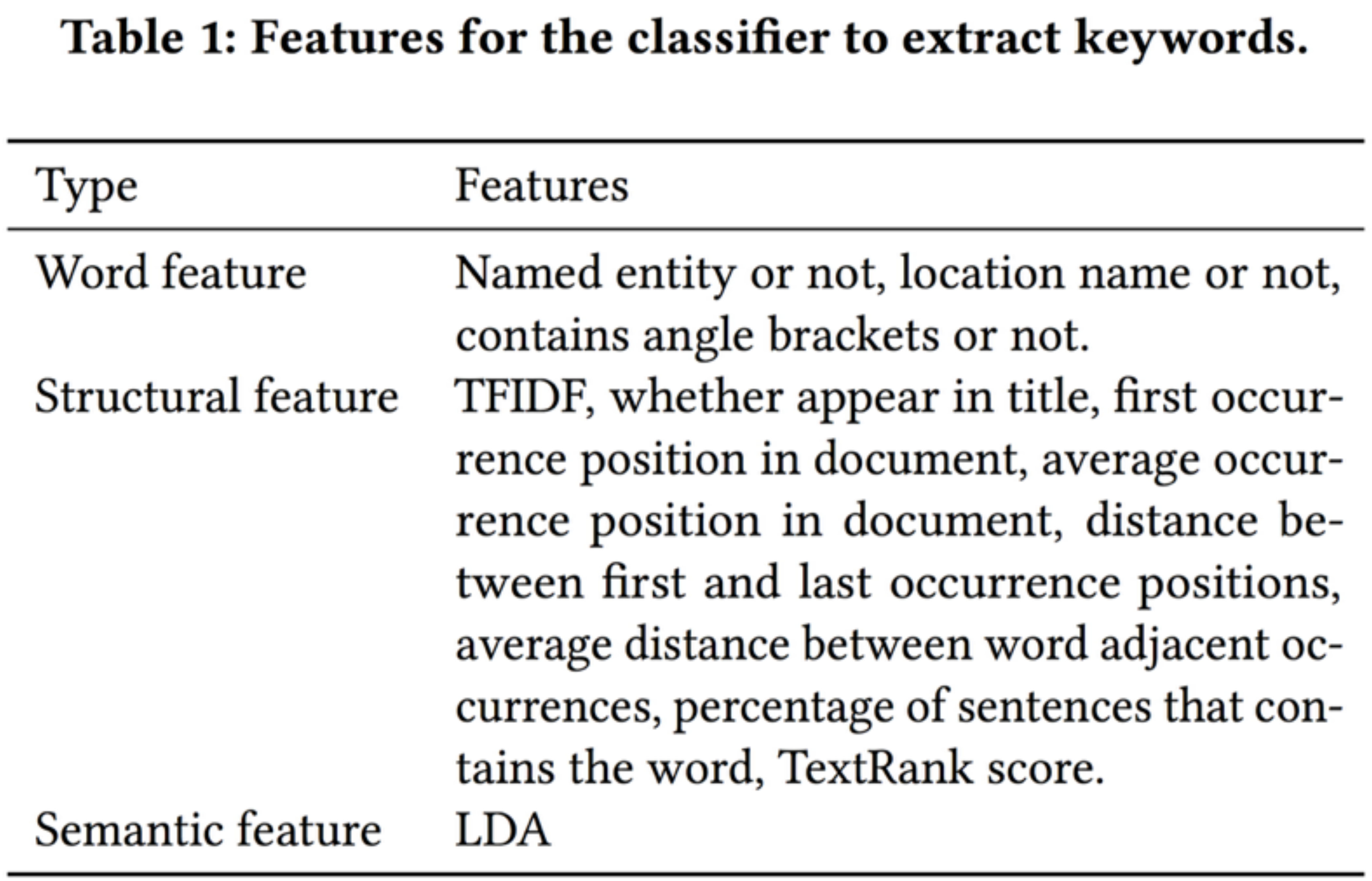
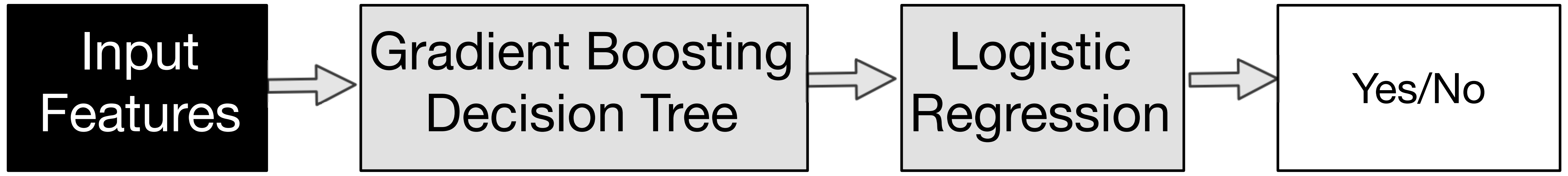
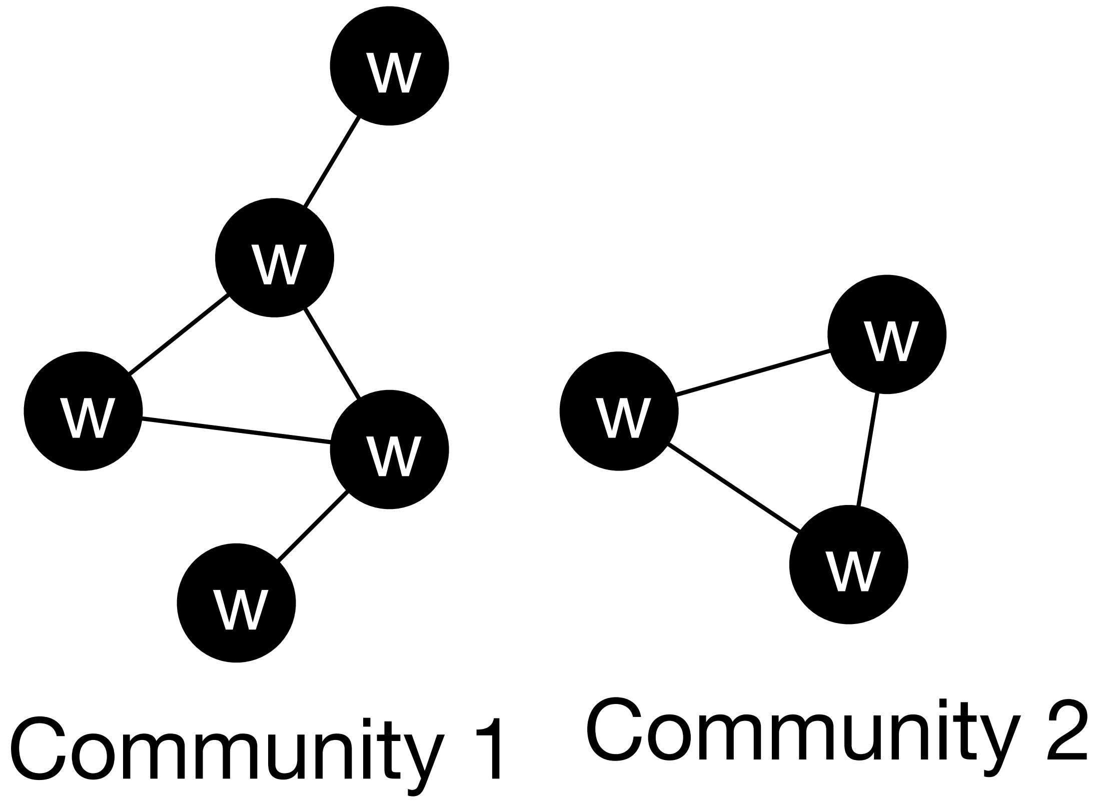


Table 1: Features for the classifier to extract keywords.

Type	Features
Word feature	Named entity or not, location name or not, contains angle brackets or not.
Structural feature	TFIDF, whether appear in title, first occurrence position in document, average occurrence position in document, distance between first and last occurrence positions, average distance between word adjacent occurrences, percentage of sentences that contains the word, TextRank score.
Semantic feature	LDA



Keyword Graph



Keyword Graph

1. Construct keyword graph
2. Community detection
3. Filtering out small sub-graphs

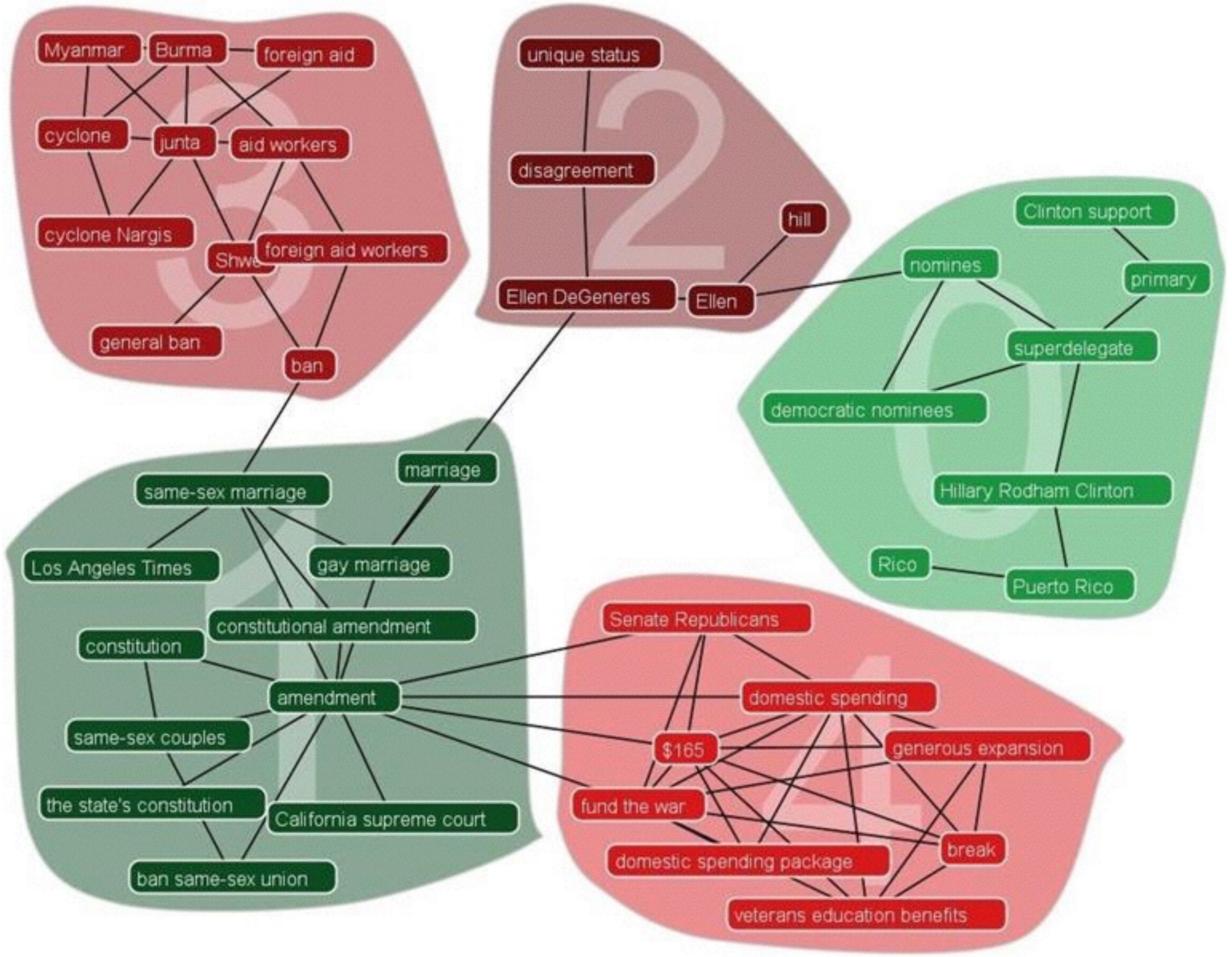
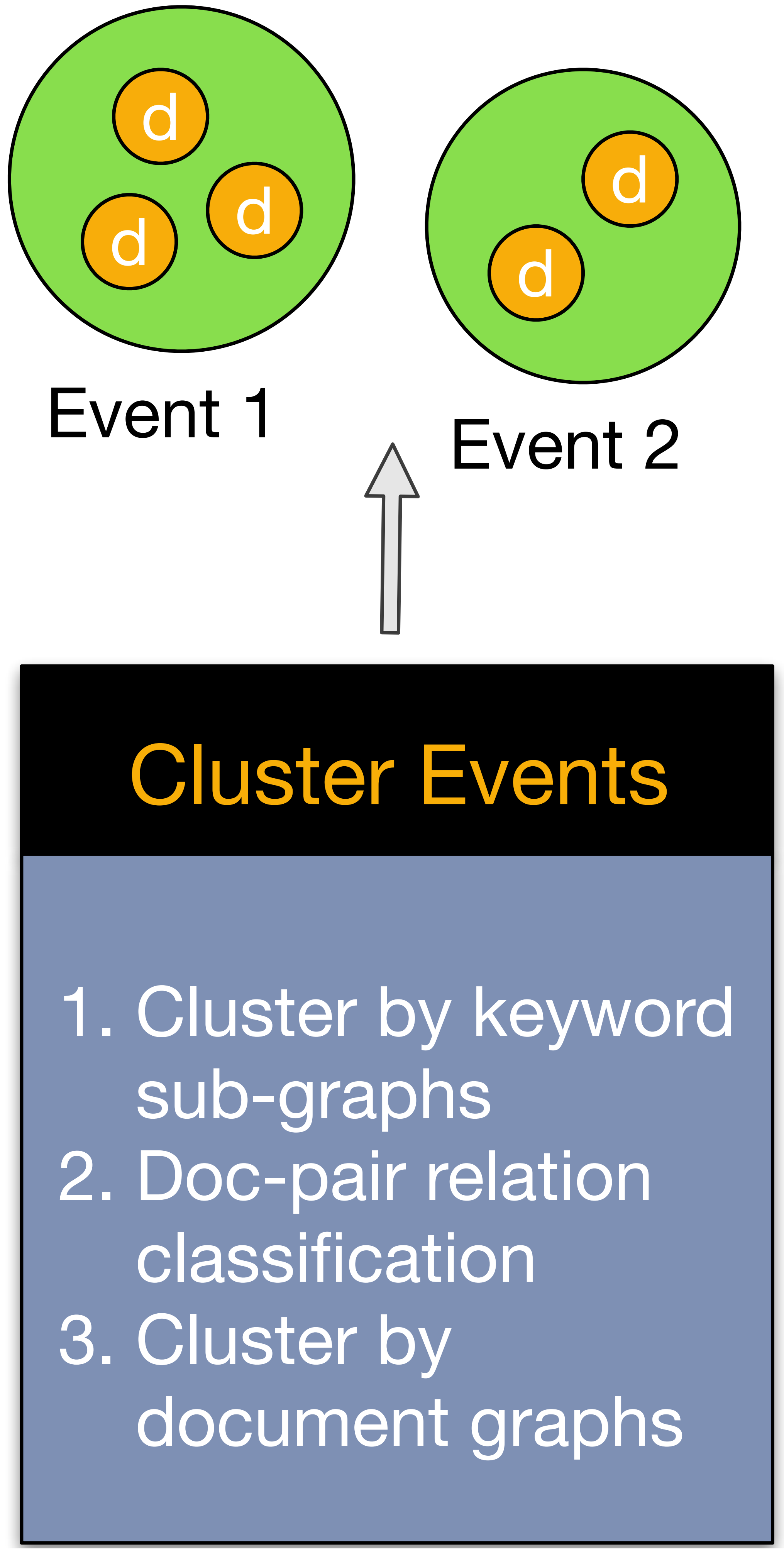


Figure from: Yukio Ohsawa, Nels E Benson, and Masahiko Yachida. 1998. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on. IEEE, 12–18.

Cluster Events



Cluster Events

1. Cluster by keyword sub-graphs
2. Doc-pair relation classification
3. Cluster by document graphs

- Cluster by **Keyword Graph**.
- Extract **doc-pair features**: title similarity measures, content similarity measures, news category, …
- Train an **SVM classifier**: input two documents features, output if they belong to same event or not.
- Community detection on **Document Graph**

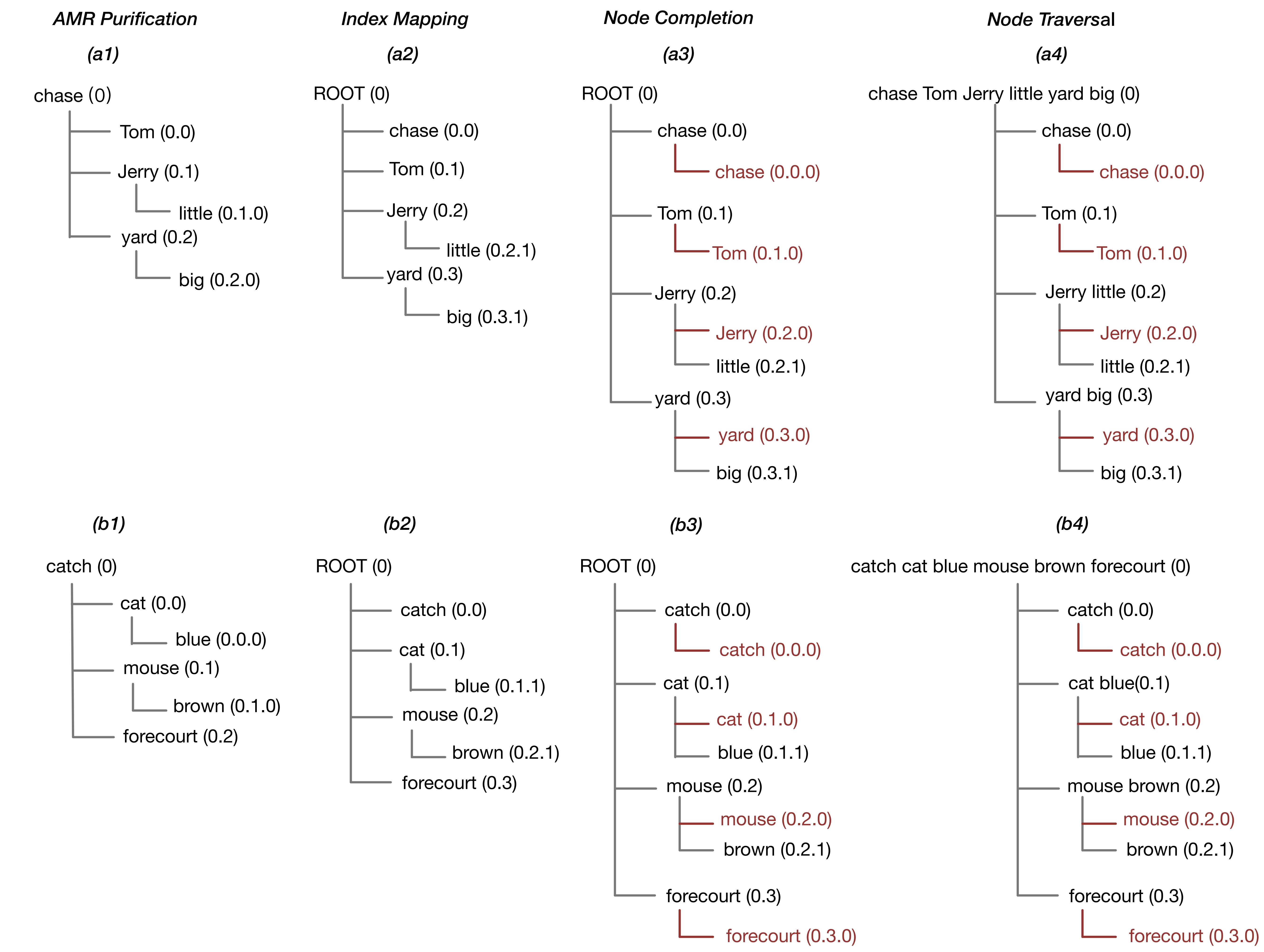
Sentence Matching based on Deep Learning

A. Original sentences

Sentence A: The little Jerry is being chased by Tom in the big yard.

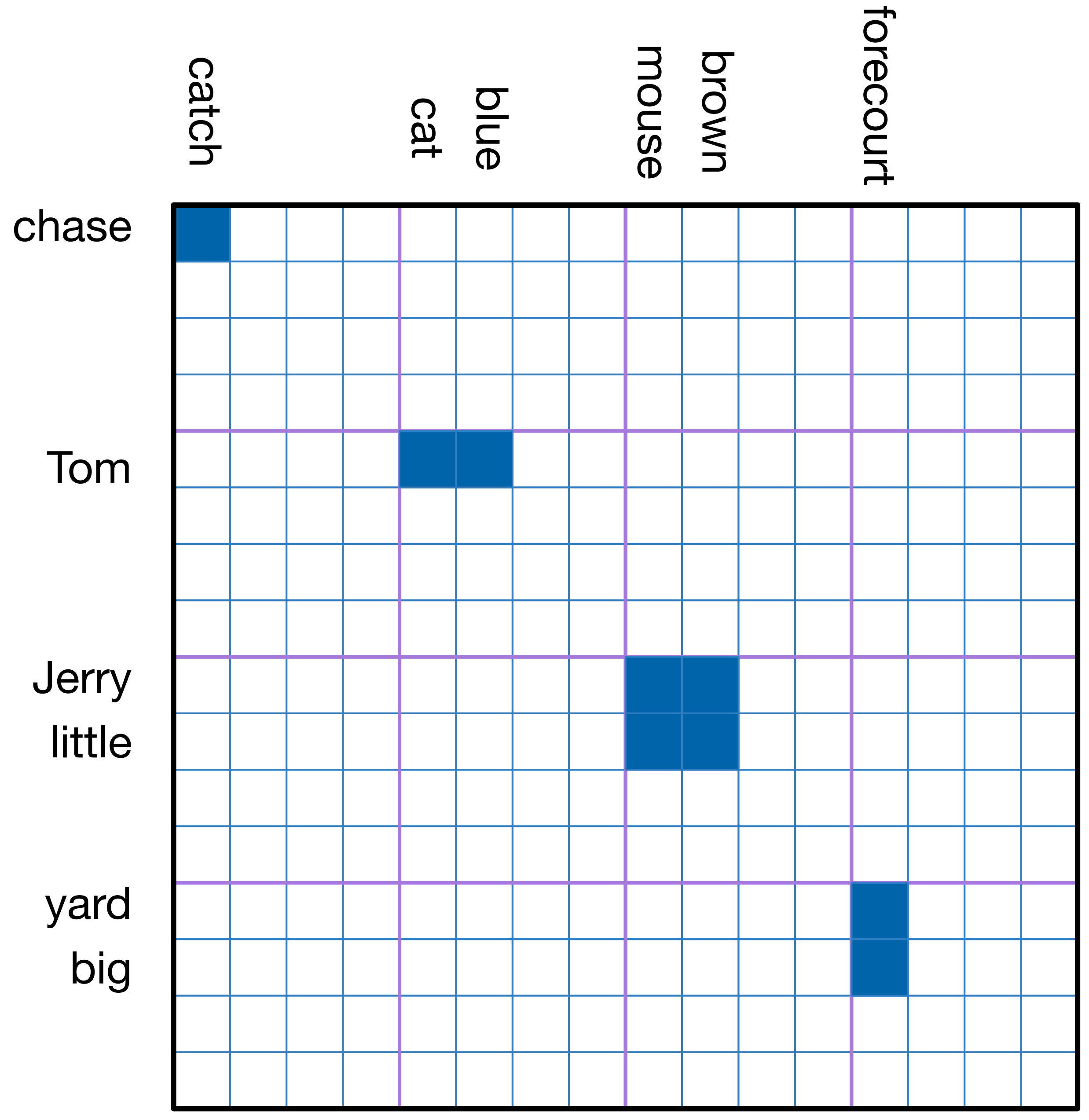
Sentence B: The blue cat is catching the brown mouse in the forecourt.

B. Sentence Factorization Tree

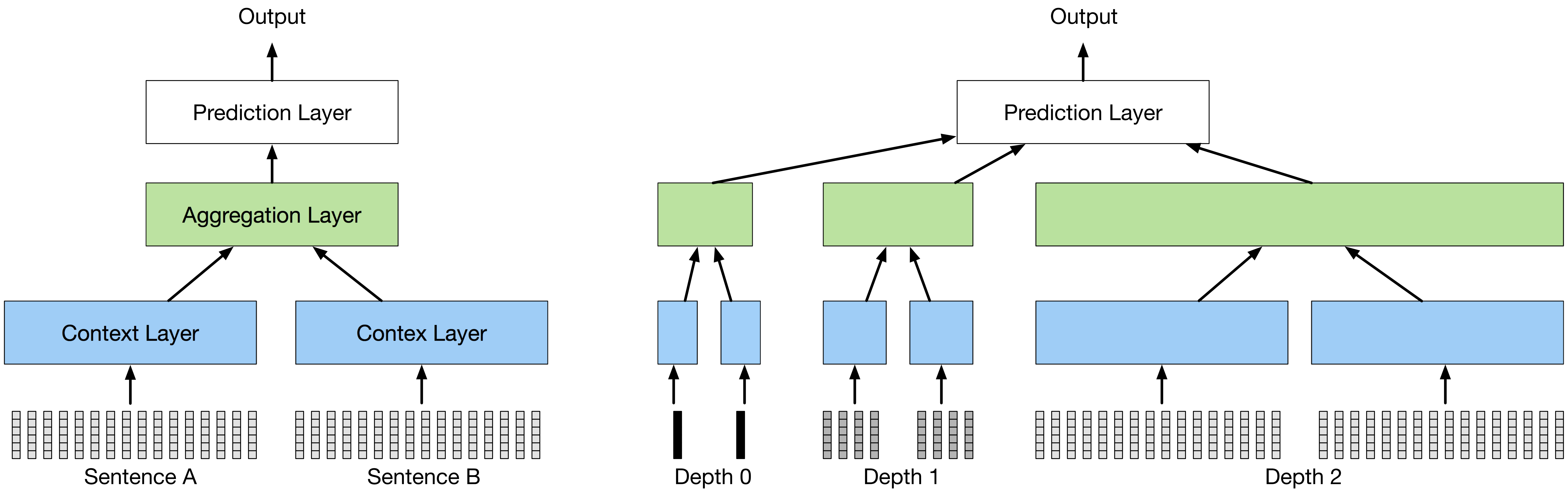


Bang Liu, Ting Zhang, Fred X. Han, Di Niu, Kunfeng Lai and Yu Xu. "Matching Natural Language Sentences with Hierarchical Sentence Factorization," in **WWW 2018.**

D. Semantic units alignments



Sentence Matching based on Deep Learning



(a) Siamese Architecture for Sentence Matching

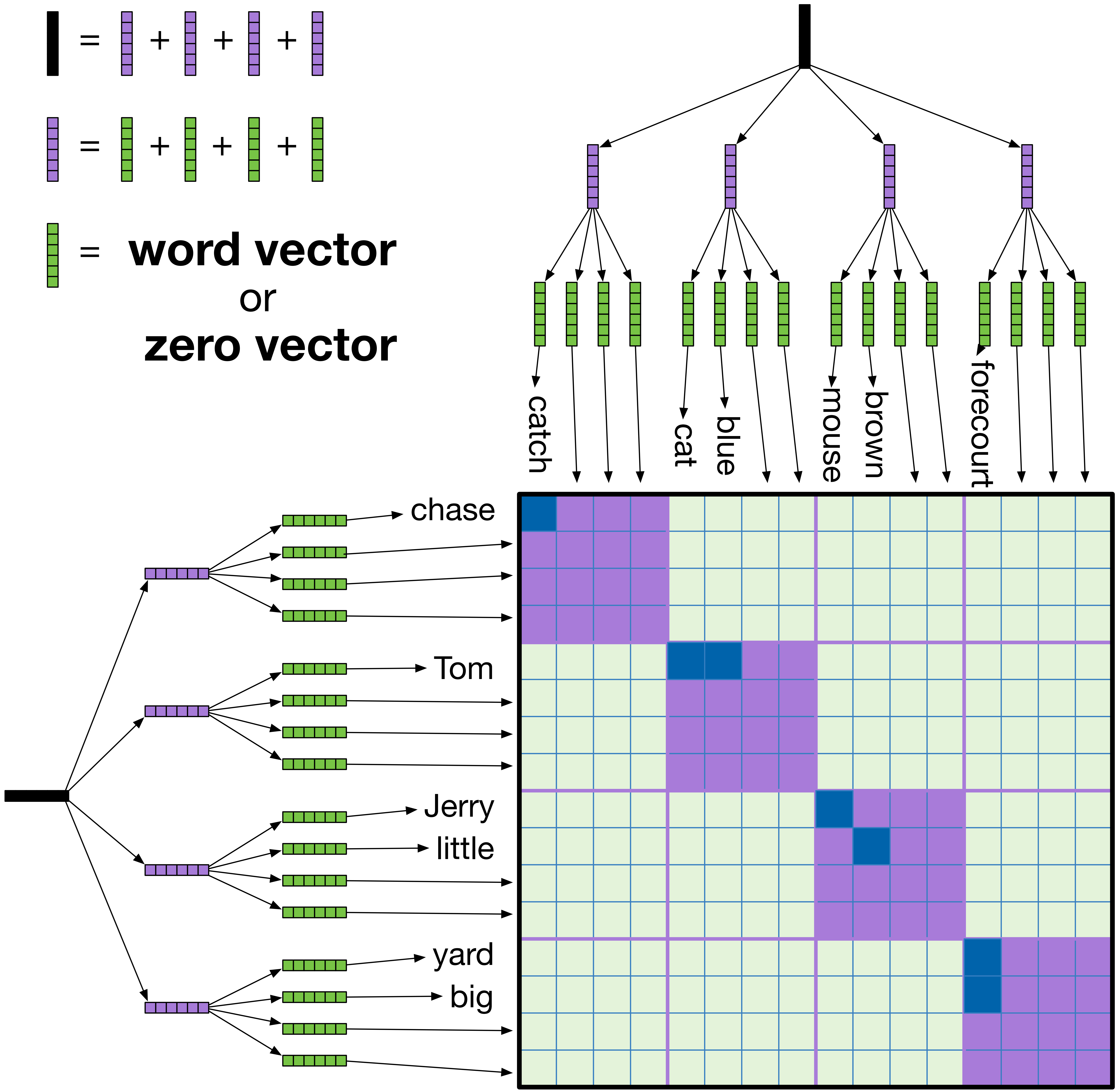
(b) Siamese Architecture with Factorized Multi-scale Sentence Representation

Figure 5: Extend the Siamese network architecture for sentence matching by feeding into the multi-scale representations of sentence pairs.

Open Source: https://github.com/BangLiu/SentenceMatching

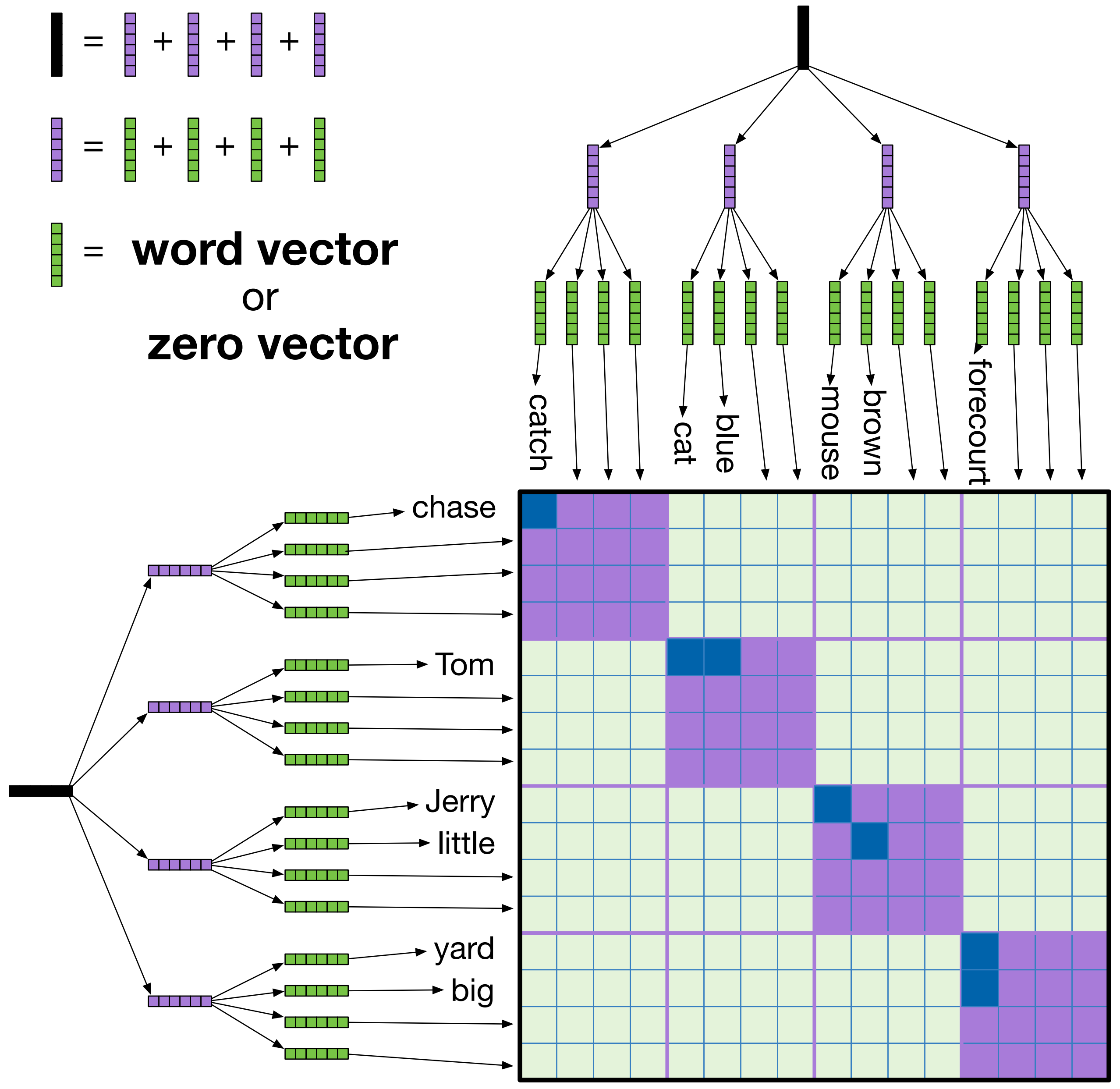
Bang Liu, Ting Zhang, Fred X. Han, Di Niu, Kunfeng Lai and Yu Xu.
“Matching Natural Language Sentences with Hierarchical Sentence Factorization,” in **WWW 2018.**

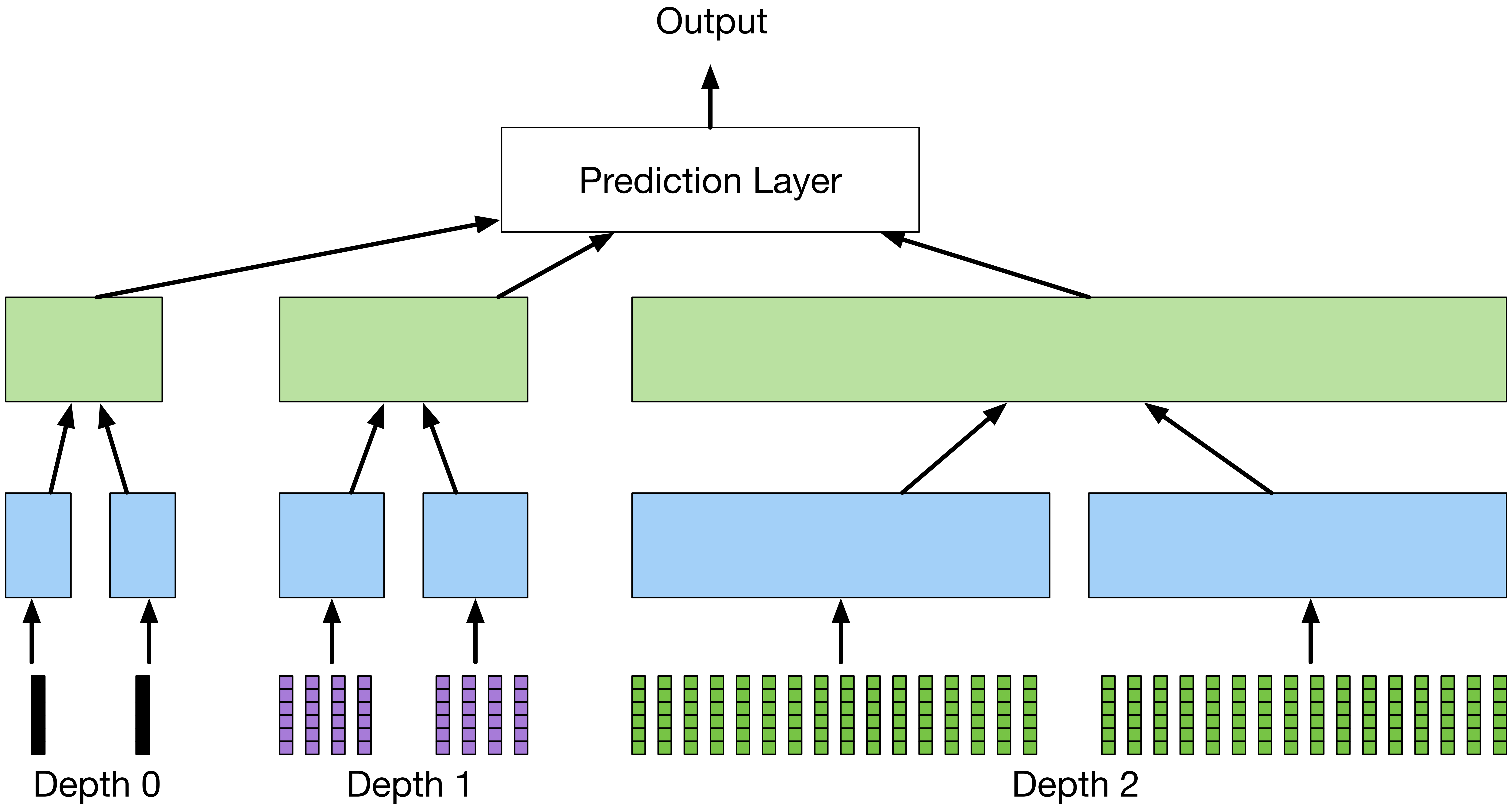
The **blue cat** is **catching** the **brown mouse** in the **forecourt**.



The **little Jerry** is **being chased** by **Tom** in the **big yard**.

Our Method: Multi-scale Sentence Matching





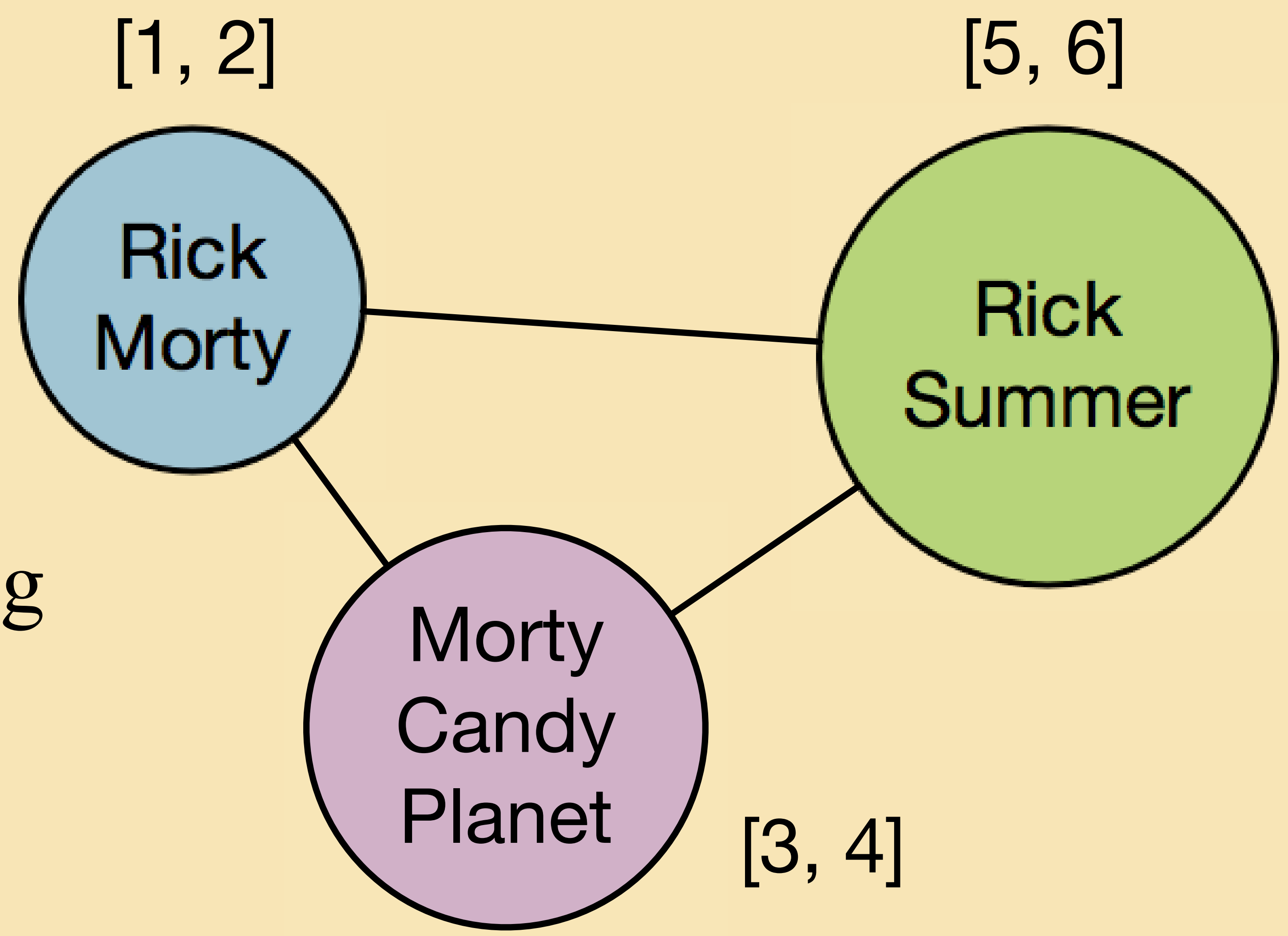
Long Document Matching

A graphical approach to long document matching —— **Concept Interaction Graph**

Text:

[1] Rick asks Morty to travel with him
in the universe.
[2] Morty doesn't want to go as Rick always
brings him dangerous experiences.
[3] However, the destination of this journey
is the Candy Planet, which is an fascinating
place that attracts Morty.
[4] The planet is full of delicious candies.
[5] Summer wishes to travel with Rick.
[6] However, Rick doesn't like to travel
with Summer.

Concept Interaction Graph:



On arXiv and under submission
arXiv:1802.07459

Matching Long Text Documents via Graph Convolutional Networks
Bang Liu, Ting Zhang, Di Niu, Jinghong Lin, Kunfeng Lai, Yu Xu

Long Document Matching

Vertex Feature
Aggregation Layer
Context Layer
Contex Layer
Sentences 1
Sentences 2
Encode each vertex in the input graph
GCN Layer
GCN Layer
Aggregate Layer
Merge Layer
Regression
Match Result
Doc 1
Doc 2
Manually Extracted Features

(a) Siamese Architecture for Text Pair Encoding on Each Vertex

(b) Architecture of Siamese Encoded Graph Convolutional Network for Long Text Pair Matching

A graphical approach to long document matching —— **Graph Convolutional Network**

Matching Long Text Documents via Graph Convolutional Networks
Bang Liu, Ting Zhang, Di Niu, Jinghong Lin, Kunfeng Lai, Yu Xu

Long Document Matching

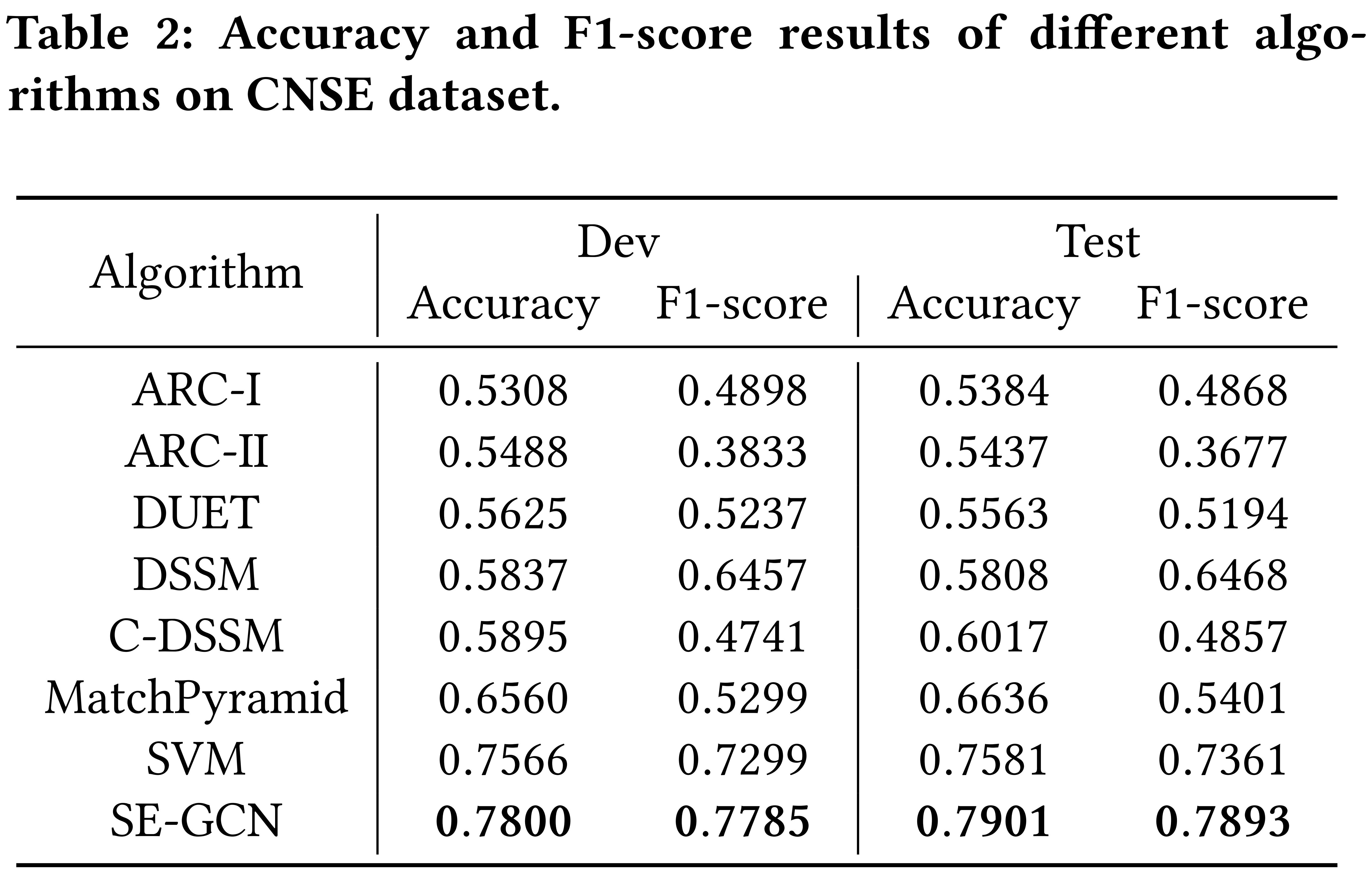


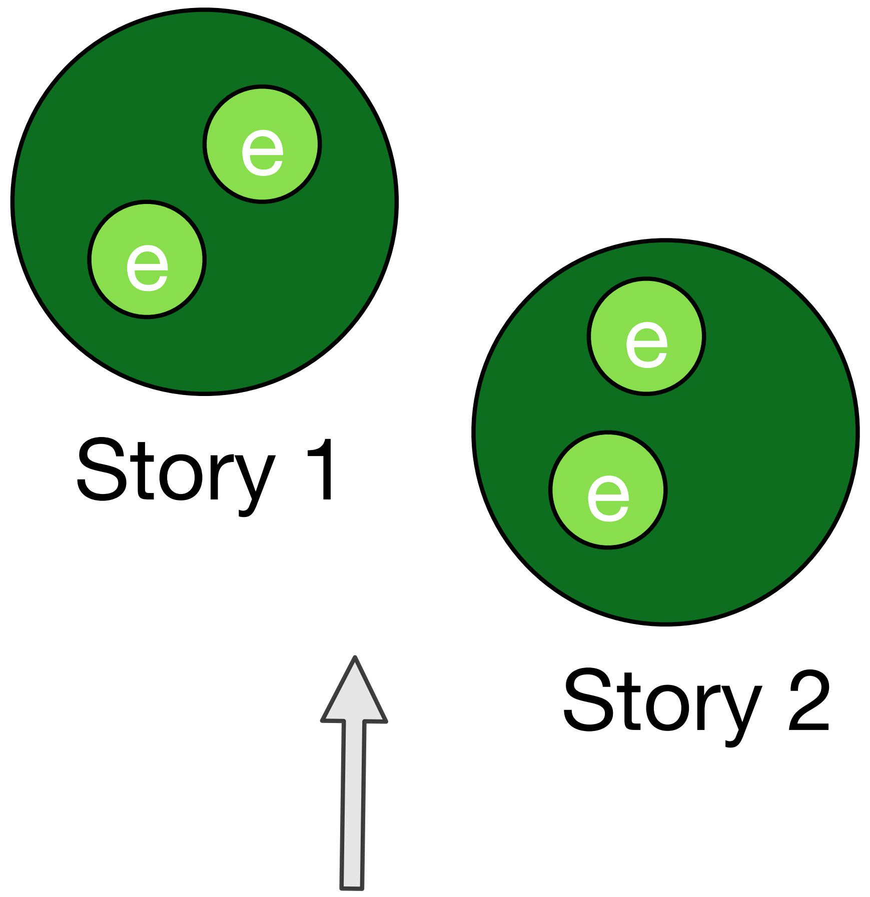
Table 2: Accuracy and F1-score results of different algorithms on CNSE dataset.

Algorithm	Dev		Test	
	Accuracy	F1-score	Accuracy	F1-score
ARC-I	0.5308	0.4898	0.5384	0.4868
ARC-II	0.5488	0.3833	0.5437	0.3677
DUET	0.5625	0.5237	0.5563	0.5194
DSSM	0.5837	0.6457	0.5808	0.6468
C-DSSM	0.5895	0.4741	0.6017	0.4857
MatchPyramid	0.6560	0.5299	0.6636	0.5401
SVM	0.7566	0.7299	0.7581	0.7361
SE-GCN	**0.7800**	**0.7785**	**0.7901**	**0.7893**

A graphical approach to long document matching —— **Graph Convolutional Network**

Matching Long Text Documents via Graph Convolutional Networks
Bang Liu, Ting Zhang, Di Niu, Jinghong Lin, Kunfeng Lai, Yu Xu

Cluster Stories

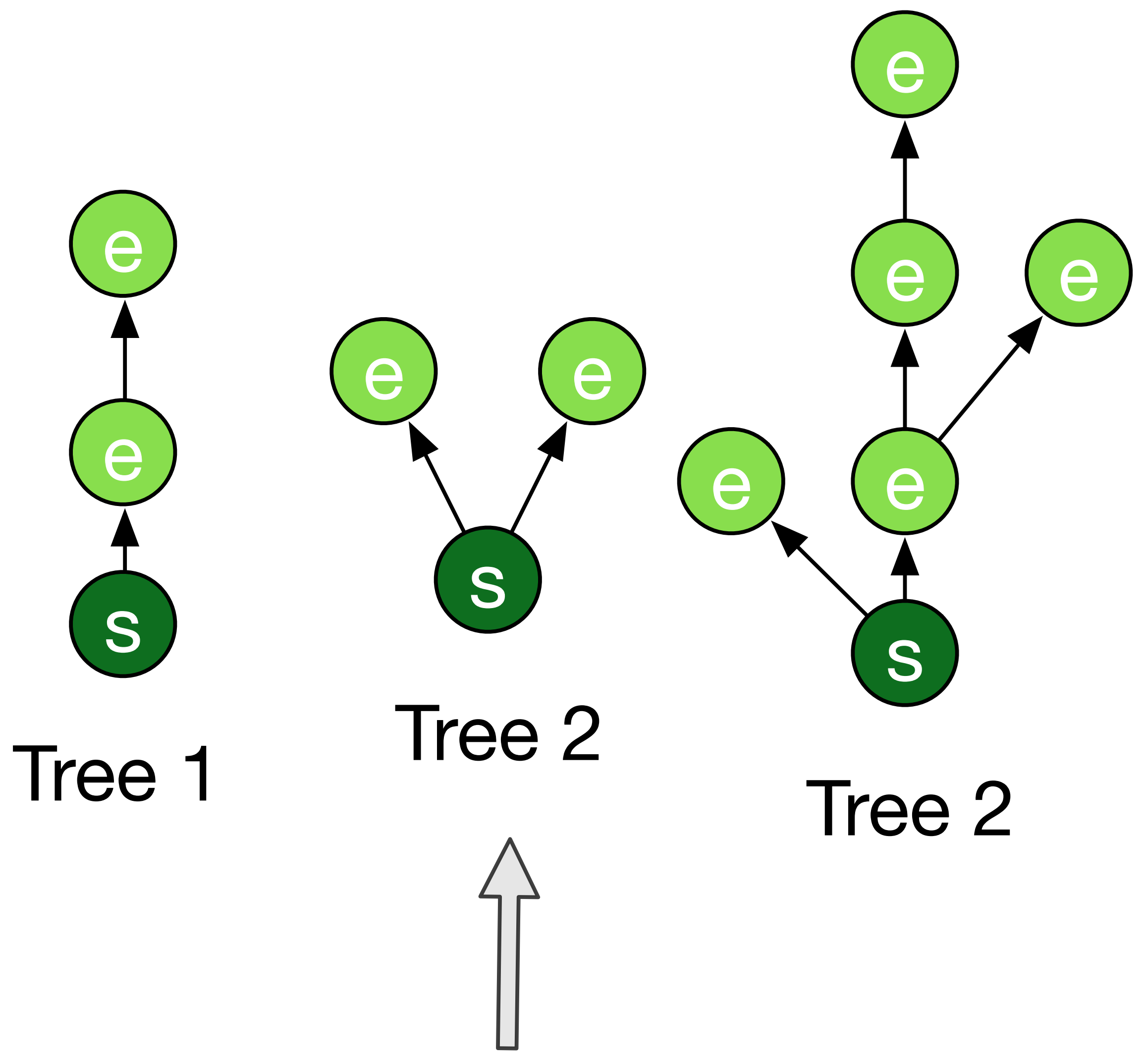


Cluster Stories

1. Find the story to which each event belongs
2. Add events to existing stories, or create new stories

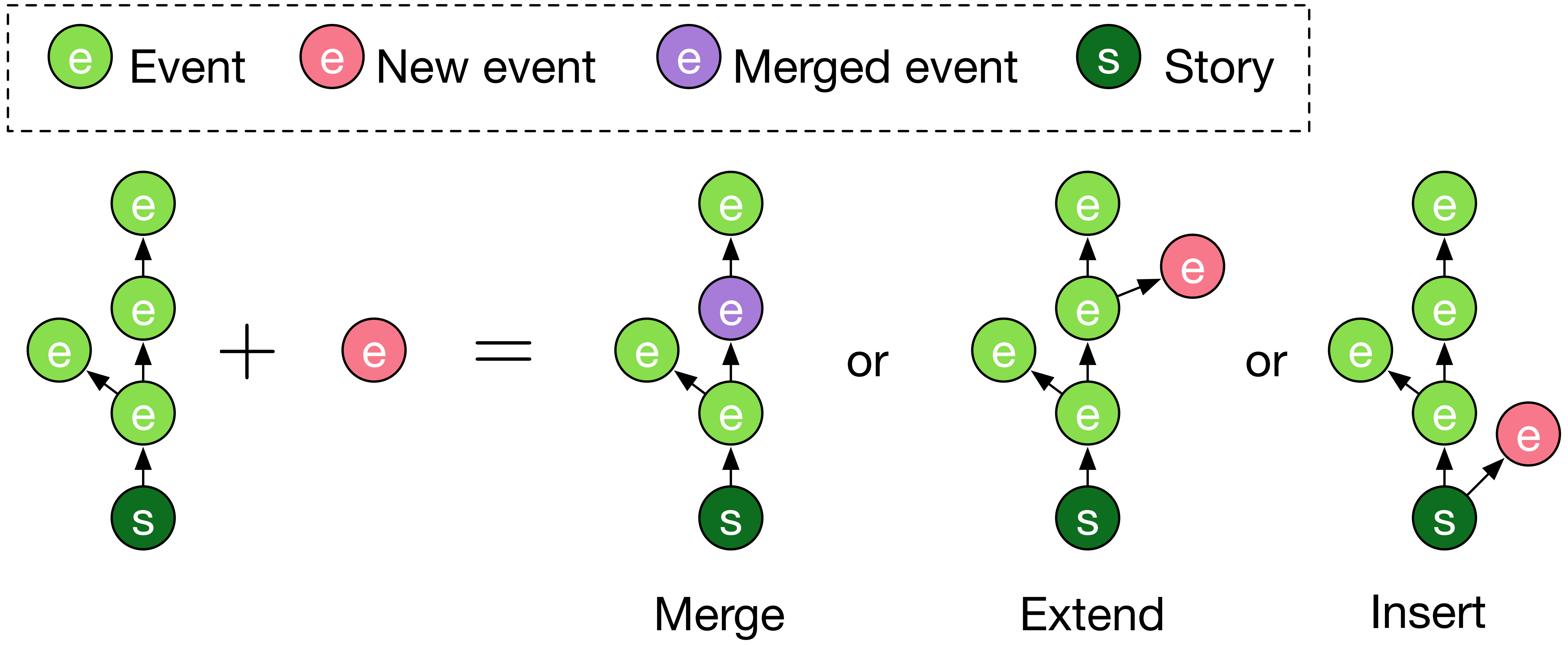
Story: multiple events that are interdependent and evolve over time form a story.

Story Structure Generation



Grow Story Forest

1. Merge same events
2. Update story tree structure with new events



Choose the best position in the tree to insert a new event node

Clustering Performance

- **LDA+Affinity Propagation**: extract 1000 dimensional LDA feature, clustering by Affinity Propagation.
- **KeyGraph**: the original KeyGraph algorithm proposed in [1], which doesn't include the second step in our approach.
- **StoryForest**: our approach.

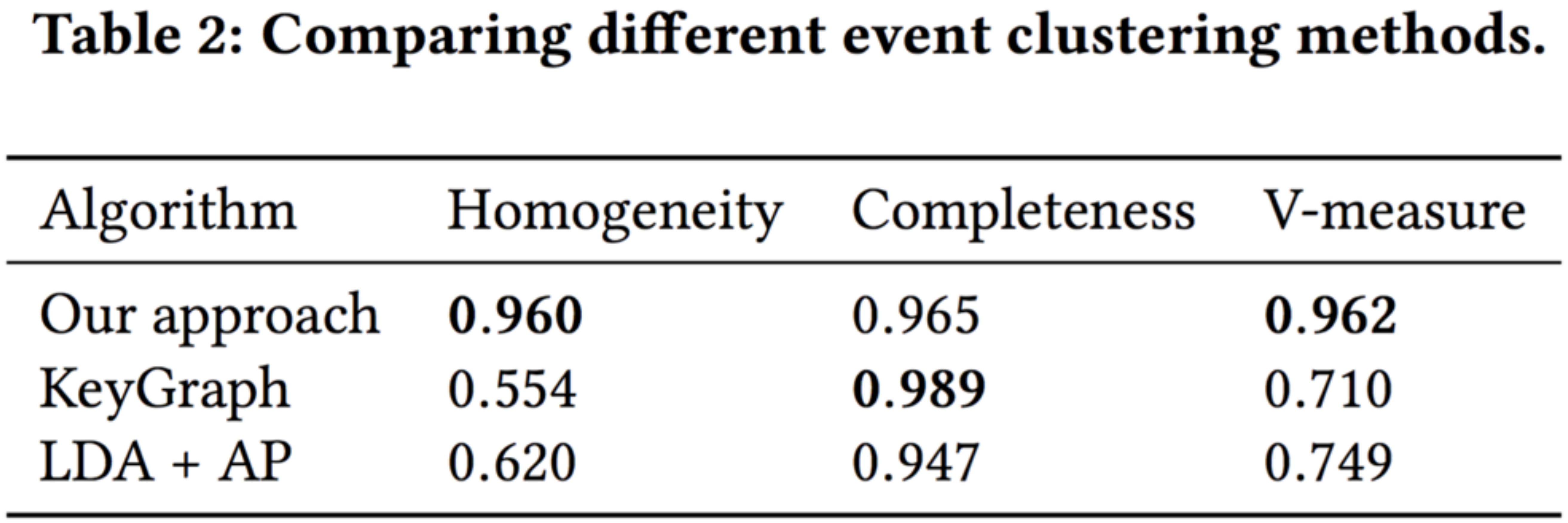


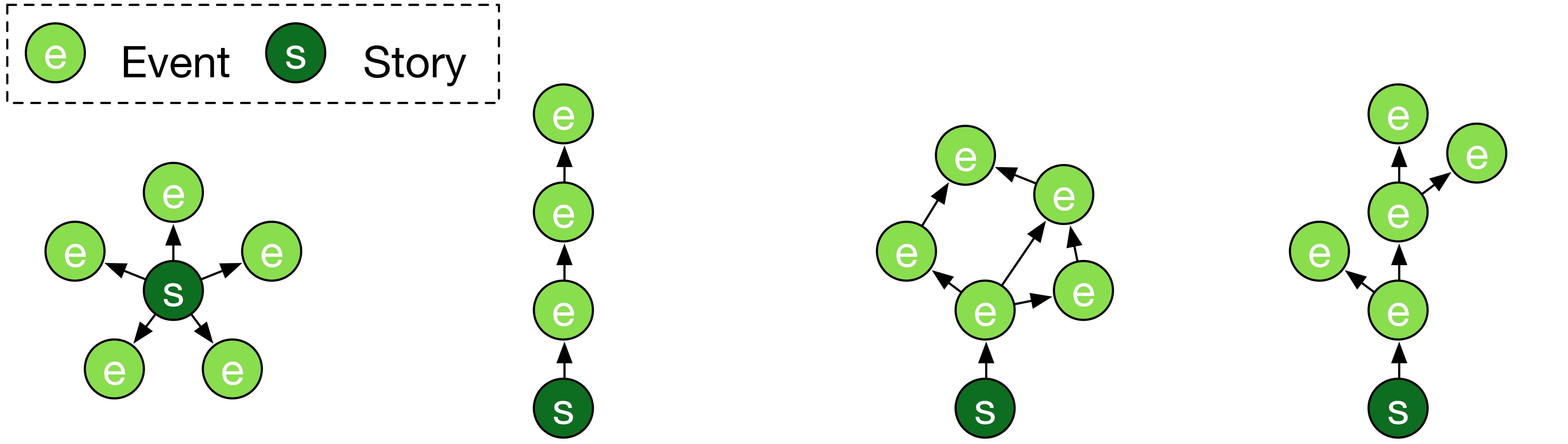
Table 2: Comparing different event clustering methods.

Algorithm	Homogeneity	Completeness	V-measure
Our approach	**0.960**	0.965	**0.962**
KeyGraph	0.554	**0.989**	0.710
LDA + AP	0.620	0.947	0.749

Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, Yu Xu. "Growing Story Forest Online from Massive Breaking News," in **CIKM 2017**.

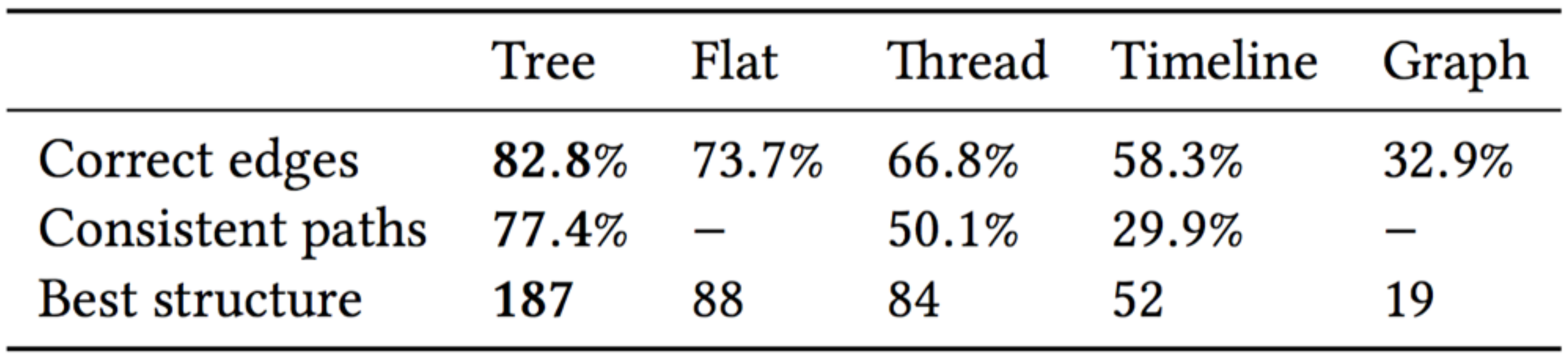
Story Structure Performance

- **Flat Cluster:** cluster by stories, no structure.
- **Story Timeline:** organizes events linearly by time.
- **Story Graph:** calculates a connection strength for each pair of events and connect the pair if the score exceeds a threshold.
- **Event Threading:** appends each event to its most similar earlier event. Similarity measured by TF-IDF.



(a) Flat structure (b) Timeline structure (c) Graph structure (d) Tree structure

Table 3: Comparing different story structure generation algorithms.

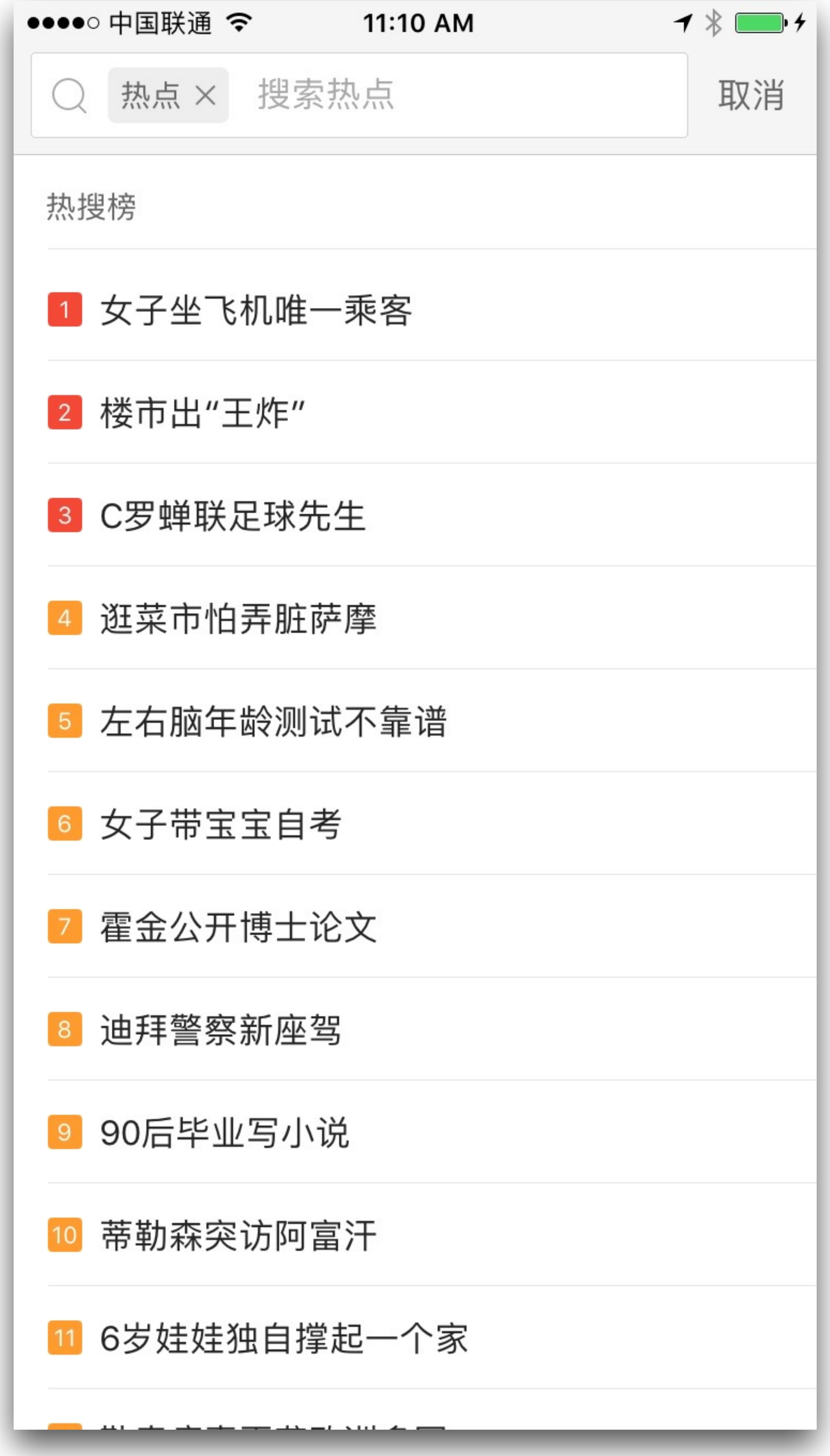


	Tree	Flat	Thread	Timeline	Graph
Correct edges	**82.8%**	73.7%	66.8%	58.3%	32.9%
Consistent paths	**77.4%**	–	50.1%	29.9%	–
Best structure	**187**	88	84	52	19

(from the **CIKM 2017** paper)

Deployed in Tencent QQ browser
The hot topic list

Dr. Hawking’s PhD thesis made public



热搜榜

1 女子坐飞机唯一乘客

2 楼市出“王炸”

3 C罗蝉联足球先生

4 逛菜市怕弄脏萨摩

5 左右脑年龄测试不靠谱

6 女子带宝宝自考

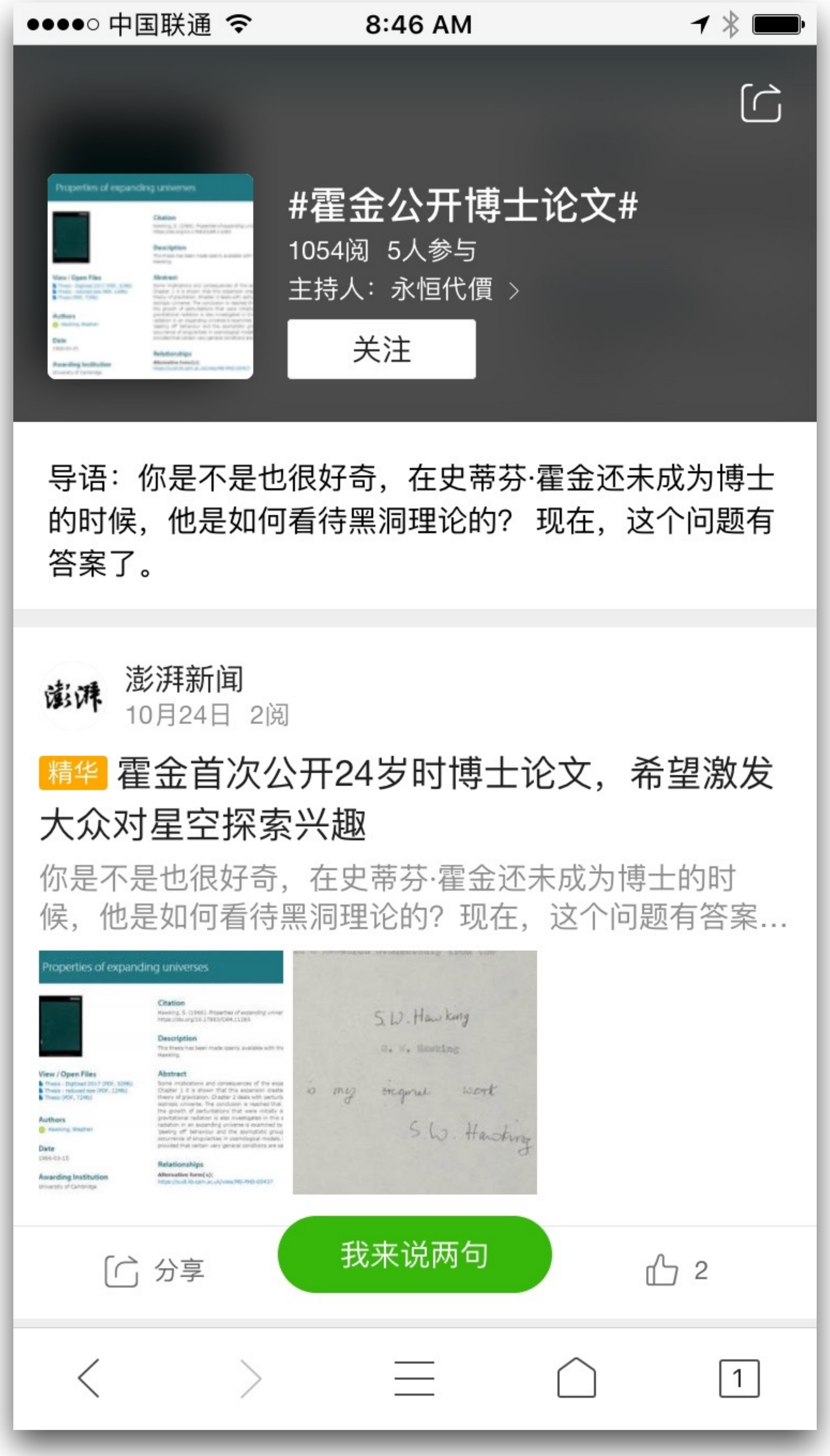
7 霍金公开博士论文

8 迪拜警察新座驾

9 90后毕业写小说

10 蒂勒森突访阿富汗

11 6岁娃娃独自撑起一个家



#霍金公开博士论文#

1054阅 5人参与

主持人：永恒代價 >

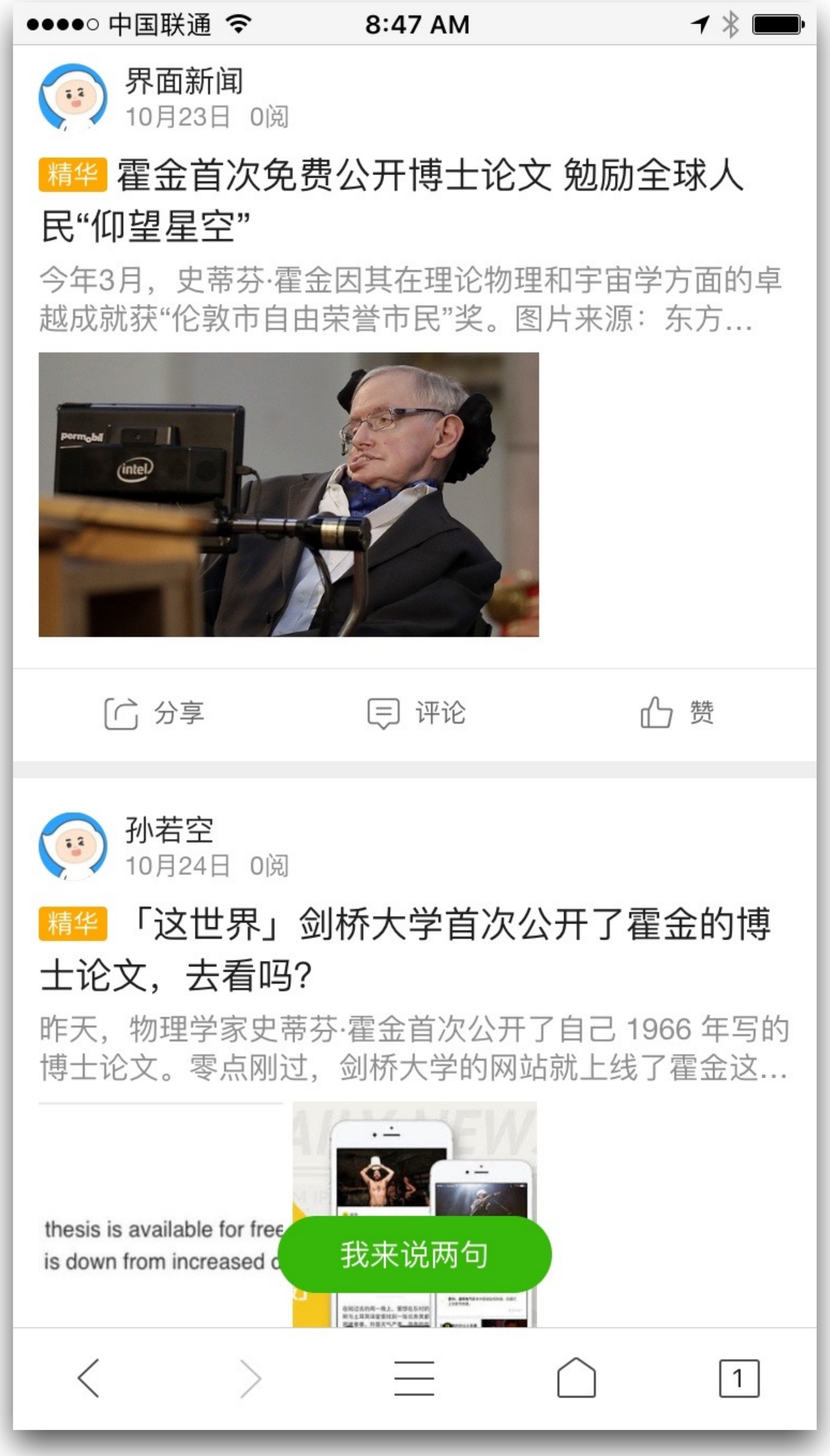
导语：你是不是也很好奇，在史蒂芬·霍金还未成为博士的时候，他是如何看待黑洞理论的？ 现在，这个问题有答案了。

澎湃新闻

10月24日 2阅

精华 霍金首次公开24岁时博士论文，希望激发大众对星空探索兴趣

你是不是也很好奇，在史蒂芬·霍金还未成为博士的时候，他是如何看待黑洞理论的？现在，这个问题有答案…



界面新闻

10月23日 0阅

精华 霍金首次免费公开博士论文 勉励全球人民“仰望星空”

今年3月，史蒂芬·霍金因其在理论物理和宇宙学方面的卓越成就获“伦敦市自由荣誉市民”奖。图片来源：东方…

孙若空

10月24日 0阅

精华 「这世界」剑桥大学首次公开了霍金的博士论文，去看吗？

昨天，物理学家史蒂芬·霍金首次公开了自己 1966 年写的博士论文。零点刚过，剑桥大学的网站就上线了霍金这…

thesis is available for free
is down from increased d

Thanks!

Q&A