数据驱动的推特新闻事件挖掘

Data Driven News Event Mining from Twitter.

报告人：黎谢鹏　指导老师：胡岩峰

目录

一、背景与意义

- 社交媒体作为社会传感器为社会科学研究提供了新的数据源。
- 新闻事件挖掘研究可以辅助社会管理者以更高的效率，更全面的信息做出正确的管理决策。

二、研究内容：

- 有监督学习效果好，但数据是瓶颈，构造可用的数据集是关键。
  - 难点：海量，噪音，冗余，多样性
- 知识产生决策，事件概念图谱是利用事件相关知识进行决策辅助的基础。
  - 难点：链接预测时图遍历算法复杂性高

二、研究内容：

新闻事件挖掘任务:

新闻事件检测系统流程

- 推文过滤
- 推文分类
- 推文聚类
- 推文排序

新闻事件数据集构建

- 相关推文检索
- 相关推文过滤与排序

新闻事件概念图谱构建

- 推文概念化
- 事件模式抽取
- 事件概念抽取

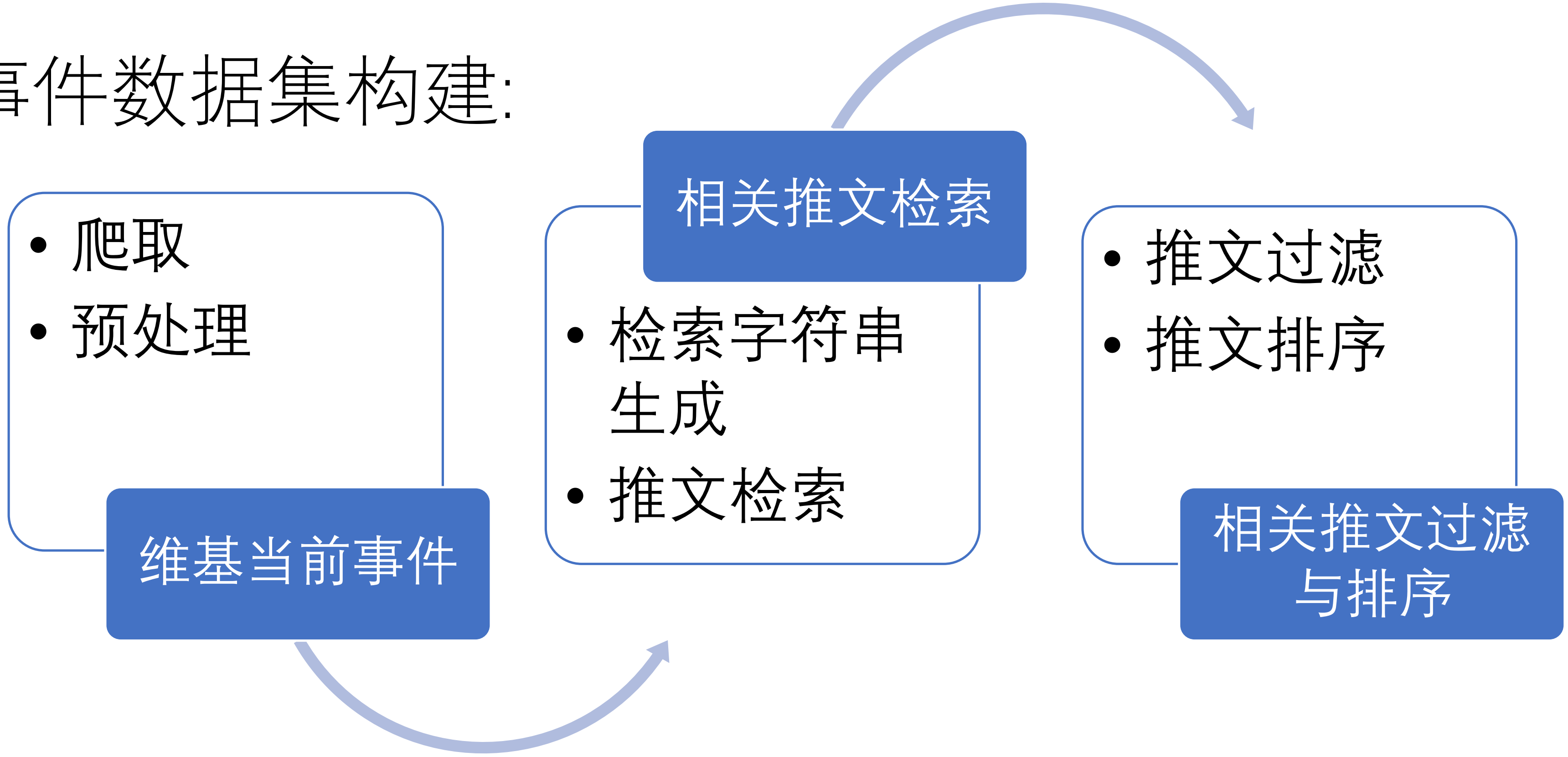
三、研究进展：

新闻事件检测系统流程

- Build a system for event detection in twitter.
- Training a classifer for Tweet classification using char-cnn and word2vec and lexical feature for tweet representation.
- Using K-means for event instance clusering, NER and SRL model for event information extraction.
- Using MMR-submodular based text summarization technology to rank relevant tweets.

三、研究进展：

新闻事件数据集构建:



Wiki current event demo

- An early-morning landslide buries 40 homes and leaves 15 people dead and 114 others missing in Aba Prefecture, Sichuan Province, China. At least 500 rescue workers are on scene, and a 2-km stretch of the river in Mao County is blocked.
- Label: ’disaster and accident’

- 中国四川省阿坝州清晨的滑坡摧毁40所房屋，造成15人死亡和114人失踪。至少有500名救援人员在现场，毛县一条2公里长的河流被封锁。
- 标签：自然灾害与事故

Related Tweets Query Generation

- NER + top_related([V,N]) combination
  - NUM:40 15 114 500
  - LOC: Aba.CITY Prefecture.LOCATION Sichuan.STATE_OR_PROVINCE Province.LOCATION   China.COUNTRY Mao.CITY County.LOCATION
  - V: buries leaves missing blocked
  - N: landslide homes people others scene stretch river
  - Related_score(word):=boe_cosine(word,'disaster and accident')
  - Combination： c_4^2
  - Query: Combination;time [-24,+24]

Related Tweets filtering and ranking

- Related_score = boe_cosine(wiki_description,tweet)
- Filtering criteria: (boe_cosine > 0.75) &(10 < num__top_tweets = 60)
- Ranking method:
  - MMR(Maximal Marginal Relevance)-submodular
  - Greedy criteria: F_mmr = graph_cut – penalty – contradiction
  - Graph_cut: lambda*sim(selected,unselected)
  - Penalty: (1-lamda)*sim(C_s^2)
  - Contradiction: beta*Log(CrossEntropy(Label_distribution(wiki_description), Label_distribution(tweet)))

Related Tweets filtering and ranking

- Demo
  - MMR-submodular based rank:
    - 0.9582722326723183 China : Death toll rises to 15 and over 120 others are missing after a landslide in south-western Sichuan province .
    - 0.9286372801720572 Rescue operation is under way in Sichuan province after more than 40 homes in Xinmo village were engulfed by lan.
    - 0.9461324580032283 More than 120 people are missing after a landslide in Sichuan province in south-western China, 40 homes were destroyed in Xinmo village.
    - 0.9379743462839512 Fifteen people were killed in a landslide in southwest China 's Sichuan Province on Saturday and about 100 were.
    - 0.9365613210034124 Dozens of homes are destroyed and at least 120 people are missing in the wake of a massive landslide in China .

Dataset Usage Demo

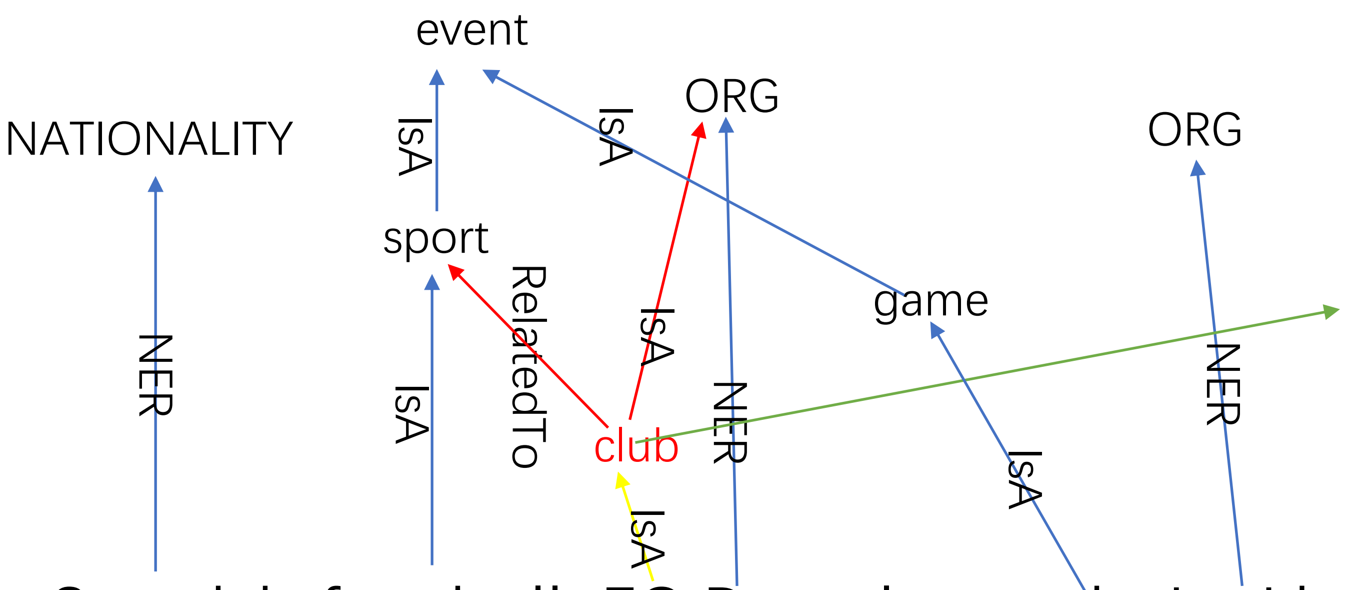
- Top 5: multi-reference news summaration dataset
- Top 15: News Event Concept Graph Building
- Top 30: Classification dataset
- Top 45: Clustering Algorithm testing

三、研究进展：

新闻事件概念图谱构建

- Probase+ConceptNet5：单词或词组表示概念
- Word2vec 候选项高速查询
- HIN（异质信息网络）：
  - G(V,E)
  - $\varphi = V \to A$
  - $\psi = E \to R$
  - $|A|$>1 or $|R|$>1
  - Meta-path（作为链接预测的特征）: $A_1 \overset{R_1}{\to} A_2 \overset{R_2}{\to} \cdots \overset{R_l}{\to} A_{l+1}$

Simple demo: instance conceptulize



Inference by metapath:
p=(c)<-[:`/r/RelatedTo`]-(a)-[:`/r/IsA`]->(b)
a: inferenced type
b: NER type
c: context concept

- In Spanish football, FC Barcelona win La Liga.
- Concept hierarchy level:
  - Level-0: event, entity [predefine]
  - Level-1: nationality, organization, sport, game, club [ner & query]
  - Level-2: football, win, club [extract & inference]
  - Level-instance: Spanish, FC Barcelona, La Liga

Simple demo: event entity & concept extraction

- Event info:
  - {‘event.type’:[’sport.football’,’game.win’],
  - ’club.instance’: [‘FC Barcelona’, ‘La Liga’],
  - ’ football.hasSubEvent’:‘win’,
  - ’event.pattern’:’ In [entity.nationality] [event.sport.football], [entity.organization.club] [event.game.win] [entity.organization.club].’}
- tweet-based event.pattern-[frequent]->event instance pattern –[frequent]-> event category pattern

Conceptulize the trigger V.&N.

- Rank_score(word)= boe_cosine(word,tweet_categorical_label)
- football,win
- MATCH p=(a)-[:`/r/IsA`|:`/r/microsoft/IsA`|:`/r/InstanceOf`*1..2]->(b) WHERE a.conceptId IN ['/c/en/football'] AND b.conceptId = "/c/en/event" RETURN extract(x IN nodes(p) | x.conceptId) as nodes,extract(x IN relationships(p) | type(x)) as rels,reduce(prob = 1.0, x IN relationships(p) | prob*x.weight) as probs ORDER BY probs DESC LIMIT 10;

Vector similarity context vs (c)<-[:`/r/RelatedTo`]-(a)

- wv.most_similar(positive=['win','football','organization'], topn=10, indexer=annoy_index)
  - [(‘football', 0.6417830884456635),
  - (‘league’, 0.6238529682159424),   #联赛
  - (‘win', 0.622621089220047),
  - (‘team’, 0.5995824038982391),   #球队
  - (‘winning', 0.5774330496788025),
  - (‘teams’, 0.5667363405227661),   #球队
  - (‘soccer’, 0.5631313323974609),   #冠军
  - (‘championship', 0.5575762391090393),
  - (‘club’, 0.55101752281189),   #俱乐部
  - (‘baseball', 0.5484375953674316)]

Concept inference by neo4j Cypher query:

- Topk=10
  vector_similarity_context=['/c/en/league','/c/en/team','/c/en/championship','/c/en/club']
- MATCH p=(a)-[:`/r/IsA`|:`/r/microsoft/IsA`|:`/r/InstanceOf`*1..1]->(b) WHERE a.conceptId IN {vector_similarity_context} AND b.conceptId = "/c/en/organization" RETURN extract(x IN nodes(p) | x.conceptId) as nodes,extract(x IN relationships(p) | type(x)) as rels,reduce(prob = 1.0, x IN relationships(p) | prob*x.weight) as probs ORDER BY probs DESC LIMIT {topk};
- Add type constraint on vector_similarity_context

Concept inference by neo4j Cypher query:

```
MATCH p=(a)-[:`/r/IsA`|:`/r/microsoft/IsA`|:`/r/InstanceOf`*1..1]->(b) WHERE a.conceptId IN ['/c/en/league','/c/en/team','/c/en/
championship','/c/en/club']
 AND b.conceptId = "/c/en/organization" RETURN extract(x IN nodes(p) | x.conceptId) as nodes,extract(x IN relationships(p) |
type(x)) as rels,reduce(prob = 1.0, x IN relationships(p) | prob*x.weight ) as probs  ORDER BY probs DESC LIMIT 10;
```

```
H p=(a)-[:`/r/IsA`|:`/r/microsoft/IsA`|:`/r/InstanceOf`*1..1]->(b) WHERE a.conceptId IN ['/c/en/league','/c/en/team','/c/en/ championshi…
```

nodes	rels	probs
["/c/en/club", "/c/en/organization"]	["/r/microsoft/IsA"]	0.012199999764561653
["/c/en/league", "/c/en/organization"]	["/r/microsoft/IsA"]	0.00039999998989515007
["/c/en/team", "/c/en/organization"]	["/r/microsoft/IsA"]	0.00039999998989515007

Level-2 concept relation:

- MATCH p=(a:Concept {conceptId:'/c/en/football'})-[*1..2]->(b:Concept {conceptId:'/c/en/win'}) RETURN extract(x IN nodes(p)|x.conceptId) as nodes,extract(x IN relationships(p)|type(x)) as rels,reduce(prob = 1.0,x IN relationships(p)|prob*x.weight) as probs ORDER BY probs DESC LIMIT 25;

nodes	**rels**	**probs**
["/c/en/football", "/c/en/game", "/c/en/win"]	["/r/IsA", "/r/RelatedTo"]	4.463354876537323
["/c/en/football", "/c/en/game", "/c/en/win"]	["/r/RelatedTo", "/r/RelatedTo"]	0.049673997903823874
["/c/en/football", "/c/en/game", "/c/en/win"]	["/r/microsoft/IsA", "/r/RelatedTo"]	0.027612898706710354
["/c/en/football", "/c/en/playing_game", "/c/en/win"]	["/r/microsoft/IsA", "/r/HasSubevent"]	0.0063687002355383715
["/c/en/football", "/c/en/playing_game", "/c/en/win"]	["/r/microsoft/IsA", "/r/Causes"]	0.005200000014156103
["/c/en/football", "/c/en/playing_sport", "/c/en/win"]	["/r/microsoft/IsA", "/r/HasSubevent"]	0.0027000000700354576
["/c/en/football", "/c/en/playing_sport", "/c/en/win"]	["/r/microsoft/IsA", "/r/Causes"]	0.0027000000700354576
["/c/en/football", "/c/en/play_game", "/c/en/win"]	["/r/microsoft/IsA", "/r/MotivatedByGoal"]	0.00039999998989515007
["/c/en/football", "/c/en/play_game", "/c/en/win"]	["/r/microsoft/IsA", "/r/HasSubevent"]	0.00019999999494757503
["/c/en/football", "/c/en/reward", "/c/en/win"]	["/r/microsoft/IsA", "/r/RelatedTo"]	0.0000174999992598896

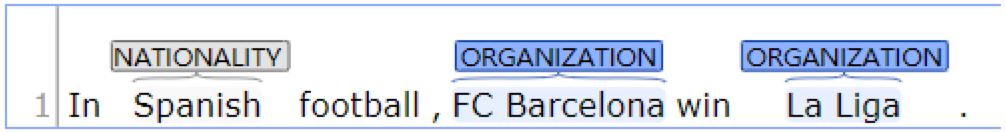
Level-2 concept relation:

- Contraint on core event relation
  - 同义词-表达多样性 :`/r/FormOf`|:`/r/Synonym`|:`/r/DerivedFrom`
  - 上下位 :`/r/IsA`|:`/r/MannerOf`
  - 因果关系 :`/r/Causes`
  - 目的关系 :`/r/MotivatedByGoal`|:`/r/Desires`
  - 包含关系 :`/r/HasSubevent`
  - 条件关系 :`/r/HasPrerequisite`|:`/r/Entails`
- Candidate pairs <s,t>:
  - <football,win>
  - <football,club>
  - <win,club>

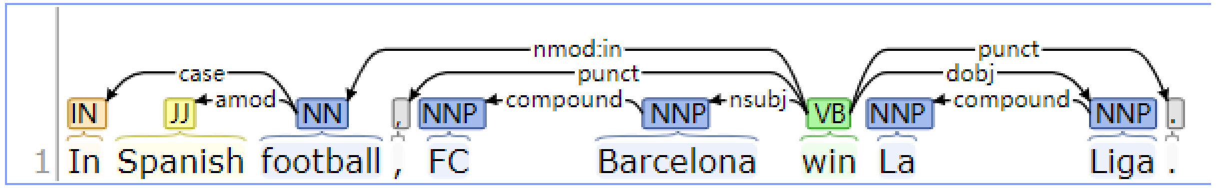
Link prediction in HIN using meta-path feature:

1.fusion of dependencies-graph & concept-graph

Named Entity Recognition:

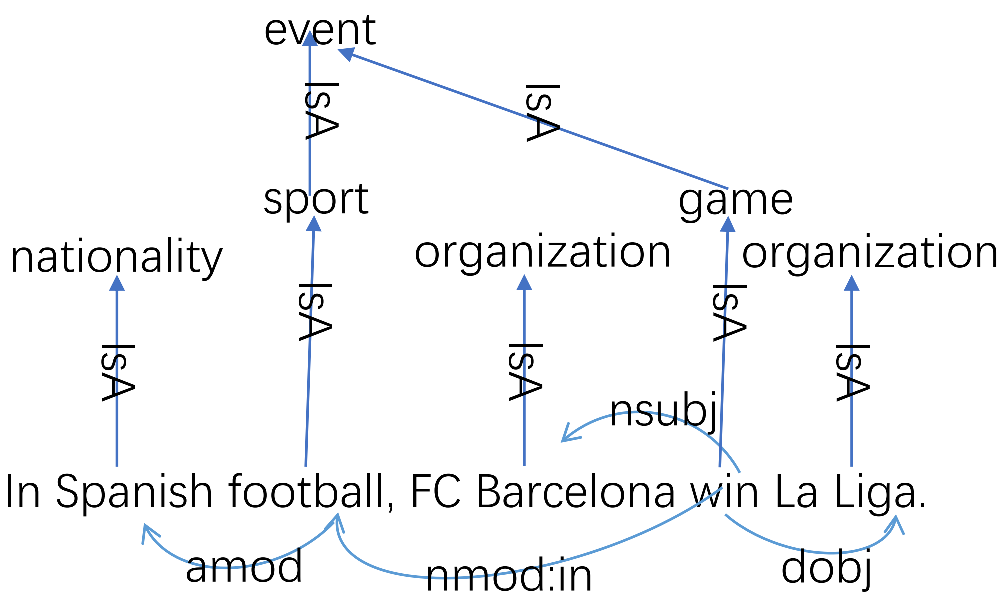


Enhanced++ Dependencies:



"nodes"	"rels"	"probs"
["/c/en/football","/c/en/sport","/c/en/event"]	["/r/IsA","/r/microsoft/IsA"]	0.1728000044822693
["/c/en/football","/c/en/game","/c/en/event"]	["/r/IsA","/r/microsoft/IsA"]	0.13930800176915525
["/c/en/football","/c/en/event"]	["/r/microsoft/IsA"]	0.0390000000804662704
["/c/en/football","/c/en/sport","/c/en/event"]	["/r/microsoft/IsA","/r/microsoft/IsA"]	0.0134870407741 88053

"nodes"	"rels"	"probs"
["/c/en/win","/c/en/game","/c/en/event"]	["/r/microsoft/IsA","/r/microsoft/IsA"]	0.0000045600002934 78373
["/c/en/win","/c/en/contest","/c/en/event"]	["/r/microsoft/IsA","/r/microsoft/IsA"]	0.0000039000000079441811
["/c/en/win","/c/en/program","/c/en/event"]	["/r/microsoft/IsA","/r/microsoft/IsA"]	0.0000011400000733695933



Link prediction in HIN using meta-path feature:

2.auto generate meta-path by random walk

- Random-walk constrain relations and weight
- <s,⋯,_,⋯,t>,r,beam_size=25
- MATCH p=(a:Concept {conceptId:s})-[*1..1]-(b:Concept {conceptId:_}) RETURN extract(x IN nodes(p)|x.conceptId) as nodes,extract(x IN relationships(p)|type(x)) as rels,reduce(prob = 1.0,x IN relationships(p)|prob*x.weight) as probs ORDER BY probs DESC LIMIT 25;

Complex demo : multi-reference dep-tree to dep-graph

- An early-morning landslide buries 40 homes and leaves 15 people dead and 114 others missing in Aba Prefecture, Sichuan Province, China. At least 500 rescue workers are on scene, and a 2-km stretch of the river in Mao County is blocked.
- China : Death toll rises to 15 and over 120 others are missing after a landslide in south-western Sichuan province .
- Rescue operation is under way in Sichuan province after more than 40 homes in Xinmo village were engulfed by lan.
- More than 120 people are missing after a landslide in Sichuan province in south-western China, 40 homes were destroyed in Xinmo village.
- Fifteen people were killed in a landslide in southwest China 's Sichuan Province on Saturday and about 100 were.
- Dozens of homes are destroyed and at least 120 people are missing in the wake of a massive landslide in China .

四、研究成果

- 新闻事件检测系统
- 新闻事件数据集

五、学位论文框架

- 第一章 引言
  - 1.1 研究背景和意义
  - 1.2 国内外研究现状与进展
  - 1.3 本文的研究方法及结构安排
  - 1.4 本文的主要创新点与贡献
- 第二章 基础理论介绍
  - 2.1 引言
  - 2.2 文本表示方法
  - 2.3 基于次模函数的文本摘要方法
  - 2.4 概念图谱与异质信息网络
  - 2.5 本章小结
- 第三章 新闻事件检测系统
  - 3.1 推文表示
  - 3.2 推文过滤与分类
  - 3.3 推文聚类与信息抽取
  - 3.4 本章小结

五、学位论文框架

- 第四章 新闻事件数据集构建
  - 4.1 相关推文检索
  - 4.2 相关推文过滤与排序
  - 4.3 本章小结
- 第五章 新闻事件概念图谱构建
  - 5.1 概念异质信息网络
  - 5.2 推文概念化
  - 5.3 事件模式抽取
  - 5.4 事件概念抽取
  - 5.5 本章小结
- 第六章 总结与展望
  - 6.1 总结
  - 6.2 展望
- 参考文献