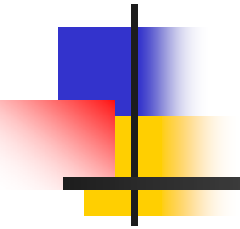
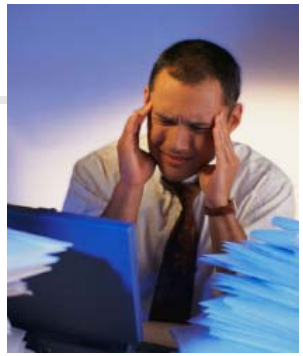
信息提取和问答式检索

赵军
jzhao@nlpr.ia.ac.cn
中国科学院自动化研究所
模式识别国家重点实验室
2007-5-31

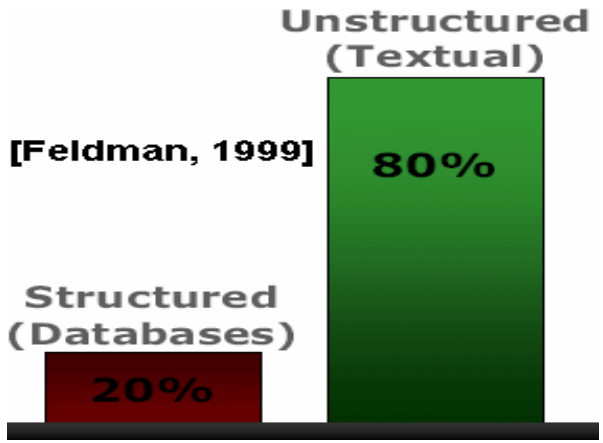


信息提取

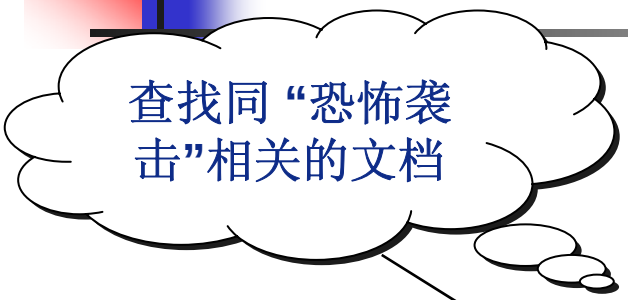
研究背景



- 互联网迅速普及和发展
  - 信息资源极大丰富
  - 但“信息过载”问题日趋严重
- 迫切需要快速、准确获取信息的技术手段
  - 信息抽取技术应运而生
    - 信息抽取—>文本信息抽取
    - 自然语言文本信息抽取<——

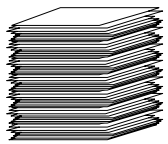


文本信息检索可以做什么



文本信息检索

www.google.com



互联网文档集

伦敦遭遇恐怖袭击_新闻中心_新浪网
伦敦遭遇恐怖袭击. ...已经公开发布声明，就是他们干的 袭击手法相似，应该就是 基地组织所为 其他恐怖团伙假冒基地组织之名 ...英国媒体披露伦敦7-7恐怖袭击与基地 无关 (0409 19:21); 英国议会就伦敦爆炸案为情报部门开脱责任 (0331 01:16) ...
news.sina.com.cn/z/ldbz05/index.shtml - 148k - 网页快照 - 类似网页

“9-11”四周年_新闻中心_新浪网
9-11” 袭击之后，越来越多的恐怖袭击事件都是打着基地的名义去干的，而这些组织未必都 受基地领导或指示...[全文] ...全球蔓延的恐怖袭击，无论伦敦还是埃及，在爆炸附近都 有中国人的身影。中外反恐专家一致认为恐怖袭击离中国并不遥远...[全文] ...
news.sina.com.cn/z/911sznj/index.shtml - 111k - 补充材料 - 网页快照 - 类似网页
[news.sina.com.cn站内的其它相关信息]

纽约遭受恐怖袭击
涉嫌参与九一一恐怖袭击的穆萨维获准为自己辩护(6月14日10:09); 华裔银行捐款纽约将 建"九一一受难华人纪念碑"(5月3日06:05); 美国承认至今未 ...组图:美国反恐战争中的 重要武器(10月8日04:28); 新闻资料：纽约世贸中心遭恐怖袭击 “之最” (9月20日10:11) ...
www.chinanews.com.cn/zhuanti/bao/index.html - 47k - 网页快照 - 类似网页

美国遭遇恐怖袭击网络IT行业受到冲击_科技时代_新浪网
有人怀疑拉丹领导下的组织正使用生化武器向美国发起恐怖袭击。对此，国际红十字会的 发言人表示，阿富汗的生物实验室“不太可能”被恐怖分子改造成制造炭疽热细菌的基地， 当然，也并非“绝不可能”。>>详细内容 [发表评论] 微软75名员工收到炭疽热病菌信件2人 ...
tech.sina.com.cn/focus/kongbu.shtml - 27k - 网页快照 - 类似网页

恐怖袭击笼罩美国_中国网
美国东部时间9月11日上午（北京时间9月11日晚上），美国纽约和华盛顿及其他一些城市 相继遭到恐怖袭击。恐怖分子劫持客机并撞向美国的经济、军事和政治中心。纽约世界贸易 中心倒塌，五角大楼部分被毁，国务院和国会山也发生爆炸。美国总统发表讲话，表示 ...

文本信息检索不能做什么

根据恐怖袭击相关的文档列出某日发生的恐怖事件

将恐怖袭击事件按照发生地点进行归类

需要文本信息抽取技术做支撑！

信息抽取概念

- 信息抽取定义[Grishman, 1997]
  - 从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。
- 最常见的信息抽取表现形式——模板填充
  - 用户定义：模板或者数据库规则
  - 输入：非结构化自然语言文本
  - 输出：从文本中抽取相关字符串填充模板

信息抽取模板填充示例

据美联社消息，当地时间7月7日清晨，英国伦敦金融中心的地铁发生6次爆炸，其中还包括一辆满载乘客的双层公共汽车。由于事发当时处于上班的高峰时期，造成了大量人员伤亡。据初步统计的数字，多起爆炸已至少造成４５人死亡、约１０００人受伤。

信息抽取模板填充示例

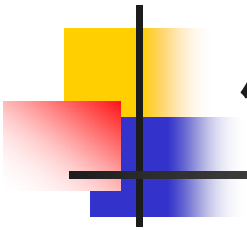
据美联社消息，当地时间7月7日清晨，英国伦敦金融中心的地铁发生6次爆炸，其中还包括一辆满载乘客的双层公共汽车。由于事发当时处于上班的高峰时期，造成了大量人员伤亡。据初步统计的数字，多起爆炸已至少造成４５人死亡、约１０００人受伤。

信息抽取

类型	地点	时间	死亡人数	受伤人数
爆炸	英国伦敦金融中心的地铁	当地时间7月7日清晨	45人	约1000人

信息提取的意义

- 对传统信息检索结果的进一步加工、精化，抽取指定的信息，以用户满意的方式输出，将传统信息检索系统转变为智能化系统，真正实现“所得即所需”。
- IE系统应用前景广泛，实现从非结构化文本到结构化信息的自动化处理，可以和数据库应用系统集成，便于数据管理和信息查询。
- 信息提取是数据挖掘、文本挖掘、自动问答、自动文摘等应用系统的重要基础。



信息抽取历史

- MUC (Message Understanding Conference)
- ACE (Automatic Content Extraction)

MUC （Message Understanding Conference，消息理解会议）

- 由美国国防高级研究计划委员会DARPA资助
- 1987～1997召开七届
- 主要是英文，后两届加了中文。
- 面向新闻信息提取，每届都设定特定的目标场景。

MUC

会议	年份	文本领域
MUC-1	1987	海军军事情报
MUC-2	1989	海军军事情报
MUC-3	1991	恐怖袭击
MUC-4	1992	恐怖袭击
MUC-5	1993	公司合资、微电子芯片制造处理
MUC-6	1995	人事职务变动
MUC-7	1997	飞机失事、航天器发射

MUC的五大评测任务

- 命名实体识别：Named Entity
- 同指关系消解：Co-reference
- 模板元素填充：Template Element
- 模板关系确定：Template Relation
- 场景模板填充：Scenario Template

命名实体识别

- 任务：识别出文本中出现的专有名称和有意义的数量短语并加以归类；
- 一般指的是三大类（实体类、时间类和数字类）、七小类（人名、地名、机构名、时间、日期、货币和百分比）命名实体。在面向新闻领域信息提取时，也涉及武器、交通工具等特殊实体。
- [*name type: person* Sam Schwartz] retired as executive vice president of the famous hot dog manufacturer, [*name type: company* Hupplewhite Inc.]. He will be succeeded by [*name type: person* Harry Himmelfarb].

同指关系消解

- 任务：识别出文本中具有同指关系的名词、名词短语和代词（MUC的定义）。
- Example:
  - *Most computational linguists* prefer *their* own parsers.
  - *Bill Clinton* is *the President of the United States*.

指代和同指的区别

- 指代（Anaphora）：也称照应，是指篇章中的一个语言单位与之前出现的语言单位存在的特殊语义关联，其语义解释依赖于前者。其中，用于指向的语言单位，称为照应语（或称指代语Anaphor），被指向的语言单位称为先行语（Antecedent）。确定照应语所指的先行语的过程称为指代消解（Anaphora Resolution）。
- 同指（Coreference）：如果照应语与先行语都指称（refer to）“现实世界”的同一对象（或实体），则表明两者具有同指关系，此时的指代消解便称为同指消解（Coreference Resolution）。

指代和同指的区别（cont.）

- 指代不一定是共指：
  - 我参观了刘博士的新房，窗户正对着花园，…
- 同指也不一定是指代：
  - 指代是话语结构分析领域的概念，是将独立的句子连接成有意义的、连贯的篇章的一种话语结构形式；
  - 同指只限定两个指称语代表现实世界的同一实体或现象，可以是跨文本的同指，而跨文本之间没有指代问题。

模板元素填充（实体属性的提取）

- 任务：从文本中识别出特定类型实体的属性特征，填充到预先定义的实体属性模板，形成实体对象；
- 例如对人物实体的模板元素抽取，需要信息抽取系统能够抽取出预先定义的人物的名称、职务、国籍等属性。

模板元素填充的评测模式(给出以下信息)

- **BNF DEFINITION.**
  - Will include the Template Element objects and slots defined in the fill rules. Primarily defines the syntax of the objects and slots.
- **FILL RULES.**
  - Will describe the reporting conditions and the semantics of each object and slot. ……
- **EXAMPLE BASE.**
- **A set of texts with accompanying filled-out templates.**

模板元素填充：Example

- **<ENTITY-0592-1>** := ENT_NAME: "BRIDGESTONE SPORTS CO."
- **ENT_TYPE:** ORGANIZATION
- **ENT_DESCRIPTOR:** "JAPANESE SPORTS GOODS MAKER"
- **ENT_CATEGORY:** ORG_CO
- **<LOCATION-0592-1> :=**
- **LOCALE:** "JAPAN"
- **LOCALE_TYPE:** COUNTRY
- **COUNTRY:** JAPAN

模板关系确定（实体关系的抽取）

<**PRODUCT_OF**> := ARTIFACT: <ENTITY>- ORGANIZATION: <ENTITY>- OBJ_STATUS: {OPTIONAL}- COMMENT: " "-	<**EMPLOYEE_OF**> := PERSON: <ENTITY>- ORGANIZATION: <ENTITY>- OBJ_STATUS: {OPTIONAL}- COMMENT: " "-
<**LOCATION_OF**> := LOCATION: <LOCATION>- ORGANIZATION: <ENTITY>- OBJ_STATUS: {OPTIONAL}- COMMENT: " "-	

模板元素填充的评测模式(给出以下信息)

- BNF DEFINITION.
  - Will include the Template Relation objects and slots defined in the fill rules. Primarily defines the syntax of the objects and slots.
- FILL RULES.
  - Will describe the reporting conditions and the semantics of each object and slot…….
- EXAMPLE BASE.
- A set of texts with accompanying filled-out templates (for Template Relations and any necessary Template Elements).

模板关系确定：Example

``Here comes my ride!" Ross shouted as the McDonnell Douglas MD Explorer came into sight.	<**PRODUCT_OF-9601290937-1**> := ARTIFACT: <ENTITY-9601290937-18> ORGANIZATION: <ENTITY-9601290937-6
<**ENTITY-9601290937-6**> := ENT_NAME: "McDonnell Douglas" ENT_TYPE: ORGANIZATION ENT_CATEGORY: ORG_CO >	<**ENTITY-9601290937-18**> := ENT_NAME: "McDonnell Douglas MD Explorer" ENT_TYPE: ARTIFACT ENT_DESCRIPTOR: "SUPERCHOPPER" ENT_CATEGORY: ART_AIR

场景模板填充

- 即事件的提取，针对某一场景中设定的一些槽提取相应的内容进行填充。

场景模板填充

据美联社消息，当地时间7月7日清晨，英国伦敦金融中心的地铁发生6次爆炸，其中还包括一辆满载乘客的双层公共汽车。由于事发当时处于上班的高峰时期，造成了大量人员伤亡。据初步统计的数字，多起爆炸已至少造成４５人死亡、约１０００人受伤。

信息抽取

类型	地点	时间	死亡人数	受伤人数
爆炸	英国伦敦金融中心的地铁	当地时间7月7日清晨	45人	约1000人

MUC的信息提取评测的工作模式

- 用户通过“信息提取模板”给出需要提取的信息；
- 用户根据信息提取模板，生成一些信息提取规则；
- 信息提取系统对文本中的候选句进行命名实体识别等浅层分析；
- 信息提取系统根据信息提取规则从候选句中提取出所需要的信息，填充“信息提取”模板的槽；
- 进行回指分析，确定信息的最终形式。

信息提取的评测

- 两个评价指标：召回率和准确率；
- 综合评价F值：是召回率和准确率（Precision）的加权几何平均值：

$$F = \frac{(1+\beta^2) \times \mathrm{Re}call \times \mathrm{Pr}ecision}{\mathrm{Re}call + \beta^2 \times \mathrm{Pr}ecision}$$

- 其中，beta是召回率和准确率的相对权重。beta等于1时，二者同样重要；beta大于1时，准确率更重要一些；beta小于1时，召回率更重要一些。在MUC系列会议中，beta取值一般为1、1/2、2。

子任务	命名实体	共指	模板元素	模板关系	场景模板	多语言
MUC-3					R<50% P<70%	
MUC-4					F<56%	
MUC-5					EJV F<53% EME F<50%	JJV F<64% JME F<57%
MUC-6	E F<97% C F<85% J F<93% S F<94%	R<63% P<72%	F<80%		F<57%	
MUC-7	E F<94% C F<91% J F<87%	F<62%	F<87%	F<76%	F<51%	
说明	R－召回率；P－精确率；F－系统F值（beta值取1）；JV－合资；ME－微电子；E－英语；C－汉语；J－日语；S－西班牙语。					

从MUC评测看信息提取的技术水平

- 这些指标也自然地反应了自然语言处理在各个层次上的难度。

- 经过七届MUC评测会议，英文系统在指定的命名实体识别方面基本达到实用水平，在受限的实体关系（TE，TR）识别方面也接近实用的水平。但在完整的信息抽取任务（ST）方面，则还有许多问题需要探索，这些问题大部分都涉及到了自然语言处理的核心难题。

MUC的成果

- MUC系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用。
- MUC定义的信息抽取任务的各种规范以及确立的评价体系在一段时间内成为信息抽取研究的事实标准。

MUC存在的问题

- MUC的这种信息提取方式有很大的领域局限性。因为没有统一格式的信息提取模板来描述用户的信息需求，因此根据信息提取模板生成信息提取规则这个过程往往需要人工劳动，信息提取技术无法推广到大规模、实用化的应用中；
- 需要研发面向开放域的、适应性强的信息提取技术：可扩展的信息抽取系统（Portable IE）。

ACE

- Automatic Content Extraction，由美国标准技术研究所NIST组织，从2000年开始到现在已经进行了五次评测，主要面向新闻领域的文本，抽取其中的实体、关系、事件；
- ACE的目标：希望建立鲁棒性的、自适应的信息提取方法。给定一些语言数据，系统就快速构建所需要信息提取系统。

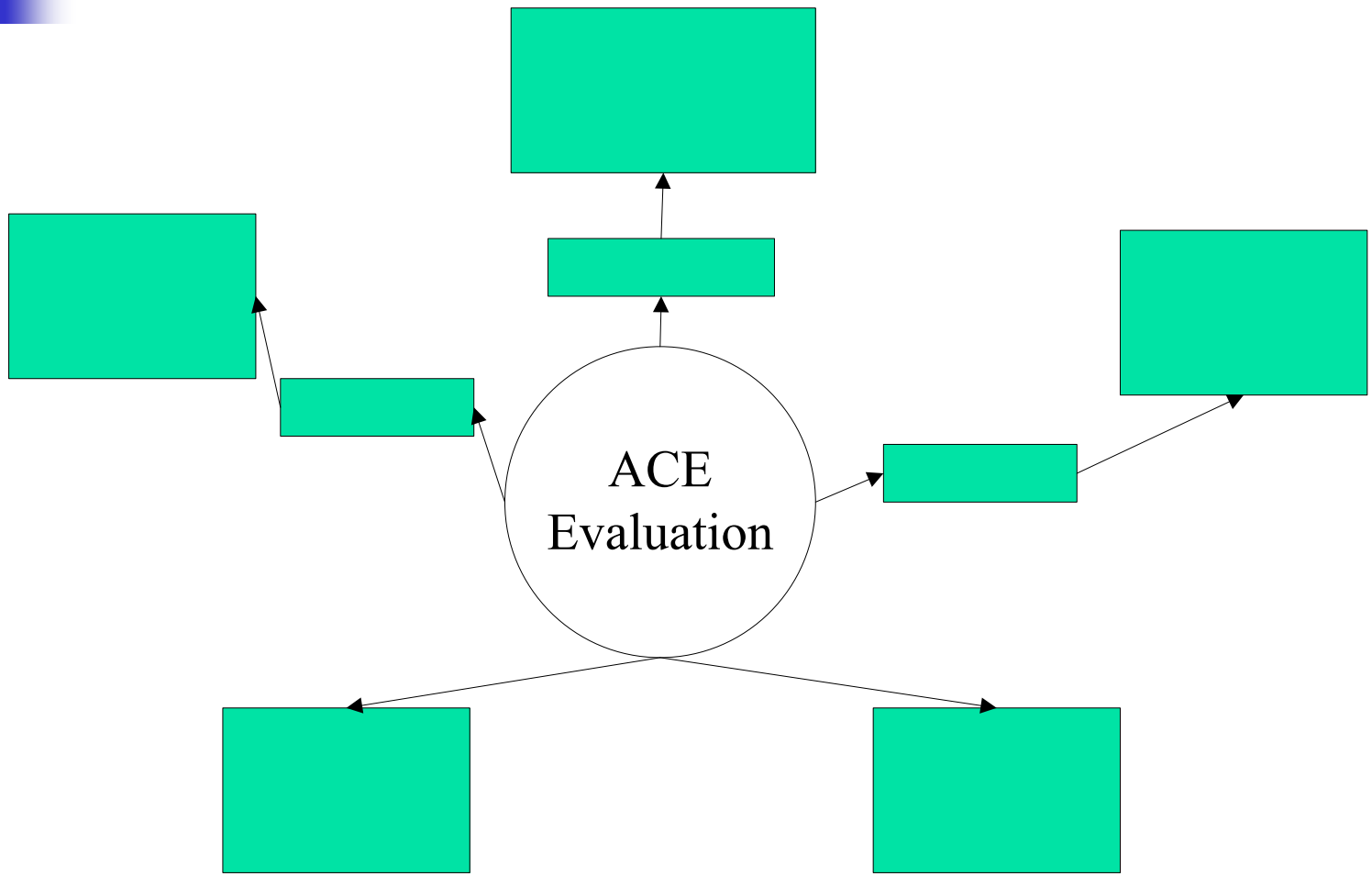
ACE与MUC的不同

- 用户信息提取的需求表示：
  - MUC：用户给定信息提取模板，信息提取系统填充模板的槽；
  - ACE：用户指定要检测的事实或事件的类别，信息提取系统给出检测这文本中这些事实或事件的出现，并进行描述；
- 信息来源：
  - MUC：书面文本；
  - ACE：信息来源更宽，不只局限于书面文本，还包括经过ASR和OCR生成的文本；

ACE相关中的几个基本概念

- 实体：A real-world person, place or thing referred to in the source language
- 数值：a quantity that provides additional information (often to relations and events)
- 时间：a temporal expression (that maybe used as an attribute to relations and events)
- 关系：a relationship between two ACE entities (these two entities are the arguments of an ACE relation)
- 事件：an event involving zero or more ACE entities, values and time expressions.
- 提及：文本中对实体、关系、事件的引用表达式。

ACE任务框架



ACE 2007 任务之一：

实体检测与跟踪

EDT的任务

- Entity Detection and Tracking，EDT；
- 检测出文本中出现的每个实体，识别出这些实体的相关信息（类别、子类、提及），并将指称同一实体的所有提及连接起来；
- 提及包括命名性的、名词性的、代词性的三种类型。同时，标注出每个提及的实体类型、子类、提及类型；
  - 命名性提及：Joe Smith;
  - 名词性提及：the guy wearing a blue shirt
  - 代词性提及：he, him
- 对于每一次提及，要求识别出表示实体的最长串；
  - （代表日本最新技术的大江户线新地铁系统），本周在（东京）投入运营。
- 子任务：EMD（Entity Mention Detection），测试系统正确识别实体提及的能力。
- 阿拉伯文、中文、英文、西班牙文

ACE-2007的EDT任务检测的7大类实体

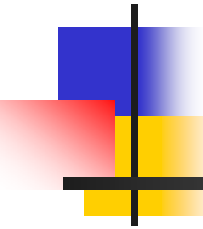
类别	子类
Person	Group, Indeterminate, Individual
Organization	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
Location	Address , Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
Geo-Political Entity	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
Facility	Airport, Building-Grounds, Path, Plant, Subarea-Facility
Vehicle	Air, Land, Subarea-Vehicle, Water, Underspecified
Weapon	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspeciified

实体检测与跟踪（续）

- Person:
  - *“John Smith”, “the butcher”, “dad”, “he”, “the family”, “the house painters”, “the linguists under the table”*
- Organization:
  - *“The Salzburg prosecutor’s office”, “the court”, “Tech Source Marine Industries in State College, PA.”, “Pennsylvania State University”, “The Red Cross”*
- GPE:
  - 一个以政治为出发点定义的地理区域，它可以指物理区域本身、也可以指该区域的政府、人民等。因为很难区分这些概念，所以单独列为一类实体。最常见的GPE是国家、省、城市等。
  - *France has an area of xxx million square kilometers.*
  - *France signed a treaty with Germany last week.*
  - *France elected a new president.*
  - *France has a gross domestic product of xxx francs.*

实体检测与跟踪（续）

- Locations:
  - 以地理或天文为出发点定义的区域，并且不组成一个政治实体的；
  - “Capital Hill”, “borders shared by Turkey, Azerbaijan, and Georgia”, “the Sun”, “the World”, “the Missouri River”, “the Southern Causasus”, “another part of the city”, “the tourist region of Kaprun”, “most places”, etc.
- Facilities:
  - 大型的、具有某种功能的人造结构，包括建筑物或类似的设施，供人居住或使用（例如：房屋、工厂、体育馆、办公室、建筑物、体育场、监狱、博物馆、空间站、仓库、加油站、停车场、街道、机场、火车站、桥梁、隧道等）。
  - *The museum* is located on Fifth Avenue.



ACE 2007 任务之二：

数值检测和识别

VAL的任务

- Value Detection and Recognition: VAL
- Task: The certain specified types of values that are mentioned in the source language data be detected and the selected information about these values be recognized and merged into a unified representation for each detected value.
- 2007年没有单位报名参加

Type	Subtype
Contact-Info	Email, Phone-number, URL
Numeric	Money, Percent

ACE 2007 任务之三：

时间识别和规范化

TERD的任务

- Time Expression Recognition and Normalization, TERD
- Task: TERD requires that certain temporal expressions mentioned in the source language data be detected and recognized in timex2 format. （时间短语的识别和规范化）
- 中文、英文、西班牙文

TIMEX2 规范

- DARPA的TIDES（Translingual Information Detection，Extraction and Summarization）计划和ACE计划制定的时间短语标注方案

- 对时间短语进行标注，分为两步：第一步是标记出时间短语（marking the Extent），也就是找到时间短语出现的开始位置和结束位置；第二步把时间短语所表示的具体时间值计算出来，或者叫做时间短语规格化（normalizing the value）。

TIMEX2 规范（cont.）

- XML结构

  <TIMEX2 attribute-list> 时间词 </TIMEX2>

- 描述4种时间类型

  点时间，段时间，模糊时间，集合时间

- 7个属性

  VAL、ANCHOR_DIR、ANCHOR_VAL、MOD、SET、NON_SPECIFIC、COMMENT

属性含义

- VAL：表示时间轴上的一个点或时间段的长度
- ANCHOR_VAL：表示锚定的时间点
- ANCHOR_DIR：表示相对于锚点的方向
- MOD：表示对时间词的修饰
- SET：集合时间标志
- NON_SPECIFIC：非特指时间标志
- COMMENT：注释

- **点时间**

  <TIMEX2 VAL=2003-8-21T11:4:7> 2003年8月21日11时04分07秒 </TIMEX2>……

- **段时间**

  <TIMEX2 VAL=P3D ANCHOR_DIR= ENDING ANCHOR_VAL =2006-06-01>过去三天</TIMEX2>阴雨绵绵

- **模糊时间**

  <TIMEX2 VAL=FUTURN _REF ANCHOR_DIR=AFTER ANCHOR_VAL=1999-7-15>近期</TIMEX2>将有一系列政策出台

- **集合时间**

  <TIMEX2 VAL=XXXX-XX-XX SET=YES>每天</TIMEX2>坚持锻炼

ACE 2007 任务之四：

关系检测与描述

RDC的任务

- Relation Detection and Characterization, RDC
- RDC的目标是检测和描述EDT输出的实体之间限定类型的关系、识别出这些关系的类别、子类等信息，并把指称同一关系的关系提及连接起来。
- 实体间的关系类型多种多样。ACE-2007的RDC测试了七大类关系检测。
- 子任务：RMD（Relation Mention Detection），用于测量系统正确识别ACE关系提及的能力。
- 英文、中文

ACE-2007的RDC任务测试的七大类关系

Type	Subtype
Agent-Artifact	User-Owner-Inventor-Manufacturer
GPE Affiliation	Citizen-Resident-Religion-Ethnicity, Org-Location
Org-affiliation	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
Part-whole	Artifact, Geographical, Subsidiary
Person-social	Business, Family, Lasting-Personal
Physical	Located, Near

关系的论元

- 每个关系都有两个论元Arg-1和Arg-2，此外，还可能有一个或多个timex2论元；
- 关系识别需要识别出每个关系的每个论元，这些论元的类型，包括：

Arg-1	Arg-2
Time-After	Time-Before
Time-At-beginning	Time-At-End
Time-Starting	Time-Ending
Time-Holds	Time-Within

关系的定义：Located/Near

- **The Located relation captures the exact location of an entity.**
  - *Ex: A military base in Germany*
  - *[Located("a military base in Germany","Germany")]*
- Near indicates that an entity is explicitly near another entity, but not actually in that location or part of that location
  - *Ex: A town some 50 miles south of Salzburg in the central Austrian Alps*
  - *[Near("a town some 50 miles south of Salzburg in the central Austrian Alps","Salzburg")]*

关系的定义：Part-whole

- **Part-Whole characterizes physical relationships between entities and their parts.**
  - *Ex: A state within the former Soviet Union*
  - *[Part-Whole("a state within the former Soviet Union","the former Soviet Union")]*
  - *Ex: The top of the mountain*
  - *[Part-Whole("the top of the mountain","the mountain")]*

关系的定义： Business/Family

- **Business captures the connection between two entities in any professional relationship, including boss-employee, lawyer-client, co-workers, political relationships, etc.**
  - ***Ex: their colleagues***
  - ***[Business("their", "their collegues")]***
- **Family captures the relation between an entity and another entity with which it is in any family position.**
  - ***Relatives of the student***
  - ***[Family("relatives of the student", "the student")]***

关系的定义： Employment

- **The relations between the organizations and persons who fill general staff positions within them.**
- *Ex: the CEO of Microsoft*
  - *[Employment("the CEo of Microsoft","Microsoft")]*
- *Ex: Mr. Smith, a senior programmer at Microsoft…*
  - *[Employment("a semior programmer at Microsoft","Microsoft")]*

关系的定义： Citizen/Resident

- Citizen/Resident describes the relation between a person and the GPE in which they have citizenship or in which they live.
- Ex: U.S. businessman Edmond Pope
  - [Citizen(“U.S. businessman Edmond Pope”,”U.S.”)]

关系提及

- 关系提及：是表达关系的一个句子、短语或用法。它必须包括该关系关联的两个实体的提及。
- 以下给出一些常用的关系提及的句法类型。
- RDC的任务：识别出这些提及，分析它的结构，找出关系的论元。

表达实体关系的句法类型:所有格

- 所有格：Possessives indicates the syntactic structure where the first noun or pronoun is in the possessive case.
- Nathan Myhrvold, Microsoft's chief scientist.
  - [Employment("Microsoft's chief scientist","Microsoft")]

表达实体关系的句法类型：介词

- Preposition indicates a taggable relation between a head noun and a prepositional phrase that modifies it.
- Ex: Officials in California are warning residents.
  - [Located("Officials in California","California")]
- Ex: The CEO of Microsoft
  - [Employment("The CEO of Microsoft","Microsoft")]

表达实体关系的句法类型：前置修饰语

- 前置修饰语（a proper adjective or proper noun premodifier）和它修饰的中心名词之间构成实体关系。
- Ex. [[[英国外交大臣] [库克]] 的发言人]
  - [Business (“英国外交大臣库克的发言人”)，”英国外交大臣库克“]
  - [Employment (“英国外交大臣”)，”英国 “]

表达实体关系的句法类型：习惯用法

- 由一些套话描述的实体关系：
  - 新华社北京电
  - [Located（”新华社“，”北京“）]

表达实体关系的句法类型：静态动词

- 青海位于青藏高原上
  - [Located（”青海“，”青藏高原“）]

表达实体关系的句法类型：分词形式

- the crowd trapped inside the compartment…
  - [Located(“the crowd trapped inside the compartment”,”the compartment”)]

ACE 2007 任务之五：

事件检测与描述

VDC的任务

- Event Detection and Characterization, VDC;
- VDC的目标是检测和描述限定类型的事件。对于每一个事件，将提取出的相关信息融合为一个统一的表示。
- 事件的类型多种多样。ACE-2004的VDC测试了5种类型的事件检测：Destruction/Damage，Creation/Improvement，Transfer of Possession or Control，Movement，Interaction of Agents；
- 事件的参与者是ACE的实体，且每个参与者都充当事件的角色。这些角色包括以下类型：施事、受事、地点、时间、来源（出发地）、目标（目的地）、其他等七种类型。

ACE-2004测试的五种事件类型

- 以受事为中心的事件：
  - Destruction/Damage：中心实体被破坏的事件；
    - 10月12号驾驶小艇[{炸毁} (美国神盾级驱逐舰科尔号)] 的两名男子的身份已经被证实。
  - Creation/Improvement：中心实体被制造或改进的事件；
    - [(代表日本最新技术的大江户线新地铁系统)，本周开始在东经{投入} 运营]。
  - Transfer of Possession or Control：中心实体的所有权被转移的事件；
    - [新加坡晋江会馆会务顾问（蔡锦淞）{受推选}为第三届会长]。
    - [截至11月底，维和部队共{逮捕}了（30名从“安全区”）[{进入}]科索沃]的阿族恐怖分子）。]

ACE-2004测试的五种事件类型（续）

- Movement：中心实体的所有权被移动的事件；
  - [（高行健）5号{飞抵}瑞典首都斯德哥尔摩机场]

- 以施事为中心的事件：
  - Interaction of Agents：An event is classified as INT when the salient entities are agents engaged in interaction.
    - [(大批阿族极端分子)在科索沃同塞尔维亚本土交界的“安全区”{袭击}塞族警察并[造成(大量人员){伤亡}]$_{BRK}$]$_{INT}$
    - [(南共体和欧盟国家的外长)在哈博罗内{举行}了两天会议。]
    - [(俄潜水员)目前正在巴伦支海和潜艇失事地区进一步{考察}艇身和海底。]

事件的论元角色的标注

- 事件的角色包括以下的类型，只标注出由实体充当的并且在事件的局部上下文环境中出现的角色。
  - 施事：
    - [(南共体和欧盟国家的外长)在哈博罗内{**举行**}了两天会议。]
  - 受事：
    - [也门当局已经{**逮捕**}了（数十名涉嫌与美国驱逐舰“科尔号”在亚丁湾被炸事件中有关的人士）]
  - 出发地：
    - [波普已经{**离开**}（监狱）][从莫斯科机场搭机{**返回**}美国]
  - 目的地：
    - [波普已经{**离开**}监狱][从莫斯科机场搭机{**返回**}（美国）]

事件的论元角色的标注（续）

- 时间：
  - [(多普)是在{半年多之前}因为被怀疑指控收集情报，从事间谍活动而被俄国联邦安全局{拘捕}的]
- 地点：
  - [(俄潜水员)目前正在（巴伦支海和潜艇失事地区）进一步{考察}艇身和海底。]
- 其他：例如工具、目的等。
  - [(香港华侨华人总会”西部大开发“经济访问团)今天乘（飞机）{抵达}贵阳]

事件的提及

- 对于每个事件，要标注出该事件的所有提及。事件提及的类型可以是名词性的或者是句子性的。
  - 名词性的事件提及：
    - [以巴长达20天的{**冲突**}]
    - [与多普的{**会晤**}]
    - [双方的{**接触**}]
  - 句子性的事件提及：
    - [也门当局已经{**逮捕**}了（数十名涉嫌与美国驱逐舰“科尔号”在亚丁湾被炸事件中有关的人士）]
    - [(波普)已经{**离开**}监狱][从莫斯科机场搭机{**返回**}美国]

ACE的评测的技术水平（Chinese）

- **EDR**

  Precision=69.1% Recall=70.5% F-measure=69.8%

- **EMD**

  Precision=89.9% Recall=85.1% F-measure=87.4%

- **RDR (relation detect & recognize)**

  Precision=54.1 % Recall=30.7 % F-measure=39.1 %

- **RMD (relation mention detect)**

  Precision=76.3 % Recall=40.6 % F-measure=53.0 %

- **TERN (Time Entity Recognize & Normalize)**

  Precision=50.8% Recall=46.0% F-measure=48.3%

ACE的评测的技术水平（English）

- **EDR task**
  - Precision=68.2% Recall=67.6% F-measure=67.9%
- **EMD task**
  - Precision=90.8% Recall=86.0% F-measure=88.3%
- **RDR task(relation detect & recognize)**
  - Precision=54.5 % Recall=28.1 % F-measure=37.1 %
- **RMD task(relation metion detect)**
  - Precision=81.6 % Recall=38.7 % F-measure=52.5%
- **TERN (Time Entity Recognize & Normalize)**
  - Precision=72.6% Recall=69.6% F-measure=71.1%
- **VDR(Event Detect & Recognize)**
  - Precision=54.5% Recall=28.1% F-measure=37.1%
- **VMD(Event Mention Detect)**
  - Precision=81.6% Recall=38.7% F-measure=52.5%

MUC和ACE的比较

- 用户信息提取的需求表示：
  - MUC：用户给定信息提取模板，信息提取系统填充模板的槽；
  - ACE：用户指定要检测的事实或事件的类别，信息提取系统给出检测这文本中这些事实或事件的出现，并进行描述；
- 信息来源：
  - MUC：书面文本；
  - ACE：信息来源更宽，不只局限于书面文本，还包括经过ASR和OCR生成的文本；
- 信息提取的方式：
  - MUC：根据用户给定的信息提取模板生成信息提取规则，然后根据信息提取规则从文本中提取信息。从信息提取模板生成信息提取规则往往需要人的干预。因此适合于受限领域的信息提取；
  - ACE：用户指定要检测的事实或事件的类别，信息提取系统给出检测这文本中这些事实或事件的出现。希望建立鲁棒性的、自适应的信息提取方法。给定一些语言数据，系统就快速构建所需要信息提取系统。

从MUC和ACE的对比看信息提取的发展方向

- 与MUC相比，ACE旨在定义一种通用的信息抽取标准，不再限定领域和场景，而是从语义的角度制订一套更为系统的信息抽取框架，这个框架将信息抽取归结为建立在一定本体论基础上的实体、关系、事件的抽取，从而适用于更广泛的领域和不同类型的文本。
- 开放域信息提取技术

开放域信息提取技术

- 难点：
  - 用户对于事实或事件的提取多种多样，描述各不相同。为了适应不同的用户信息提取需求，这些技术必须是面向开放域的、可以自适应的。
- Solutions：
  - 提出通用的实体体系、关系体系和事件体系，针对通用的体系，研究开发实体识别、关系识别和事件识别技术。
  - 根据不同用户的信息提取需求，将他所需要提取的实体、事实和事件映射到通用的实体类别、关系类别和事件类别。如果是领域特定的一些实体、关系和事件，在通用的体系中找不到对应的化，则这部分领域特定的部分需要重新开发。例如通讯领域的产品名识别等等。

开放域信息提取技术中的主要任务（一）

- 实体识别：
  - 通用的实体体系的建立：除了人名、地名、机构名等常规命名实体外，还有一些实体对于信息提取也非常有用。
  - 命名实体识别技术：
    - 实体简称、别称等的识别；
  - 其他实体识别技术：
    - 名词组块识别；

开放域信息提取技术中的主要任务（二）

- 实体关系识别：
  - 实体关系体系的建立；
  - 实体关系识别技术：
    - 理论上，实体关系识别依赖于句子的句法语义依存分析技术；
    - 在没有句法语义分析技术之前，一些简单的模式匹配技术可以解决一些问题。

开放域信息提取技术中的主要任务（三）

- 事件识别：
  - 事件体系的建立；
  - 事件识别技术：
    - 理论上，实体关系识别依赖于句子的句法语义依存分析技术（特别是语义角色标注技术）；
    - 在没有句法语义分析技术之前，一些简单的模式匹配技术可以解决一些问题。

开放域信息提取技术中的主要任务（四）

- 同指消解：
  - 很难的问题；
  - 在提出汉语回指问题的解决方案之前，需要在大规模语料库中进行回指语言现象的定量调查。

开放域信息提取技术中的主要任务（五）

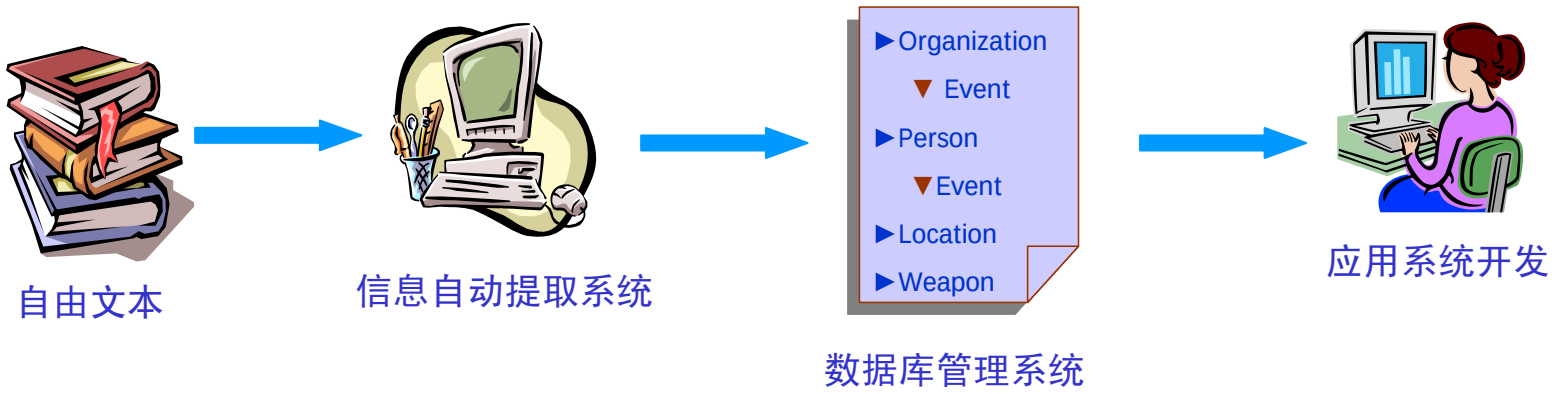
- 信息提取技术的自适应：
  - 根据不同用户的信息提取需求，将他所需要提取的实体、事实和事件映射到通用的实体类别、关系类别和事件类别；
  - 如果是领域特定的一些实体、关系和事件，在通用的体系中找不到对应的话，则这部分领域特定的部分需要重新开发。例如通讯领域的产品名识别等等。

信息提取用到的自然语言处理技术

- 命名实体的识别和标注；
- 基本组块的识别和标注；
- 基于框架理论的汉语动词词典建设；
- 基于框架理论的汉语句法语义分析器；
- ……

商务领域中文信息抽取

- 目标：面向商务领域文本，抽取指定类型的实体以及其间关系，进而抽取新品上市、产品价格上涨下跌等等商务事件，建立数据库供用户进行快速信息查询，并服务于单文档或多文档文摘应用系统。



任务一：实体的抽取

- 抽取实体类型限定为产品名、机构名、人名、地名、数量短语、时间短语等
- 产品属性的提取

任务三：实体关系的抽取

- 抽取产品和其他实体之间的关系
  - **Product_of**（产品，公司）
  - **Price_of**（数量短语，产品）
  - **Located**（公司，地名）

实体、实体属性、实体关系抽取的实例：人名档案

Name:	威廉·亨利·盖茨
EnglishName:	**William Henry Gates III**
Aliases:	比尔·盖茨
Gender:	男
Birthday:	**1955.10.28**
BirthPlace:	美国西雅图
Email:	**billgates@microsoft.com**
Position:	微软公司董事长
Education：	哈佛大学肄业
Event-involved:	退学 创办微软公司
Snippet:	威廉·亨利·盖茨 (William Henry Gates III)昵称比尔·盖茨 (Bill Gates) １９５５年１０月２８日出生……

任务四：实体链接

- 包括两个子任务，一是将两个具有共指关系的实体进行合并，一是将具有共指关系实体的各个属性进行合并，链接为统一的一个实体表达

任务五：事件抽取和合并

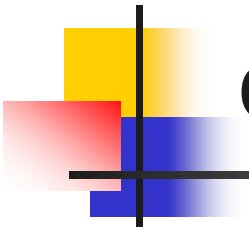
- 针对商务领域有代表性的场景，设计场景模板，并进行填充，实现事件的抽取和合并

Event:	股票跌
Share_Info:	朗讯公司股票
Location_Info:	纽约股市
Price_Past:	
Price_Cur:	4.31美元
Fall Rate:	9.8%
Date_Info:	2002年3月19日
Reason:	
Snippet:	朗讯公司……3月19日该股票创历史最低水平……朗讯的股票下挫了9.8%，以4.31美元收盘……朗讯股票是纽约股市中较为活跃的一支股票……

Event:	产品价格下调
Product_Info:	奇瑞·风云基本型
Location_Info:	全国
Price_Past:	8.79万元
Price_Cur:	7.8万元
Fall Rate:	11%
Date_Info:	2003年3月24日
Snippet:	…奇瑞·风云…全线下调产品价格。其中基本型由原来的8.79万元调至7.8万元……

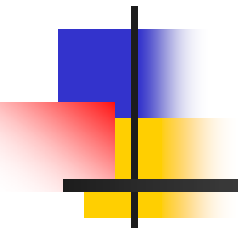
信息提取的发展方向

- 文本信息提取：
  - 开放域信息提取技术；
  - 主观性信息提取技术
- Web信息提取；
- 多媒体信息提取；
- 信息集成



Questions

- 信息提取的应用在哪里？



问答式检索

研究背景

- 信息的两大特点
  - 海量
  - 有效利用率低
- 传统信息检索系统存在着先天性不足
  - 无法通过几个关键词来表达用户的意图
  - 基于关键词匹配的检索方法，其结果不如意
  - 返回的相关信息太多，或者说相关的无用的信息太多
  - 基本上是一种文档检索
- 问答式检索系统（Question Answering System）

问答系统的定义

- Question Answering (QA) is an interactive human computer process that encompasses understanding a user information need, typically expresses in a natural language query; retrieving relevant documents, data, or knowledge from selected sources; extracting, qualifying and prioritizing available answers from these sources; and presenting and explaining responses on an effective manner;

  ----------by Mark Maybur

- 定义
  - 输入：自然语言的问句，非关键词的组合

    谁获得1987年的诺贝尔文学奖？
  - 输出：直接答案，非文档集合

    约瑟夫·布罗茨基

QA在自然语言理解中的作用

- 美国认知心理学家G.M.Ulson认为，判别计算机理解自然语言的4个标准是：QA，Summarization, Paraphrase, MT. 计算机只要达到以上标准之一，就认为它理解了自然语言。
- 自然语言理解：词语层面、句子层面、篇章层面、篇章之间、语言之间的基本问题，在QA中都会出现。另一方面，QA和信息检索密切相关，信息检索中的基本问题在QA中也同样存在。
- QA和MT一样，是NLU研究人员永远追求的一个目标，它的研究会带动NLU的发展。
- 当然，QA研究本身也有重要的应用价值。

问答系统的分类

- 答案源
  - 基于大规模语料库的问答系统；
  - 基于网络的问答系统；
  - 基于知识库的问答系统；
  - FAQ
  - 单文本问答系统（阅读理解系统）
- 领域

  开放域问答系统，限定领域问答系统
- 答案

  词、短语、句子、段落、文摘

各类问答系统的特点

- **基于大规模真实文本固定语料库问答系统**：从预先建立的大规模真实文本语料库中查找答案的。这类问答系统的缺点是无法涵盖用户所有类型提问的答案，却能够提供一个优良的算法评测平台，适合我们对不同问答技术的比较研究；
- **基于网络的问答系统**：从互联网中查找提问的答案。虽然它是在真实环境下研发的问答技术，却不适合评价各种问答技术的优劣；
- **单文本问答系统**：也称之为阅读理解式的问答系统，它是从一篇给定的文章中查找答案。系统在“阅读”完一篇文章后，根据对文章的“理解”给出用户提问的答案。这种系统非常类似于我们在学习英语时做的阅读理解；

各类问答系统的特点（cont.）

- **基于结构数据库的问答系统**：从一个预先建立的结构化的数据库中查找提问的答案。所以该系统可以具有较强推理能力，但建立大规模的结构知识库是一个非常困难的问题。
- **基于常用提问集的问答系统**：在已有的“提问—答案”对集合中找到与用户提问相匹配的提问，并将其对应的答案返回给用户。
- **限定领域问答系统**：用户提问只能限定在某一特定领域。
- 目前的研究主要集中在“基于大规模真实文本固定语料库的开放域问答系统”方面，而实用系统的开发主要是“基于网络的问答系统”和“基于常问提问集的问答系统”上。

问答式检索系统

- 英文问答系统

  AskJeeves: www.ask.com

  AnswerBus: www.answerbus.com/index.shtml

  START: www.ai.mit.edu/projects/infolab

  EasyAsk：www.easyask.com

  AnswerLogic：www.answerlogic.com

  AnswerFriend：www.answerfriend.com

- 中文问答系统
  - 爱问知识人：http://iask.sina.com.cn/

- http://www.faqs.org/faqs/

问答式检索评测

- 目前的评测都是面向大规模文本库的开放域问答式检索
- 评测现状
  - 英文：TREC （Text Retrieval Conference）
    - NIST, DARPA,
  - 日文： NTCIR (NII-NACSIS Test Collection for IR Systems) Project
    - evaluation workshops designed to enhance research in Information Access (IA) technologies including information retrieval, question answering, text summarization, extraction, etc
    - Japan Society for Promotion of Science (JSPS) 和National Center for Science Information Systems (NACSIS)联合实施

问答式检索评测 (cont.)

- 欧洲： CLEF (Cross-Language Evaluation Forum)
  - Sixth Framework Programme of the European Commission
  - 欧洲语言，单语言，跨语言
- 中文：到目前为止，还没有一个大范围的、公认的评测平台

TREC-英语问答评测QA

- 1999年TREC-8到2005年TREC-14，QA评测进行了7届。
- 文本库：The AQUAINT collection consists of 1,033,461 documents taken from the New York Times, the Associated Press, and the Xinhua News Agency newswires.
- 问题集：301个，其中230个factoid questions, 56个list questions, 65个Other questions.

TREC—QA Track评测任务

- TREC QA Track评测任务在不断地变化，大致包括以下几类：
  - Factoid任务：测试系统对基于事实、有简短答案的提问的处理能力。例如，Where is Belize located? 而那些需要总结、概括的提问不在测试之列。例如，如何办理出国手续？如何赚钱？等。
  - List任务：要求系统列出满足条件的几个答案。在TREC2003之前，任务要求被测试系统给出不少于给定数目的实例，如：Name 22 cities that have a subway system。TREC2003要求系统要给出满足条件的尽可能多实例，如：List the names of chewing gums。
  - Definition任务：要求系统给出某个概念，术语或现象的定义、解释。例如：What is Iqra?等。

TREC—QA Track评测任务（cont.）

- Context任务：测试系统对相关联的系列提问的处理能力，即对提问i的回答还依赖对提问j（i>j）的理解。例如：a、佛罗伦萨的哪家博物馆在1993年遭到炸弹的摧毁？b、这次爆炸发生在那一天？c、有多少人在这次爆炸中受伤？
- Passage任务：TREC2003提出的新任务。和其他任务不同的是，它对答案的要求偏低，不需要系统给出精确答案，只要给出包含答案的一个字符序列(a small chunk of text that contains an answer)。

TREC-QA评测（评测指标）

- TREC QA Track的评测指标主要有平均排序倒数（Mean Reciprocal Rank，简称MRR）、准确率（Accuracy）、CWS（Confidence Weighted Score）等
- MRR：

$$MRR = \sum_{i=1}^{N} 1 / \text{标准答案在系统给出的排序结果中的位置}$$

  - 如果标准答案存在于系统给出的排序结果中的多个位置，以排序最高的位置计算；如果标准答案不在系统给出的排序结果中，本题得0分。N表示测试集中的提问个数。
- CWS：

$$CWS = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{前}i\text{个提问中被正确回答的提问数}}{i}$$

  - N表示测试集中的提问个数。

NTCIR-日文问答评测

- Question Answering Challenge（QAC）
- QAC-1 (2002), QAC-2 (2004), QAC-3 (2005)
- 三个子任务：
  - 任务1：每个提问，系统给出五个按概率大小排列的答案列表；采用MRR打分标准；系统必须给出支持每个答案的文档。
  - 任务2：每个提问，系统给出一个答案；如果某个提问在语料中有几个答案，系统须给出所有答案，且必须给出支持每个答案的文档。
  - 任务3：这个任务评测系统对关联提问的处理能力；关联提问是指提问之间可能有互指关系、省略等，类似TREC中的Context Task；系统必须给出支持每个答案的文档。

NTCIR-日文问答评测测试集（cont.）

- 文本库：
  - QAC1：Mainichi Newspaper（1998~1999）；
  - QAC2：Mainichi Newspaper和Yomiuri Newspaper（1998~1999）；
- 提问数：
  - QAC1：任务1—200个；任务2—200个；任务3—40个；
  - QAC2：任务1—200个；任务2—200个；任务3—200个；

CLEF—欧洲多语言问答系统评测

- 2003,2004,2005,2006年设立多语言问答系统评测专项；
- CLEF QA Track定义了单语和多语两个任务：
  - 单语言任务：指输入提问是某种语言，输出的答案就是这种语言。
    - 2003年：Dutch，Italian，Spanish；
    - 2004年：Dutch，French，German，Italian，Spanish ；
  - 多语言任务：指输入提问可以是任何一种语言，但是系统给出的答案必须是英语文本。
    - 2003年：Dutch, French, German, Italian, Spanish；
    - 2004年：Dutch, French, German, Italian, Spanish, others。

问答技术现状

- 信息检索 + 信息抽取
- 信息检索 + 模式匹配
- 信息检索 + 自然语言处理技术
- 基于统计翻译模型的问答技术

信息检索+信息抽取

- **方面描述**：从问句中提取关键词语，用信息检索的方法找出包含候选答案的段落或句子，然后基于问答类型用信息抽取的方法在这些段落和句子中找出答案
- **算法核心**：段落或者句子级的排序
- **排序算法**：
  - 把关键词分为：普通关键词，短语，扩展关键词…
  - 不同的关键词对句子的排序的贡献不同
  - Score = wo*O + we*E + wb*B + wt*T + …
- **算法特点**：
  - 优点：技术相对成熟，易于开发
  - 缺点：准确率一般，不能推理
- **代表**：新加坡国立大学Hui Yang等人研发的系统

信息检索+模式匹配

方法描述：

- 基本思想是：对于某些提问类型（某人的出生日期、某人的原名、某物的别称等），问句和包含答案的句子之间存在一定的问答模式，该方法在信息检索的基础上根据这种问答模式找出答案。因此如何自动获取某些类型提问的尽可能多的答案模式是其中的关键技术。
- 例如，询问“某人生日年月日”类提问的部分答案模式如下：
  - 1.0 <NAME>（<ANSWER> -）
  - 0.85 <NAME> was born on <ANSWER>
  - 0.6 <NAME> was born in <ANSWER>
  - 0.59 <NAME> was born <ANSWER>
  - 0.53 <ANSWER> <NAME> was born
  - 0.50 – <NAME>（<ANSWER>
  - 0.36 <NAME>（<ANSWER> -

信息检索+模式匹配（cont.）

- 包括两阶段的任务：
  - 离线阶段：获取问答模式
  - 在线阶段：首先判断当前提问属于哪一类，然后使用这类提问的所有模式来对抽取的候选答案
- 模板获取方法：
  - 表层字符串匹配[Deepak Ravichandran, 2002]
  - 深层句法分析[Dekang Lin, 2000]
  - 人们已经开始把注意力从原来的基于深层文本分析方法转移到基于字符的表层的文本分析技术上。
- 算法特点：
  - 优点：对于某些类型的问题(如生日问题等)有良好的效果
  - 缺点：无法表达长距离、复杂关系，没有推理能力
- 代表：俄罗斯InsightSoft-M公司Martin Soubbotin等人研发的系统

信息检索+自然语言处理技术

- 尽管IR + IE算法技术相对成熟
- 尽管IR + Pattern Matching算法对某些类问题很有效
- 但要提高问答系统的性能，理解问句、解释答案，必须结合自然语言处理的技术[Eric Nyberg et al. 2002]；也就是在对问句和答案句进行句法语义分析的基础上进行问答。
- 现阶段，自然语言处理的技术还不成熟，对句子的深层句法、语义分析还不能达到实用的效果。因此，大多数系统都是基于对句子进行浅层分析，获得句子的浅层句法、语义表示，作为对前两种方法的补充和改进
- 算法特点
  - 优点：能够从语法、语义的角度解析答案
  - 缺点：技术还不成熟

信息检索+自然语言处理技术（cont.）

- 典型系统：美国Language Computer Corporation公司Sanda Harabagiu等人研发的系统，该系统在TREC QA Track 评测中获得好成绩，且具有较大的领先优势。

基于统计翻译模型的问答技术

- **方法描述**：把提问句看作答案句在同一语言内的一种翻译；
- **特点**：过分依赖于训练集。

四类问答技术的比较分析

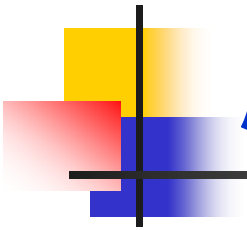
- **基于信息检索和信息抽取的问答技术：**相对简单，容易实现。但它以基于关键词的检索技术(也可被称为词袋检索技术)为重点，只考虑离散的词，不考虑词之间的关系。因此无法从句法关系和语义关系的角度解释系统给出的答案，也无法回答需要推理的提问。
- **基于模式匹配的问答技术：**虽然对于某些类型提问(如定义，出生日期提问等)有良好的性能，但模板不能涵盖所有提问的答案模式，也不能表达长距离和复杂关系的模式，同样也无法实现推理。

四类问答技术的比较分析(cont.)

- **基于自然语言处理的问答技术**：可以对提问和答案文本进行一定程度的句法和语义分析，从而实现推理。但目前自然语言处理技术还不成熟，除一些浅层的技术(命名实体识别，汉语分词、词性标注等)外，其他技术还没有达到实用的程度。所以，这种技术的作用还有限，只能作为对前两种方法有效的补充。
- **基于统计翻译模型的问答技术**：在很大程度上依赖训练语料的规模和质量，而对于开放域问答系统，这种大规模训练语料的获取是非常困难的。

前三类方法在Factoid子任务中的名次

	IR + IE	IR + Pattern Match	IR + IE + NLP
2000	-	-	1
2001	-	1	2
2002	3	2	1
2003	2	-	1



总结

- 基于字符表层的文本分析技术(例如模板技术)必须和快速、浅层自然语言处理技术有效结合，才能获得性能优良的问答系统。

可以使用于QA的NLP技术

- 命名实体识别 (Named Entity Recognition)
- 句法分析 (Syntactic Parsing)
- 逻辑表示 (Logic Form)
- 互指关系 (Co-reference)
- 复述（同义互训）(Paraphrase)
- ……

命名实体识别

对于问答系统非常重要

- 缩小了候选答案的范围
- 参加TREC的几乎所有系统均使用了NER技术

如何使用NER技术

- 在段落或者句子的排列阶段：问答系统首先根据查询关键词进行检索，然后对于检索出来的段落或句子重新进行排序：当某个句子包含所期望的实体时，则给句子适当的加分。
- 在候选答案的抽取阶段：大多数的问答系统都是在答案抽取阶段使用命名实体的技术，答案抽取模块只抽取和期望答案类型一致的实体作为答案，而命名实体不参与句子或段落的排序。

句法分析

- 短语结构分析或依存结构分析的结果是得到句子的短语结构句法树或依存结构句法树。在句子排序或答案抽取阶段，使用更合理的句法信息。

- 例子：
  - 提问：Who killed Lee Harvey Oswald?
  - 文本：Belli's clients have included Jack Ruby, who killed John F. Kennedy assassin Lee Harvey Oswald and Jim and Tammy Bakker.
  - 候选答案包括Jack Ruby和John F. Kennedy，如果采用基于词袋的检索问答技术，系统很有可能返回John F. Kennedy，因为John F. Kennedy和查询关键词killed、Lee Harvey Oswald的距离更近。但是，如果引入句法信息，系统只会返回答案Jack Ruby。因为Jack Ruby在文本中是killed的逻辑主语，Lee Harvey Oswald是killed的逻辑宾语，这和问句的句法结构完全相似。

句法分析 (cont.)

- 方法的局限性
  - 依赖句法分析器的结果
  - 通过比较提问和文本的句法树来抽取答案虽然提高了系统的性能，但这种基于句法树分析的方法还是非常浅层的。因为对句法树的分析基本上就是合一(Unification)运算，比较两棵句法树的相似性，无法回答那些需要语义信息才能回答的提问。

逻辑表示 (Logic Form)

- 通过逻辑表示方法，表示问句和答案句的某些语义，并支持某种程度的推理。
- 问句和文本同时转化成统一的Logic Form（QLF和ALF），通过对QLF和ALF的运算来抽取答案。Logic Form最大的特色是它结合WordNet可以表达语义知识，实现推理功能，这也是LCC系统在TREC QA评测中取得好成绩的主要原因。

逻辑表示 (cont.)

问句：谁杀了林肯？

文本：林肯被布斯刺杀身亡。

问句逻辑表示：谁 (x1) & 杀 (e1,x1,x2) & 林肯 (x2)

文本逻辑表示：

林肯 (x1) & 布斯 (x2) & 刺杀 (e1,x2,x1) & 身亡(e2, x1, x3)

词汇链推理规则：

刺杀 (e2,x2,x1) -----〉杀(e2,x2,x1)

Paraphrase技术

- Paraphrase是指用不同的词汇-句法结构表达同样的意思。词汇链就是一种特殊的Paraphrase：词汇Paraphrase。Paraphrase可以解决因提问和答案的表述不同给问答系统的设计带来的麻烦。

- 例：
  - When did Colorado become a state?
  - (1a) Colorado became a state in 1876.
  - (1b) Colorado was admitted to the Union in 1876.
  - Who killed Abraham Lincoln?
  - (2a) John Wilkes Booth killed Abraham Lincoln.
  - (2b) John Wilkes Booth ended Abraham Lincoln's life with a bullet.

Paraphrase技术（cont.）

- 如果上述两个提问的答案都是以(1a)(2a)的形式来表述的，问答系统可以使用非常简单的技术(命名实体识别技术)就可以找出答案。但是如果答案以(1b)(2b)的形式出现，问答系统要找到答案将是非常困难的。但是通过Paraphrase技术获得Paraphrase规则，问答系统就能容易地找出提问的答案。

小结

- 自然语言处理领域和信息检索的一个重要分支和新兴的研究热点，其”通过系统化、大规模地定量评测推动研究向前发展”的发展轨迹，以及某些成功启示，都极大地推动了自然语言处理研究的发展，促进了NLP研究与应用的紧密结合。
- 目前的问答技术也不成熟，问答系统能够处理的提问非常有限，系统的性能离实用的目标还很远。

小结（cont.）

- 存在一些问题：
  - 目前我们的问答系统基本上都是针对具有简短答案的事实问题研发的，但这样的系统在实际应用中到底能够解决用户真正关心问题的百分之多少，或者说我们应该研究哪种类型问答系统。这点非常值得我们去研究。
  - 重视大规模的公开评测技术，以评测推动问答技术的发展。现阶段对于汉语问答技术的研究，我们迫切需要一个公开、公认、合理的问答评测平台。
  - 从问答技术的研究角度看，我们需要重视基于字符表层的文本分析技术和基于自然语言处理技术的有效结合，扬长避短。

NLPR做的一些工作

- 构建中文问答系统评测环境
- 系统原理图
- 中文问答系统的提问分类技术
- 支撑中文问答系统的命名实体识别技术
- 基于聚类的中文问答系统句子检索模型
- 中文问答模式学习及其应用
- 工作总结

评测目的

- 通过系统化、大规模的定量评测推动研发向前发展的研究方法和技术路线受到越来越多的研发人员的重视
  - 比较不同方法的优劣，获得结论
  - 以系统化、大规模测试为基础，推动研究的发展
  - 通过对真实环境的模拟，加速研究成果转化为产品
  - 发展适当且具应用性的评估技术

问答评测平台的组成环节

- 建立语料库
  - 固定语料库的语料来源和大小： 互联网网页
  - 语料大小：1.8GB
- 建立测试集
- 建立打分机制
  - 参照TREC, NICIR：主要指标是MRR

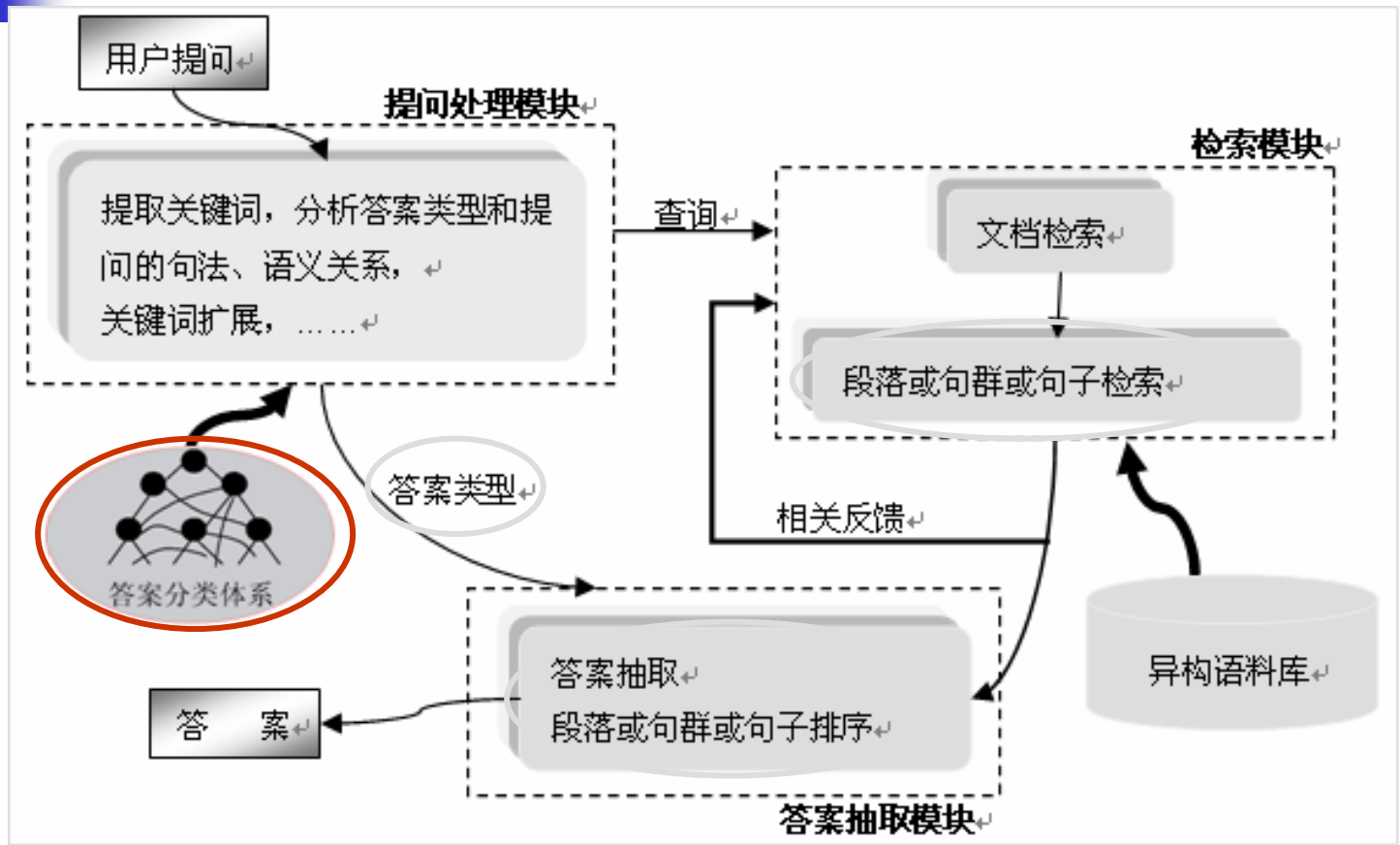
建立测试集

- 建立测试的原则
  - 全面性；真实性；无歧义性
- 测试集的来源
  - 自然语言搜索网站日志
  - 问答题库
  - 实验室工作人员的兴趣提问
  - 对英语提问(TREC)的翻译
- 现阶段的测试集的大小
  - 收集7050个测试问题，以及部分提问(1000个)在语料库中的答案

NLPR做的一些工作

- 构建中文问答系统评测环境
- **系统原理图**
- 中文问答系统的提问分类技术
- 支撑中文问答系统的命名实体识别技术
- 基于聚类的中文问答系统句子检索模型
- 中文问答模式学习及其应用
- 工作总结

问答系统原理图


用户提问
提问处理模块
提取关键词，分析答案类型和提问的句法、语义关系，
关键词扩展，……
查询
检索模块
文档检索
段落或句群或句子检索
答案类型
答案分类体系
相关反馈
异构语料库
答案抽取
段落或句群或句子排序
答　案
答案抽取模块

提问分类技术

- 系统原理图
- 构建中文问答系统评测环境
- 中文问答系统的提问分类技术
  - Chinese Question Classification from Approach and Semantic Views. In the Proceedings of AIRS2005,Korea.
- 支撑中文问答系统的命名实体识别技术
- 基于聚类的中文问答系统句子检索模型
- 中文问答模式学习及其应用
- 工作总结

为何要提问分类？

- 问答系统的四大组成部分
  - 提问处理（提问分类）
  - 文本检索，句子检索
  - 候选答案抽取
  - 答案选择
- 提问分类是整个问答系统的基石，D. Moldovan工作说明问答系统的错误大约有 36.4%是由于提问分类的错误导致的

提问分类的相关工作

- 主要是英语提问的分类研究
- 采用的是Hierarchical分类体系,且针对的是提问的语义类别进行分类
- 分类算法主要包括
  - Support Vector Machine, SnoW, Language Model, Maximum Entropy Model, K-means, Handcrafted Rules

中文提问分类体系的设计

- 提问分类的作用/遵守三个原则
  - 提问类型应该确定后续候选答案提取模块和答案选择模块应该采用的技术类型，这是提问的技术分类
  - 提问类型应该确定提问答案的语义类型，限制候选答案的产生，这是提问的语义分类
  - 设计的提问分类体系在现有的资源和工具上是可实现的，并且具有一定的普遍性和覆盖度。

中文提问分类体系

Question Type	#	Question Type	#	Question Type	#
Approach Categories					
ABBR	42	SYNONYM	190	LIST	234
CH-ABBR	11	BIRTHDAY	52	REASON	129
EN-ABBR	39	BIRTHDAY-PLACE	25	MANNER	90
ABBR-EX	92	BOOK-AUTHOR	71	CONTRAST	41
CH-ABBR-EX	6	REAL-NAME	41	FUNCTION	39
EN-ABBR-EX	7	CAPITAL-PLACE	48	DESCRIPTION	128
YES-NO	10	POPULATION	52	COMPONENTS	51
TRANS-TO-OTHERS	19	WHY-FAMOUS	79	CAUSE-OF-DEATH	41
TRANS-TO-EN	22	DEFINITION	257	OTHER-APPROACH	
TRANS-TO-CH	25				
Semantic Categories					
OTHER-TEMP	399	PARTY	20	MUSIC-INSTRU	35
DURATION	123	SPORTS-TEAM	35	BOOK-NAME	48
SEASON	28	UNIVERSITY	37	MOVIE-NAME	62
YEAR	206	MAGNEWS	18	PRODUCT	11
MONTH	37	BANK	26	PHONE-NUMBER	37
DATE	90	OTHER-ENTITY	641	ZIP-CODE	33
TIME	16	DYNASTY	48	EMAIL	4
AGE	48	LANGUAGE	56	URL	13
PERSON	769	ANIMAL	129	OTHER-NUMBER	426
OTHER-PLACE	606	PLANT	58	MONEY	67
CONTINENT	73	OCCUPATION	19	SPATIAL-NUMBER	209
COUNTRY	334	HUMAN-FOOD	50	SPEED	41
PROVINCE	99	BODY-PART	32	WEIGHT	44
CITY	189	DISEASE	26	ACCELERATION	25
BODY-OF-WATER	142	SPORT	36	ORDINAL	34
ISLAND	35	COLOR	51	PERCENTAGE	68
MOUNTAIN	45	UNIT	27	TEMPERATURE	56
SPHERE	72	MONETARY-UN	37	RANGE-NUMBER	19
OTHER-ORG	148	NATINOALITY	31		

中文提问分类体系特点

- 提问的类别
  - 28 Approach Types
  - 56 Semantic Types
- 问答系统技术路线决定的：基于字符表层的文本分析技术的有效性和快速、浅层自然语言处理技术的必要性必须紧密结合
- 中文提问分类体系特点
  - 两个Parallel结构的提问分类体系
  - 技术类别确定该提问可以使用高性能的模板技术，而语义类别又可以限制模板技术的机械性匹配
- Example: 莫扎特是哪年出生的？
  - 技术类属于C-BIRTHDAY类别，语义类属于C-YEAR类别

提问的分类算法

- 基于SVM的分类算法
- 特征空间
  - 基本特征 (Word and POS)
  - 结构特征 (Bi-gram or Dependency Relation)
  - 词汇语义特征 (Thesaurus and Named Entity Types)
- Hybrid Feature Weighting

$$t_{ik} = \beta^{d} \times \sqrt[\eta]{\chi^{2}_{max}} \times TF \times IDF$$

χ^{2}_{max} 反应特征与类别之间的局部属性关系，IDF则是特征与整个训练语料之间的整体属性关系; β 是对距离d的惩罚因子

提问分类算法的性能指标

- 实验语料
  - 模型训练：**6350** 个提问
  - 模型测试：**700**个提问
- FSET6: Word + POS + Named Entities + Thesaurus + Bi-Gram