8.1 Question Answering

首先从几个著名的问答系统引入，如 Siri、Ask Jeeves（ask.com，其实是支持语音的搜索引擎）、WolframAlpha（解决自然语言描述的数学问题）、Watson（IBM 的成果，在 Jeopardy game 问答比赛中击败人类代表）。

接下来展示了一些数据集中的问题，不再赘述。问题分类（Question Type Taxonomy）是一个值得关注的话题，它有助于优化回答的质量，常见的问题类型如：

- Yes/no vs. wh-，wh-即 what、when、where、why、how 等；
- Factual vs. Procedural，factual questions 是针对客观事实的问题，procedural 则需要考虑提问的上下文等具体信息；
- Single answer vs. multiple answers
- Objective vs. Subjective，objective 指有着客观的答案，subjective 指之能做出主观的回答；
- Context-specific (personalized) vs. generic
- Known answer in the collection, y/n，即该问题是否在数据集的支持范围之内。

此外，在 8.2 节还举例说明了一些其他类型的问题，如 Definitional（询问某事物的定义）、List（列举一些实例）、Crosslingual（问题与参考材料语言不同）、Series（一系列上下文相关的问题）。

早期的问答系统主要关注 factual, short-answer questions，常常是基于信息检索（information retrieval）的方法，用到的知识和推理相对较少。

接下来介绍了几个早期的问答系统，它们的作者和时间参考第 15 页 ppt。首先是 Eliza，它不能给出有意义的回答，但是它给出了一种思路。Eliza 所做的的只是识别用户语言的结构（simple pattern matching），然后利用用户给出的信息生成问题反问用户。例如 I am always tired.->Why do you say you are always tired?

然后介绍了 Lunar 系统，它针对 Apollo 11 moon rocks（阿波罗 11 发回的月球岩石的相关知识）这一特定领域。值得注意的是它使用了 ATN （augmented transition network）这一语法模型，并能够通过语义推理（基于 a procedural-semantics framework）将语法结构变为逻辑表达。此外，它具有规模 3500 的词典，13000 个实体构成的知识库。它回答的问题如 How many breccias（角砾岩） contain olivine（橄榄石）。作为备注，ATN 由 1972 年 Bill Woods

的论文提出。

随后是 SHRDLU 系统，它是一个识别用户指令并操纵虚拟块体（立方体 cube、金字塔 pyramid 等）的系统。它具有推理能力（例如回答 Find a block which is taller than the one you are holding and put it into the box），能够处理问题之间的上下文（例如回答 What does the box contain）。

最后简单介绍了 Start 系统，与前几个不同它是开放领域系统，并且是最早的基于网络的系统（the oldest web based system）。

8.2 Evaluation of QA, System Architecture

本节首先介绍问答系统的评价。一个数据集是 TREC（Text retrieval evaluation conference）提供的，该会议由 NIST 运营（National Institute for Standard and Technology in the United States）。该数据集提供问题（693 questions in 2000）以及蕴含答案的文本集（2 GB），而且这些问题都是基于事实的（factual）且不必要进行推理的，即纯粹的信息检索算法也可以用它来评价。问题的答案一般是简短的。

接下来介绍评价指标。常用的评价指标是 MRR（Mean Reciprocal Rank，平均倒数排列），公式为，其中只有正确答案对应的 $rank_i$ 才可以参与计算：

$$\mathrm{MRR} = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{rank_i}\right)$$

例如问题 What is the capital of Canada，给出的 5 个答案依次为 1.Toronto, 2.Ottawa, 3.Albany, 4.Philadelphia, 5.Ottawa（第 2、5 为正确答案）。则 MRR=1/2=0.5，第 5 个虽然正确但无贡献。若使用 TRR（total reciprocal rank），则 TRR=1/2+1/5=0.7。另外，近年的改进版本进一步考虑 Confidence-weighted score，即需要算法给定每个答案的置信度。这样，如果给出了正确答案但对其置信度较低，则对 MRR 的贡献也将很小。

然后列举了近年来在 TREC 中表现出色的参赛者（但并没有给出其得分，有点奇怪）。

接下来讨论问答系统的架构。首先一个直观的认识是搜索引擎（信息检索方法）可以用来回答问题，但是它给出的结果一般是可能含有答案的文档列表，因此需要其他的处理。

一种经典的架构是：

1. 根据问题构建查询（Query modulation, best paraphrase of a NL question given the syntax of a search engine），一个例子是 Who wrote Hamlet → author | wrote Hamlet；

2. 文档检索（Document retrieval）；

3. 文档、段落或语句的排列（Sentence ranking），涉及到的方法如 n-gram matching；

4. 答案提取（Answer extraction），可能使用 question type classification、phrase chunking（即尽量保证短语完整、避免被分裂）等技术；

5. 答案排列（Answer ranking）。

这里介绍了问题分类。之前也提到过，该过程有利于获得更准确的答案，例如对于 Who wrote Hamlet，能够将答案限定在人名、作家等类别中将十分有利。这里介绍了两个问题类别体系（taxonomy of question type），分别是

- SYN-classes，from IBM's answer selection system or Ansel，包含约 20 个类别，如图 8.2.1 所示；
- UIUC（University of Illinois Urbana Champagne） Question Types，包含比 SYN 更多的类别，如图 8.2.2 所示；

QA-token	Question type	Example
PLACE	Where?	In the Rocky Mountains
COUNTRY	Where? What country?	United Kingdom
STATE	Where? What state?	Massachusetts
PERSON	Who?	Albert Einstein
ROLE	Who?	Doctor
NAME	Who? What? Which?	The Shakespeare Festival
ORG	Who? What?	The U.S. Post Office
DURATION	How long?	For 5 centuries
AGE	How old?	30 years old
YEAR	When? What year?	1999
TIME	When? What time?	In the afternoon
DATE	When? What date?	July 4th, 1776
VOLUME	How big? How large?	3 gallons
AREA	How big? How large?	4 square inches
LENGTH	How long? How big?	3 miles
WEIGHT	How heavy? How big?	25 tons
NUMBER	How many?	1,134.5
RATE	How much? What percentage?	50 per cent
MONEY	How much?	4 million dollars

图 8.2.1

- **ENTITY**: entities
  - animal: animals
  - body: organs of body
  - color: colors
  - creative: inventions, books and other creative pieces
  - currency: currency names
  - dis.med.: diseases and medicine
  - event: events
  - food: food
  - instrument: musical instrument
  - lang: languages
  - letter: letters like a–z
  - other: other entities
  - plant: plants
  - product: products
  - religion: religions
  - sport: sports
  - substance: elements and substances
  - symbol: symbols and signs
  - technique: techniques and methods
  - term: equivalent terms
  - vehicle: vehicles
  - word: words with a special property
- **ABBREVIATION**: abbreviation
  - abb abbreviation
  - exp expression abbreviated
- **DESCRIPTION**: description and abstract concepts
  - definition: definition of sth.
  - description: description of sth.
  - manner: manner of an action
  - reason: reasons
- **HUMAN**: human beings
  - group: a group or organization of persons
  - ind: an individual
  - title: title of a person
  - description: description of a person
- **LOCATION**: locations
  - city: cities
  - country: countries
  - mountain: mountains
  - other: other locations
  - state: states
- **NUMERIC**: numeric values
  - code: postcodes or other codes
  - count: number of sth.
  - date: dates
  - distance: linear measures
  - money: prices
  - order: ranks
  - other: other numbers
  - period: the lasting time of sth.
  - percent: fractions
  - speed: speed
  - temp: temperature
  - size: size, area and volume
  - weight: weight

图 8.2.2

8.3 节的开始指出，Question Classification 可以看做经典的分类问题，提取特征并利用机器学习可以处理。结合它的特点也可以构建规则，例如利用正则表达式匹配特定模式并推理出问题的类别。需要注意的是，一个问题对应到类别可能是歧义的，即该问题可能同时属于多个类别。

本节的最后列举了几个 UIUC 关于问题分类的论文链接。

8.3 QA System Architecture

开始介绍的是 Question Classification，已经总结在了上一节的最后。

对于 Query Modulation（Query Formulation），关键在于去适应特定的搜索引擎。通常需要进行的处理是：

- 忽略原问题中某些词，这些词也被称为 stop words；
- 用一个或一组词替换掉原问题中的某个词。

对于 Passage Retrieval，常用于检索的特征有：

- 与查询中的词匹配（Proper nouns that match the query）；
- 查询中的词出现在彼此的附近（Near each other）；
- 匹配到所需的实体类型（Entities that match the expected answer type）。

对于 Answer Retrieval，是在上一步检索到的文章、段落中进一步识别所需的句子、短语，会使用到诸如 NER（name entity recognition）的技术，常用的特征有：

- 该实体距离查询中的词汇的距离（Distance to query words），显然在查询词汇附近的实体更有可能成为答案；
- 匹配所需的答案类型（Answer type）；
- 词相似性，例如 Wordnet similarity、词向量相似性；
- Redundancy（过多，冗余），在答案列表中出现的频率高更可能是真正的答案，这主要是自然语言中的 paraphrase（意译，改述）和 variability（易变性）导致的。

最后专门讨论了开放领域的、基于网络的系统（open-ended web-based system），指出它们与传统基于本地全文的系统（traditional corpus-based system）有显著不同，主要体现在：

- 规模显著扩大（Significantly larger corpus）；

- 人工预处理更为困难（No pre-annotation is possible）。

另外还要认识到，对于各种问答系统而言，搜索引擎都只能作为一部分（partially helpful），主要原因有

- Stop words 的过程会忽略重要信息（例如 who）；
- 不能很好地利用 Question types；
- 查询本身有其限制（Restrictions on queries），如长度限制等；
- Issues with reliability（可靠性）, timeliness（合时、时效性）, inaccurate（不准确的）answers。

8.4 Question Answering Systems

本节介绍一些近年来的问答系统，这些系统使用到了之前介绍的方法或策略。

首先是 AnSel（John Prager et al. 1999），这是 IBM 参赛 TREC 的系统。它在构建数据集时使用了预注释（Predictive Annotation）的方法，即提前识别和标注了数据集中的实体。Radev 指出这是一种介于 Knowledge-based 与 Knowledge-poor 的方法，前者将数据集表达为具有复杂从属关系的知识库，后者基于某些文本相似性进行检索，8.4.1 是一个示意图。

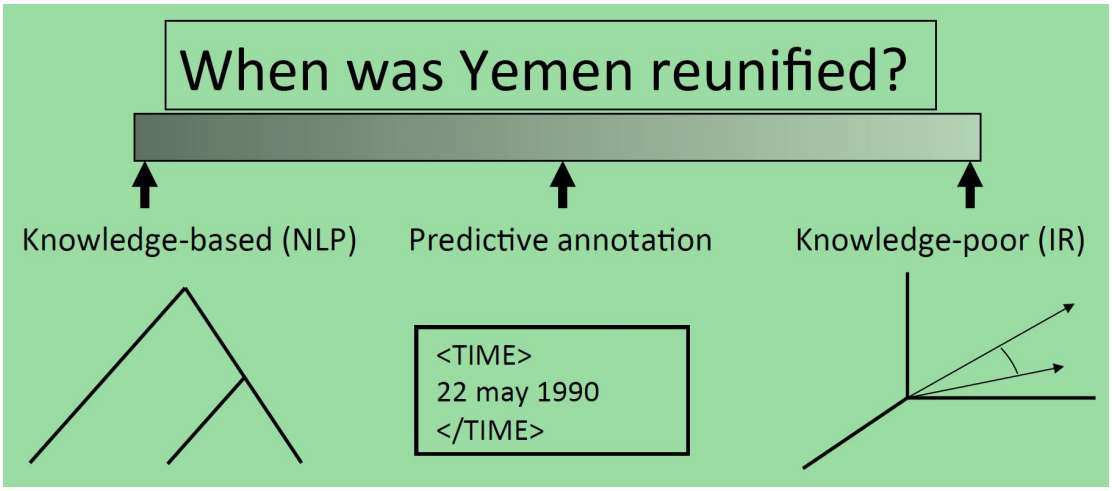

图 8.4.1

基于这种方式，系统可以根据所需的答案类型缩小候选空间。在得到候选答案后，AnSel 对每个答案的特征进行 Logistic Regression，利用所得评分对候选结果（称为 span）排序，使用到的特征有：

1. Average distance，span 与查询中单词在位置上的平均距离；

2. Not in query，在 span 中出现但没有在查询中出现的单词数目，注意一般而言该数值越大越好，因为出现在查询中的单词反而不太可能出现在答案中；

3. Frequency，该 span 在整个候选段落集合（passage set）中出现的次数；

4. Sscore（search engine relevance score），搜索引擎计算出的 span 所在段落（passage）与查询的相关性；

5. Number，该 span 在整个（GuruQA 的）搜索结果列表中的位置是第几个；

6. Rspanno（relative span number），该 span 在当前段落中的位置是第几个:

7. Count，该 span 所在段落内有多少个 span（重复的只记一次）；

8. Type，该 span 所属的实体类型在候选类型中排在的位置。例如对于 who 问题，候选类型可能是 PERSON、ORG（组织）、NAME（一般指人名以外的名称）、ROLE，注意排在前面的表示更有可能的，那么人名 John Snow 的 Type 取值就为 1。

简单介绍了 IONAUT（Abney Cass et al. 2000）。它使用的 passage retrieval 组件是 START，（Salton, Buckley），entity recognition 组件是 partial parser（Abney Cass），entity classification 简单地使用了 8 中问题类型。

Mulder（Kwok et al. 2001）是第一个大规模的网络问答系统（First large-scale Web QA system），主要组件有：

- Maximum entropy parser (Charniak)
- PC-Kimmo for POS and morphological analysis for unknown words
- Link parser (a type of dependency link parser, by Sleator and Temperley)
- Google as the underlying search engine.

此外，使用了 tokenization（标记化）技术，即 identify phrases in quotes（引用、引述），以及 Query transformations（即 Query Modulation）。

简单介绍了 NSIR（Radev et al. 2002），它使用了一种 Probabilistic phrase reranking 来确定短语的实体类型。简单来说就是利用训练样本得到参数 P(qtype|signature)，signature 是短语对应的词性序列（POS sequence），qtype 是短语的实体类型。NSIR 使用的搜索引擎有 AlltheWeb, NorthernLight, Altavista, Google。

简要介绍了 AskMSR（Michelle Banko et al. 2002），它的基本结构如图 8.4.2。

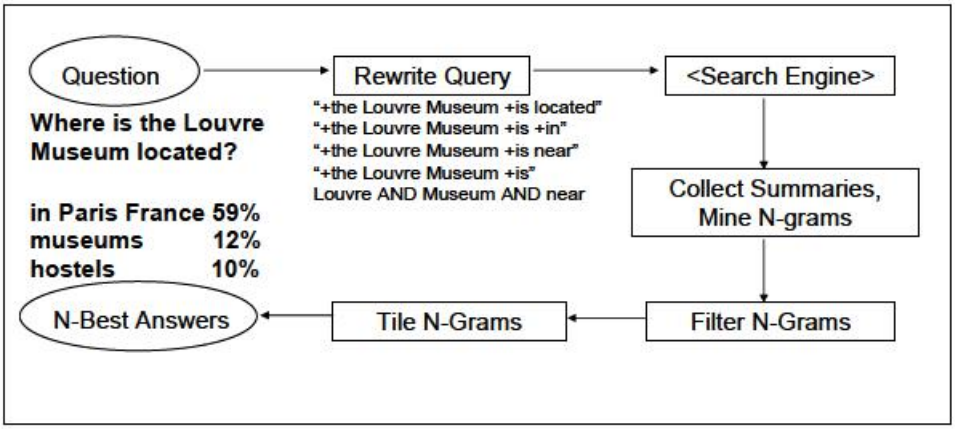

图 8.4.2

需要说明的是 tile n-gram 是指将较短的 gram 连接成较长的，即 Combining A B C and B C D into A B C D，如”Mr. Charles” 和”Charles Dickens” 组合成“Mr. Charles Dickens”。

简要介绍了 Echihabi and Marcu 系统，它是基于 noisy-channel 模型的，即给定问题 S，从数据集中找到 P(q|S)最大的答案 q。显然这需要一种有效的 sentence simplification，这样才能有效地捕捉模式来训练概率模型。

8.5 Question Answering Systems 2/2

LCC（Harabagiu Moldovan et al. 2003）与之前的系统不同，它是基于语义表达和推理的系统。首先，它利用逻辑形式转换（logic form transformations）将数据集中的句子表达为语义逻辑。这样，给定问题后就可以利用语义规则（semantic axioms）进行推理。显然这种方式不局限于问题本身包含的词汇。一个例子如图 8.5.1。

Example:
Heavy selling of Standard & Poor's 500-stock index futures in Chicago relentlessly beat stocks downward.

LF:
heavy_JJ(x1) & selling_NN(x1) & of_IN(x1,x6) & Standard_NN(x2) & &_CC(x13,x2,x3) & Poor_NN(x3) &'s_POS(x6,x13) & 500-stock_JJ(x6) & index_NN(x4) & future_NN(x5) & nn_NNC(x6,x4,x5) & in_IN(x1,x8) & Chicago_NN(x8) & relentlessly_RB(e12) & beat_VB(e12,x1,x9) & stocks_NN(x9) & downward_RB(e12)

Q1394: What country did the game of croquet originate in ?
Answer: Croquet is a 15th-century French sport that has largely been dominated by older, wealthier people who play at exclusive clubs.
Lexical chains:
(1) game:n#3 → HYPERNYM → recreation:n#1 → HYPONYM → sport:n#1
(2) originate_in:v#1 → HYPONYM → stem:v#1→ GLOSS → origin:n#1 → GLOSS → be:v#1

图 8.5.1

QASM（Radev & al. 2001）是一种 Noisy channel model，在将自然问题转换为查询的过程中涉及下面三种操作（channel operators）：

- DELETE, e.g., delete prepositions, stop words

- REPLACE, e.g., replace a noun phrase with a WordNet expansion
- DISJUNCT, e.g., replace a noun phrase with a disjunction

Ravinchandran 和 Hovy 在 2002 的系统是一种半监督的方式，可以利用有限的已知问题和答案自动学习简单的特征（Find patterns that contain both the question and the answer terms），过程与之前 POS 问题中的类似。例如对 Mozart was born in 1756，它可能学习到 <NAME> was born on <BIRTHDATE>的模式。

Watson（David Ferrucci et al. 2010）是一个更为复杂的问答系统。它使用了 DeepQA，使得 Watson 可以基于大规模的开放性知识回答自然问题，它的硬件系统也十分强大。

值得注意的是，它可以使用不同类型的知识源（结构化和非结构化数据、知识表达与推理引擎）。对于候选答案可以给出有效的置信度，这使得它可以在反应速度和准确度上取得平衡，以胜任抢答环境。此外，它的问题类型体力庞大，包含约 2500 种，其中 200 种是常用的。

具体的知识参考 Ferrucci et al. 2010. Building Watson: An Overview of the DeepQA Project. AI Magazine. Fall 2010. 59-79.

接下来列举了问答系统面临的挑战：

1. 歧义（Word Sense Disambiguation）；

2. 问题中的代指关系（Coreference Resolution）使得理解问题更为困难；

3. 语义标记（Semantic Role Labeling），例如识别出句子中的谓语（main predicate in sentences）后可进一步明确问题结构和含义，便于寻找答案；

4. 时效性问题（Temporal questions），如何确保答案是与时俱进的；

5. 问题的讨论范围，例如 Categories on Jeopardy，Watson 在 Jeopardy 中就曾因为没能理解问题给出的限定范围而回答错误。