

# 美团大脑：知识图谱的建模方法及其应用

## (<https://tech.meituan.com/2018/11/01/meituan-ai-nlp.html>)

📅 2018年11月01日    仲远    ✎ 13651字    ⌚ 28分钟阅读

作为人工智能时代最重要的知识表示方式之一，知识图谱能够打破不同场景下的数据隔离，为搜索、推荐、问答、解释与决策等应用提供基础支撑。美团大脑围绕吃喝玩乐等多种场景，构建了生活娱乐领域超大规模的知识图谱，为用户和商家建立起全方位的链接。我们美团希望能够通过对应用场景下的用户偏好和商家定位进行更为深度的理解，进而为大众提供更好的智能化服务，帮大家吃得更好，生活更好。

近日，美团 AI 平台部 NLP 中心负责人、大众点评搜索智能中心负责人王仲远博士受邀在 AI 科技大本营做了一期线上分享，为大家讲解了美团大脑的设计思路、构建过程、目前面临的挑战，以及在美团点评中的具体应用与实践，其内容整理如下：

## 知识图谱的重要性



近年来，人工智能正在快速地改变人们的生活，可以看到各家科技公司都纷纷推出人工智能产品或者系统，比如说在 2016 年，谷歌推出的 AlphaGo，一问世便横扫整个围棋界，完胜了人类冠军。又比如亚马逊推出的 Amazon Go 无人超市，用户只需下载一个 App，走进这家超市，就可以直接拿走商品，无需排队结账便可离开，这是人工智能时代的“新零售”体验。又比如微软推出的 Skype Translator，它能够帮助使用不同语言的人群进行实时的、无障碍的交流。再比如说苹果推出的 Siri 智能助理，它让每一个用苹果手机的用户都能够非常便捷地完成各项任务。所有这些人工智能产品的出现都依赖于背后各个领域技术突飞猛进的进展，包括机器学习、计算机视觉、语音识别、自然语言处理等等。

### 美团点评在人工智能的布局

- 美团点评AI平台部NLP（自然语言处理）中心于2018年2月正式成立
- 愿景：用人工智能帮大家吃得更好，生活更好
- NLP (Natural Language Processing)：语言是人类智慧的结晶，自然语言处理是人工智能中最为困难的问题之一，其核心是让机器像人类一样理解和使用语言

用户发表的评价

机器阅读这条评价，充分理解用户的喜怒哀乐

一个商家2242条评价

机器代替用户快速阅读大量评价，并总结归纳商家的情况，供用户参考

#### 人工智能助理

用户餐饮娱乐的决策需求

基于对用户评价和商家的深刻理解，帮助用户快速完成决策

作为全球领先的生活服务电子商务平台，美团点评在人工智能领域也在积极地进行布局。今年 2 月份，AI 平台部 NLP 中心正式成立，我们的愿景是用人工智能帮大家吃得更好，生活更好。语言是人类智慧的结晶，而自然语言处理是人工智能中最为困难的问题之一，其核心是让机器能像人类一样理解和使用语言。

我们希望在不久的将来，当用户发表一条评价的时候，能够让机器阅读这条评价，充分理解用户的喜怒哀乐。当用户进入大众点评的一个商家页面时，面对成千上万条用户评论，我们希望机器能够代替用户快速地阅读这些评论，总结商家的情况，供用户进行参考。未来，当用户有任何餐饮、娱乐方面的决策需求的时候，美团点评能够提供人工智能助理服务，帮助用户快速的进行决策。

### 人工智能两大技术驱动力

#### 深度学习（隐性模型）

- 面向某一个具体任务（如下围棋，识别猫，人脸识别，语音识别等）
- 需要海量训练数据
- 需要强大的计算力

#### 知识图谱（显性模型）

- 可广泛用于不同任务
- 从海量数据中进行知识学习和挖掘
- 可理解、可解释，类似人类的思考方式

	深度学习	知识图谱
场景示例	人脸识别	语音助手
目前进展	在一些任务上已经接近或超过人类	在知识量上超过人类，在知识推理上不如人类
任务范围	面向具体任务，难以迁移	广泛适用于不同任务
可解释性	较难解释	可解释性强
数据量	海量训练数据	海量知识数据
未来趋势	未来深度融合	

所有这一切，都依赖于人工智能背后两大技术驱动力：深度学习和知识图谱。我们可以将这两个技术进行一个简单的比较：



我们将深度学习归纳为隐性的模型，它通常是面向某一个具体任务，比如说下围棋、识别猫、人脸识别、语音识别等等。通常而言，在很多任务上它能够取得非常优秀的结果，同时它也有非常多的局限性，比如说它需要海量的训练数据，以及非常强大的计算能力，难以进行任务上的迁移，而且可解释性比较差。

另一方面，知识图谱是人工智能的另外一大技术驱动力，它能够广泛地适用于不同的任务。相比深度学习，知识图谱中的知识可以沉淀，可解释性非常强，类似于人类的思考。



我们可以通过上面的例子，来观察深度学习技术和人类是如何识别猫的，以及它们的过程有哪些区别。

2012 年，Google X 实验室宣布使用深度学习技术，让机器成功识别了图片中的猫。它们使用了 1000 台服务器，16000 个处理器，连接成一个 10 亿节点的人工智能大脑。这个系统阅读了 1000 万张从 YouTube 上抽取的图片，最终成功识别出这个图片中有没有猫。

我们再来看看人类是如何做的。对于一个 3 岁的小朋友，我们只需要给他看几张猫的图片，他就能够很快识别出不同图片中的猫，而这背后其实就是大脑对于这些知识的推理。

2011 年，Science 上有一篇非常出名的论文叫《How to Grow a Mind》。这篇论文的作者来自于 MIT、CMU、UC Berkeley、Stanford 等美国名校的教授。在这篇论文里，最重要的一个结论就是：**如果我们的思维能够跳出给定的数据，那么必须有 Another Source Of Information 来 Make Up The Difference。**

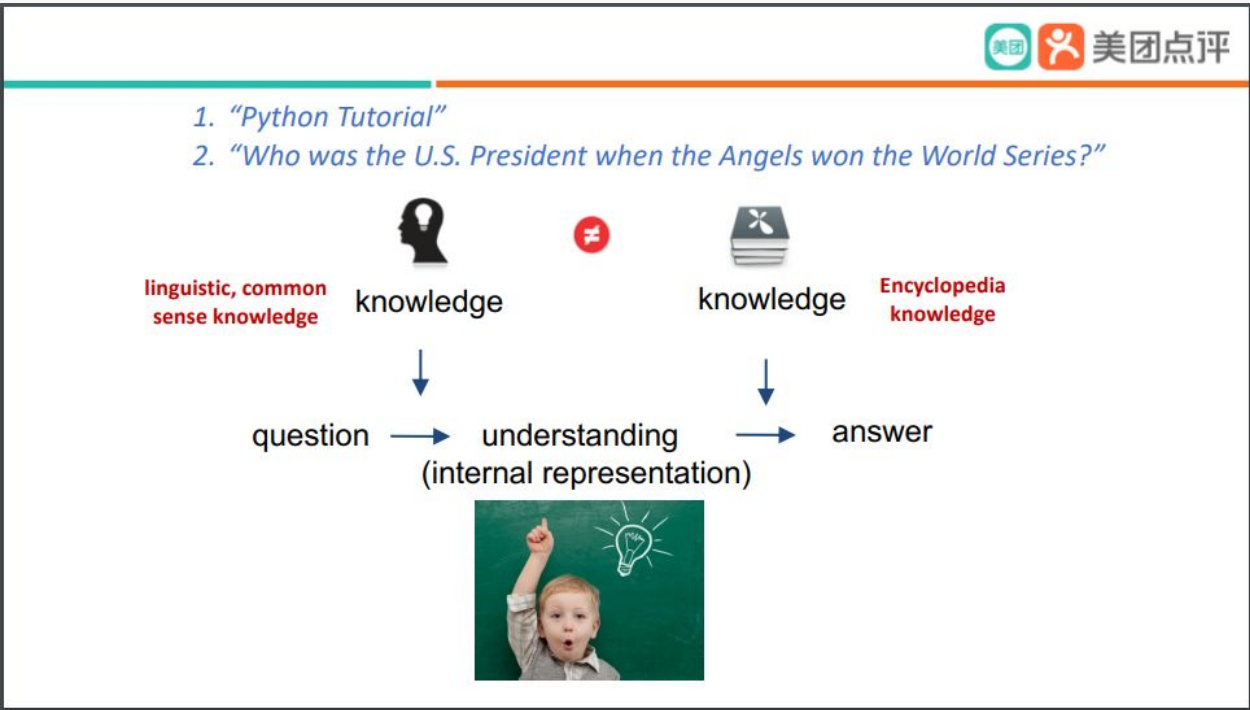
这里的知识语言是什么？对于人类来讲，其实就是我们从小到大接受的学校教育，报纸上、电视上看到的信息，通过社交媒体，通过与其他人交流，不断积累起来的知识。

近年来，不管是学术界还是工业界都纷纷构建自家的知识图谱，有面向全领域的知识图谱，也有面向垂直领域的知识图谱。其实早在文艺复兴时期，培根就提出了“知识就是力量”，在当今人工智能时代，各大科技公司更是纷纷提出：**知识图谱就是人工智能的基础。**

全球的互联网公司都在积极布局知识图谱。早在 2010 年微软就开始构建知识图谱，包括 Satori 和 Probase。2012 年，Google 正式发布了 Google Knowledge Graph，现在规模已经达到 700 亿左



右。目前微软和 Google 拥有全世界最大的通用知识图谱，Facebook 拥有全世界最大的社交知识图谱，而阿里巴巴和亚马逊则分别构建了商品知识图谱。



如果按照人类理解问题和回答问题这一过程来进行区分，我们可以将知识图谱分成两类。我们来看这样一个例子，如果用户看到这样一个问题，“Who was the U.S. President when the Angels won the World Series?”相信所有的用户都能够理解这个问题，也就是当 Angels 队赢了 World Series 的时候，谁是美国的总统？

这是一个问题理解的过程，它所需要的知识通常我们称之为 Common Sense Knowledge（常识性知识）。另外一方面，很多网友可能回答不出这个问题，因为它需要另外一个百科全书式的知识。

因此，我们将知识图谱分成两大类，一类叫 Common Sense Knowledge Graph（常识知识图谱），另外一类叫 Encyclopedia Knowledge Graph（百科全书知识图谱）。这两类知识图谱有很明显的区别。针对 Common Sense Knowledge Graph，通常而言，我们会挖掘这些词之间的 Linguistic Knowledge；对于 Encyclopedia Knowledge Graph，我们通常会在乎它的 Entities 和这些 Entities 之间的 Facts。

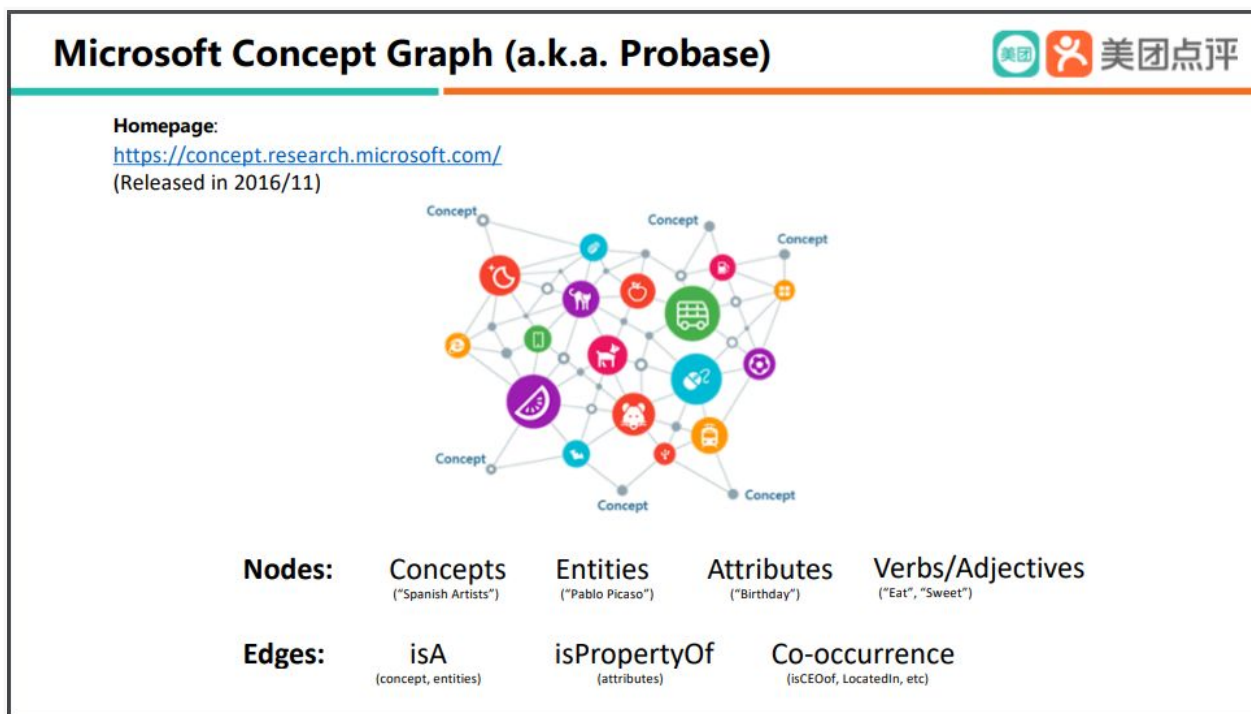
对于 Common Sense Knowledge Graph，一般而言我们比较在乎的 Relation 包括 isA Relation、isPropertyOf Relation。对于 Encyclopedia Knowledge Graph，通常会预定义一些谓词，比如说 DayOfBirth、LocatedIn、SpouseOf 等等。

对于 Common Sense Knowledge Graph 通常带有一定的概率，但是 Encyclopedia Knowledge Graph 通常就是“非黑即白”，那么构建这种知识图谱时，我们在乎的就是 Precision（准确率）。

Common Sense Knowledge Graph 比较有代表性的工作包括 WordNet、KnowItAll、NELL 以及 Microsoft Concept Graph。而 Encyclopedia Knowledge Graph 则有 Freepase、Yago、Google Knowledge Graph 以及正在构建中的“美团大脑”。

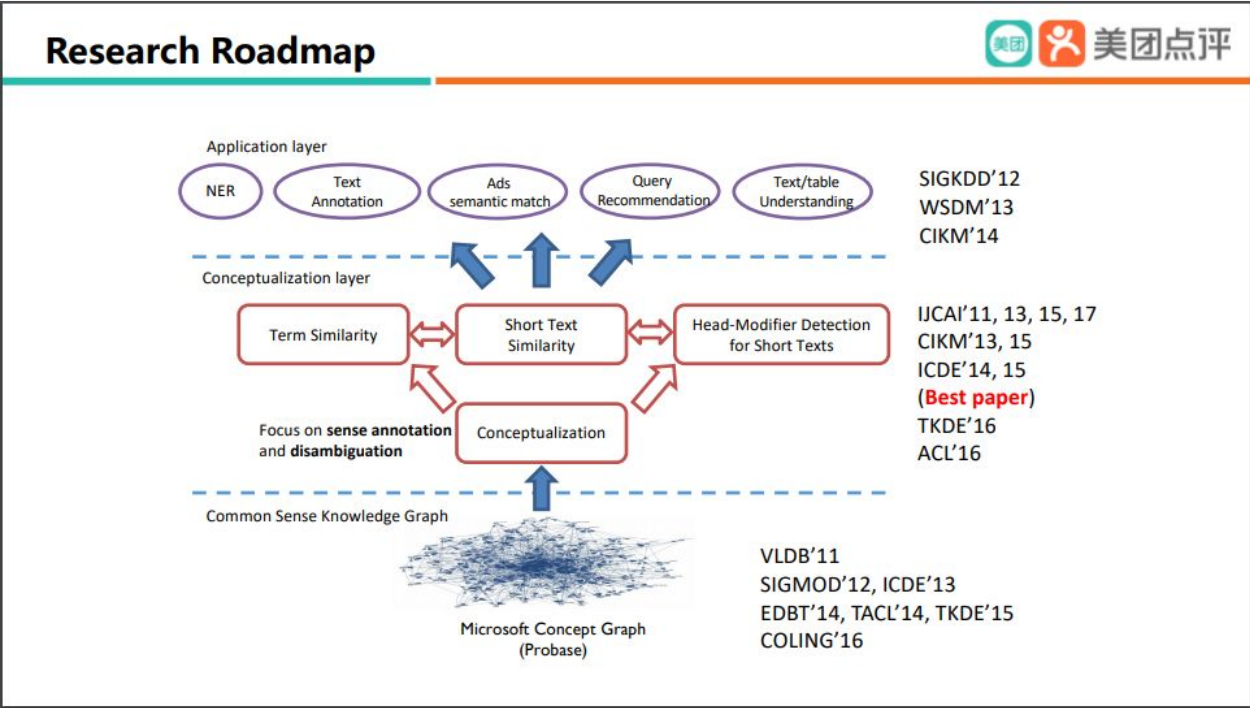
这里跟大家介绍两个代表性工作：1) Common Sense Knowledge Graph：Probase；2) Encyclopedia Knowledge Graph：美团大脑。

## 常识性知识图谱（Common Sense Knowledge Graph）



Microsoft Concept Graph 于 2016 年 11 月正式发布，但是它早在 2010 年就已经开始进行研究，是一个非常大的图谱。在这个图谱里面有上百万个 Nodes（节点），这些 Nodes 有 Concepts（概念），比如说 Spanish Artists（西班牙艺术家）；有 Entities（实体），比如说 Picasso（毕加索）；有 Attributes（属性），比如 Birthday（生日）；有 Verbs（动词），有 Adjectives（形容词），比如说 Eat、Sweet。也有很多很多的边，最重要的边，是这种 isA 边，比如说 Picasso，还有 isPropertyOf 边。对于其他的 Relation，我们会统称为 Co-occurrence。

这是我们在微软亚洲研究院期间对 Common Sense Knowledge Graph 的 Research Roadmap（研究



路线图)。当我们构建出 Common Sense Knowledge Graph 之后，重要的是在上面构建各种各样的模型。我们提出了一些模型叫 Conceptualization（概念化模型），它能够支持 Term Similarity、Short Text Similarity 以及 Head-Modifier Detection，最终支持各种应用，比如 NER、文本标注、Ads、Query Recommendation、Text Understanding 等等。

到底什么是 Short Text Understanding？常识怎么用在 Text Understanding 中？下面我们可以看一些具体的例子：

# Add Common Sense to Computing

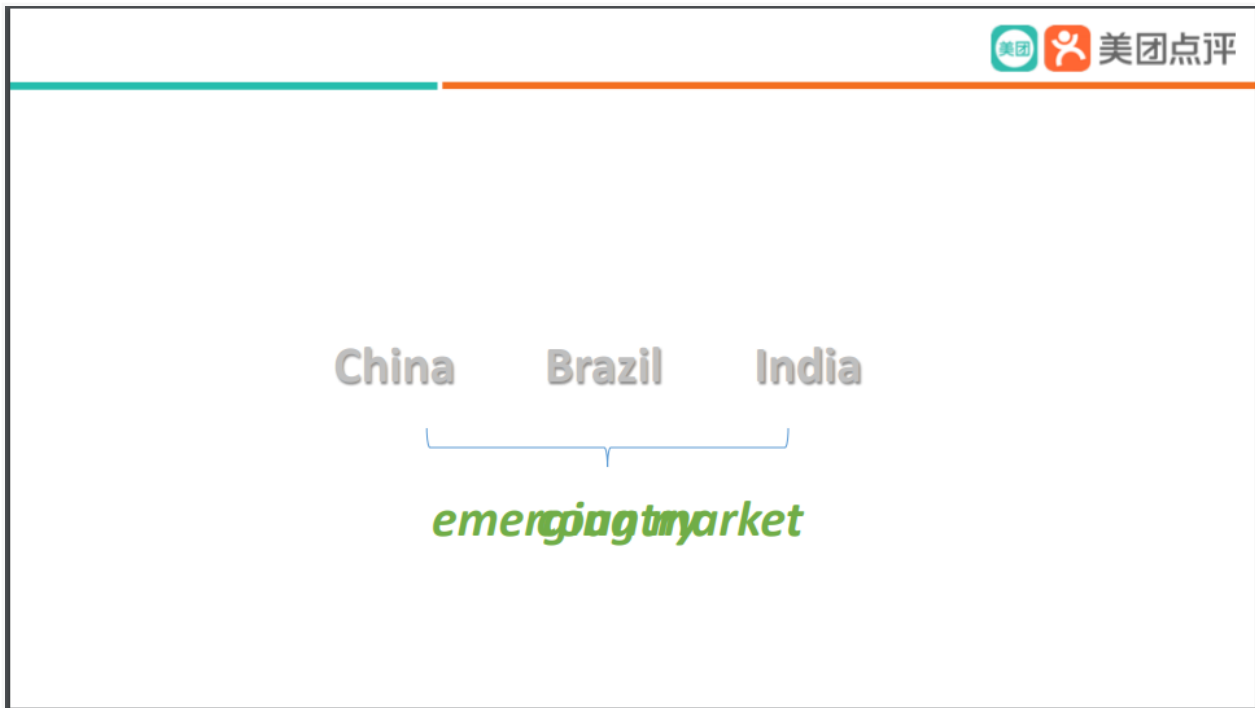
Pablo Picasso

25 Oct 1881

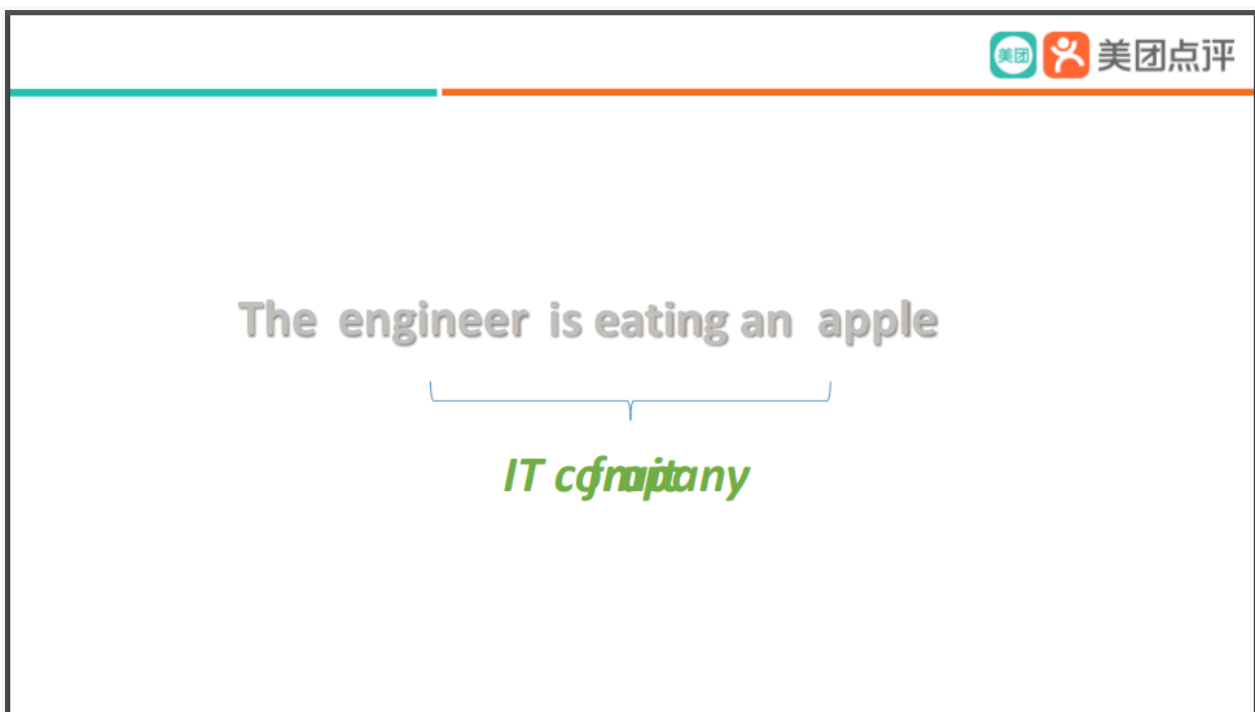
Spanish

当大家看到上面中间的文本时，相信所有人都能够认出这应该是一个日期，但是大家没办法知道这个日期代表什么含义。但如果我们再多给一些上下文信息，比如 Picasso、Spanish 等等，大家对这个日期就会有一些常识性的推理。我们会猜测这个日期很可能是 Picasso 的出生日期，或者是去世日期，这就是常识。





比如说当我们给定 China 和 India 这两个 Entity 的时候，我们的大脑就会做出一些常识性的推理，我们会认为这两个 Entity 在描述 Country。如果再多给一个 Entity: Brazil，这时候我们通常会想到 Emerging Market。如果再加上 Russia，大家可能就会想到“金砖四国”或者“金砖五国”。所有这一切就是常识性的推理。



再比如，当我们看到 Engineer 和 Apple 的时候，我们会对 Apple 做一些推理，认为它就是一个 IT Company，但是如果再多给一些上下文信息，在这个句子里面由于 eating 的出现，我相信大家的大脑也会一样地做出常识推理，认为这个 Apple 不再是代表 Company，而是代表 Fruit。

所以，这就是我们提出来的 Conceptualization Model，它是一个 Explicit Representation。我们希望它能够把 Text，尤其是 Short Text，映射到 Millions Concepts，这样的 Representation 能够比较容易让用户进行理解，同时能够应用到不同场景当中。

## Conceptualization

• **Conceptualization:** An **explicit** representation for the **short text**

$$P(\text{concept} \mid \text{short text})$$

• **Short text** is *sparse, noisy, and ambiguous*

• **Explicit** means

- Conceptualization results can be *easily understood* by human beings
- Conceptualization model can be *easily customized* for different scenarios

## Conceptualization

• **Conceptualization:** An **explicit** representation for the **short text**

ShortText: pear apple

[Show Parameters](#)

Elapsed Time - 00:00:00.0140014

pear [25/fruit]		apple [25/fruit]	
25/fruit	0.5724769	25/fruit	0.5718007
fruit	0.0562338	fruit	0.1323192
fresh fruit	0.02918897	fresh fruit	0.05945546
tree fruit	0.02126049	tree fruit	0.01657355
dried fruit	0.01165293	dried fruit	0.01593271
seasonal fruit	0.01160144	seasonal fruit	0.01553914
juice	0.01062348	juice	0.01540397
hard fruit	0.01011614	fruit juice	0.01309836
climacteric fruit	0.009254614	hard fruit	0.01297377
fruit juice	0.009048668	climacteric fruit	0.01062575
sweet fruit	0.008924312	sweet fruit	0.01033749
9405/food	0.1058999	9405/food	0.1241537
food	0.03129783	food	0.06844553
high fiber food	0.008663075	high fiber food	0.01028461
ingredient	0.007349567	ingredient	0.009757149
high fiber food	0.004695967	fresh food	0.004133756
fresh food	0.004038723	hard food	0.004045118
hard food	0.00355713	high-fiber food	0.00373333

"pear apple"

ShortText: ipad apple

[Show Parameters](#)

Elapsed Time - 00:00:00.6470667

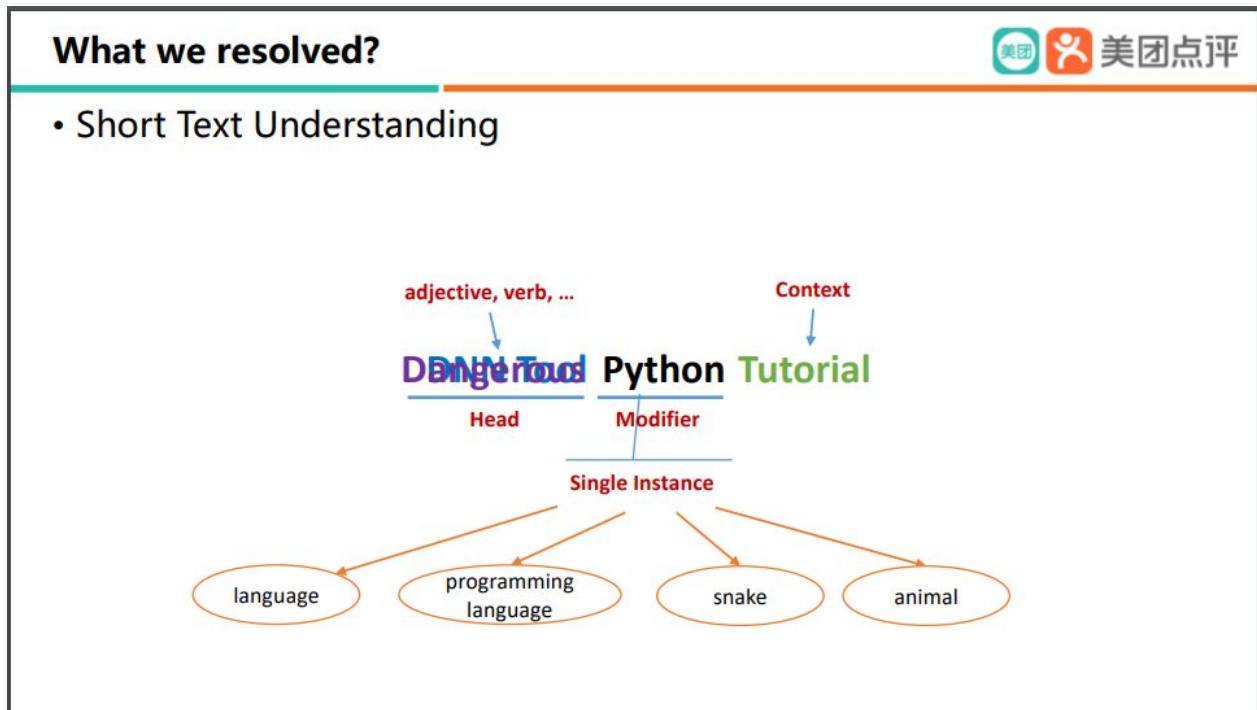
ipad [15/mobile device/device]		apple [1/technology company/company]	
15/mobile device/device	0.8672805	1/technology company/company	0.9623328
mobile device	0.01889746	technology company	0.005182603
apple device	0.01723156	computer manufacturer	0.005060604
tablet device	0.01718674	tech company	0.004826283
ios device	0.01706666	innovative company	0.00475833
portable device	0.01540841	computer company	0.004570935
gadgets	0.0121459	tech giant	0.004452952
handheld device	0.0105827	technology giant	0.004435323
digital device	0.01037961	successful company	0.00422191
multimedia device	0.00983045	tech stock	0.004143819
wireless device	0.009496055	software company	0.004138673
3/apple product/product	0.1443738	1853/laptop brand/top brand name/brand	0.82506031
apple product	0.01561871	laptop brand	0.0004202249
apple's product	0.005314387	iconic brand	0.0004185871
electronic product	0.00442585	great brand	0.0004146745
hot new apple product	0.004225718	global brand	0.0004121742
apple's high technology product	0.00422373	big brand	0.0004073361
popular apple product	0.004221935	strong brand	0.0003894416
iconic product	0.004196884	popular brand	0.0003739756
revolutionary product			0.0003660904
digital product			0.0003555858
popular product			0.0003552496

"ipad apple"

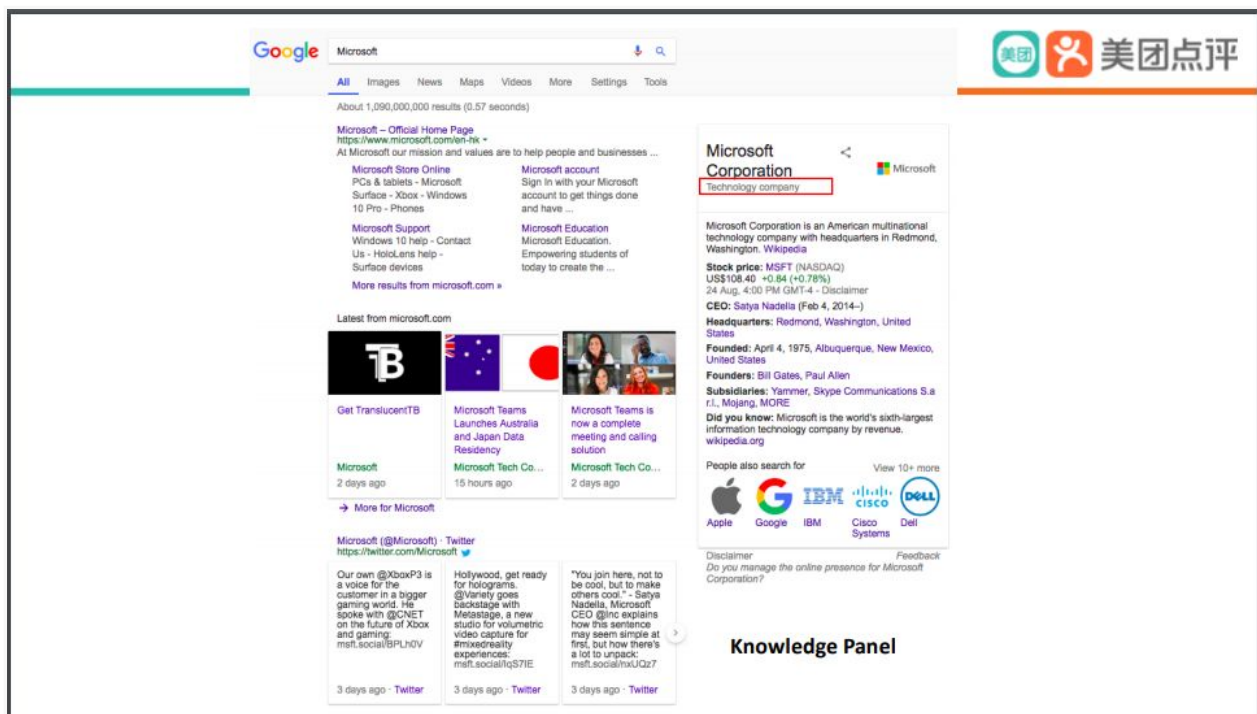
在这一页 PPT 中，我们展示了 Conceptualization 的结果。当输入是 Pear 和 Apple 的时候，那么我们会将这个 Apple 映射到 Fruit。但是如果是 iPad Apple 的时候，我们会将它映射到 Company，同时大家注意这并不是唯一的结果，我们实际上是会被映射到一个 Concept Vector。这个 Concept Vector 有多大？它是百万级维度的 Vector，同时也是一个非常 Sparse 的一个 Vector。

通过这样的一个人 Conceptualization Model，我们能够解决什么样的文本理解问题？我们可以看这样一个例子。比如说给定一个非常短的一个文本 Python，它只是一个 Single Instance，那么我们会希望将它映射到至少两大类的 Concept 上，一种可能是 Programming Language，另外一种 Snake。当它有一些 Context，比如说 Python Tutorial 的时候，那么这个时候 Python 指的应该是 Programming Language，如果当它有其他的 Adjective、Verb，比如有 Dangerous 时，这时候我们就会将 Python 理解为 Snake。





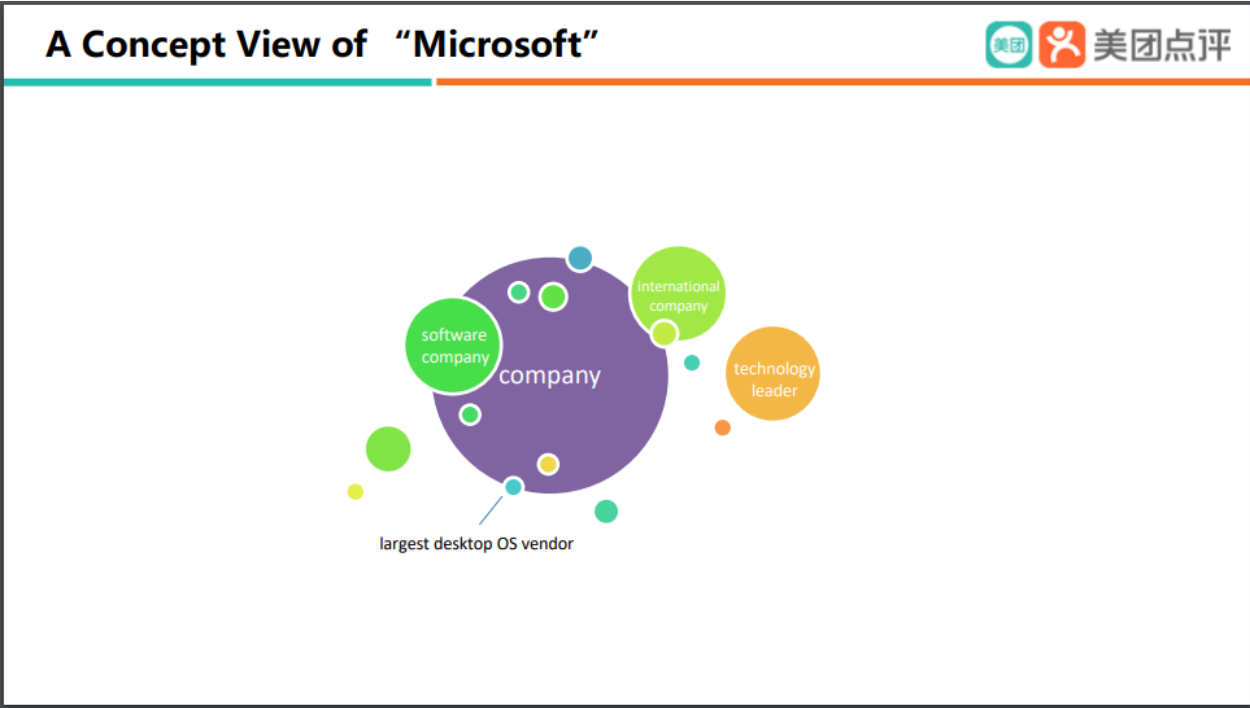
同时如果在一个文本里面包含了多个的 Entity，比如说 DNN Tool、Python，那么我们希望能够检测出在这个文本里面哪一个是比较重要的 Entity，哪一个是用来做限制的 Entity。



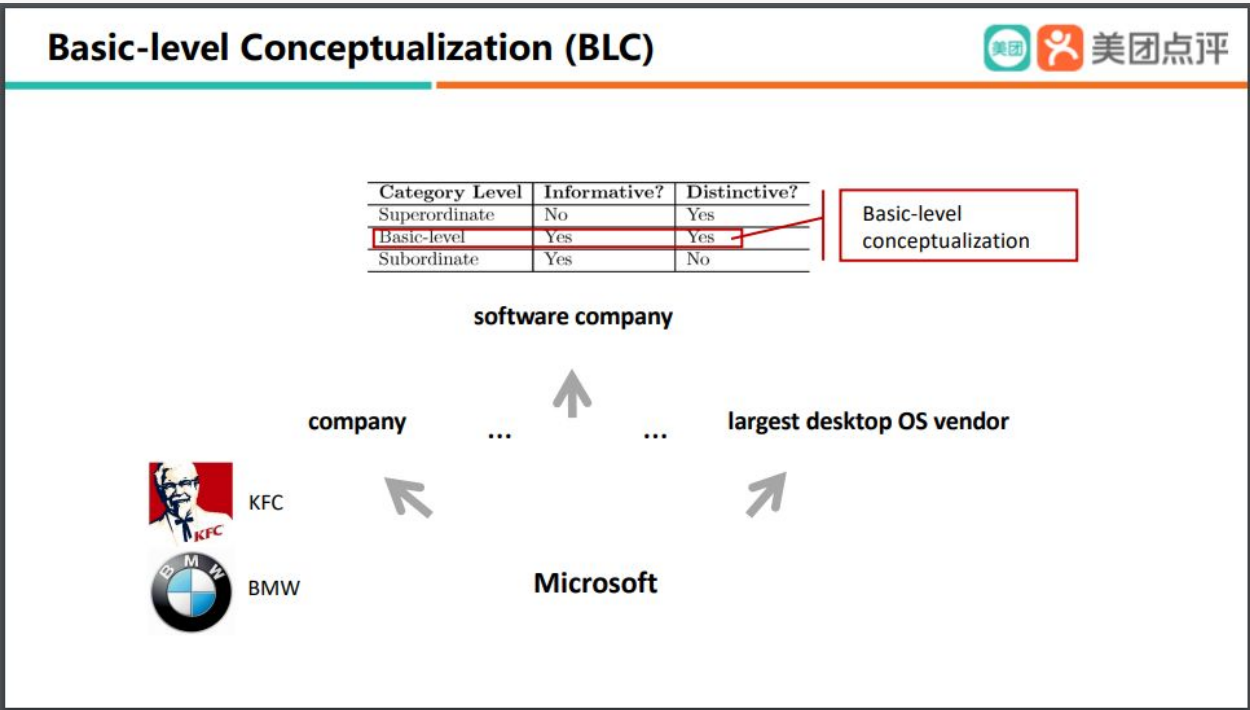
下面我们将简单地介绍一下，具体应该怎么去做。当我们在 Google 里搜一个 Single Instance 的时候，通常在右侧会出现这个 Knowledge Panel。对于 Microsoft 这样一个 Instance，我们可以看到这个红色框所框出来的 Concept，Microsoft 指向的是 Technology Company，这背后是怎么实现的？

我们可以看到，Microsoft 实际上会指向非常非常多的 Concept，比如说 Company，Software Company，Technology Leader 等等。我们将它映射到哪一个 Concept 上最合适？

如果将它映射到 Company 这个 Concept 上，很显然它是对的，但是我们却没办法将 Microsoft 和 KFC、BMW 这样其他类型的产品区分开来。另外一方面，如果我们将 Microsoft 映射到 Largest Desktop OS Vendor 上，那么这是一个非常 Specific 的 Concept，这样也不太好，为什么？因为这



个 Concept 太 Specific，太 Detail，它可能只包含了 Microsoft 这样一个 Entity，那么它就失去了 Concept 的这种抽象能力。



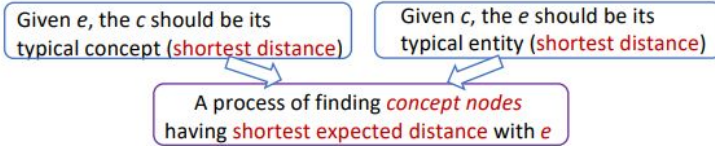
所以我们将 Microsoft 映射到一个既不是特别 General（抽象），又不是一个特别 Specific（具体）的 Concept 上。在语言学上，我们将这种映射称之为 Basic-level，我们将整个映射过程命名为 Basic-level Conceptualization。

我们提出了一种计算 Basic-level Conceptualization 的方法，其实它非常简单而且非常有效。就是将两种的 Typicality 做了一些融合，同时我们也证明了它们跟 PMI 和 Commute Time 之间的一些关联。并且在一个大规模的数据集上，我们通过 Precision 和 NDCG 对它们进行了评价。最后证明，我们所提出来的 Scoring 方法，它在 NDCG 和 Precision 上都能达到比较好的结果。最重要的是，它在理论上是能够对 Basic-Level 进行很好的解释。

## Using $Rep(e, c)$ for BLC



- Our measure  $Rep(e, c) = P(c|e) * P(e|c)$  means:



- (With PMI) If we take the logarithm of our scoring function, we get:

$$\log Rep(e, c) = \log P(c|e) * P(e|c) = \log \frac{P(e, c)}{P(e)} * \frac{P(e, c)}{P(c)} = \log \frac{P(e, c)^2}{P(e)P(c)} = PMI(e, c) + \log P(e, c) = PMI^2$$

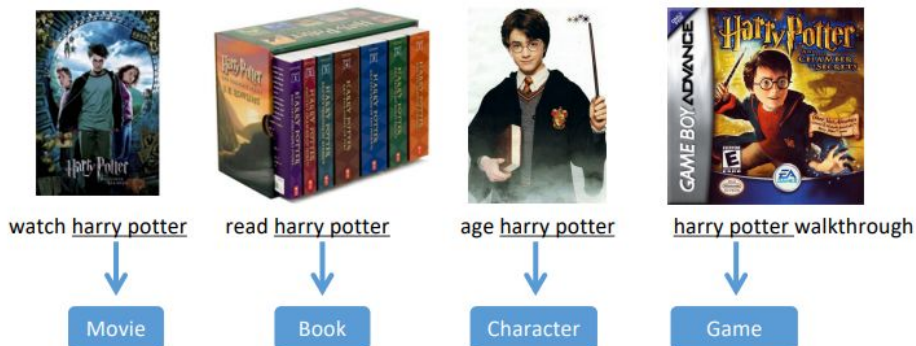
- (With Commute Time) The commute time between an instance  $e$  and a concept  $c$  is:

$$\begin{aligned} Time(e, c) &= \sum_{k=1}^{\infty} (2k) * P_k(e, c) = \sum_{k=1}^T (2k) * P_k(e, c) + \sum_{k=T+1}^{\infty} (2k) * P_k(e, c) \\ &\geq \sum_{k=1}^T (2k) * P_k(e, c) + 2(T+1) * (1 - \sum_{k=1}^T P_k(e, c)) = 4 - 2 * Rep(e, c) \end{aligned}$$

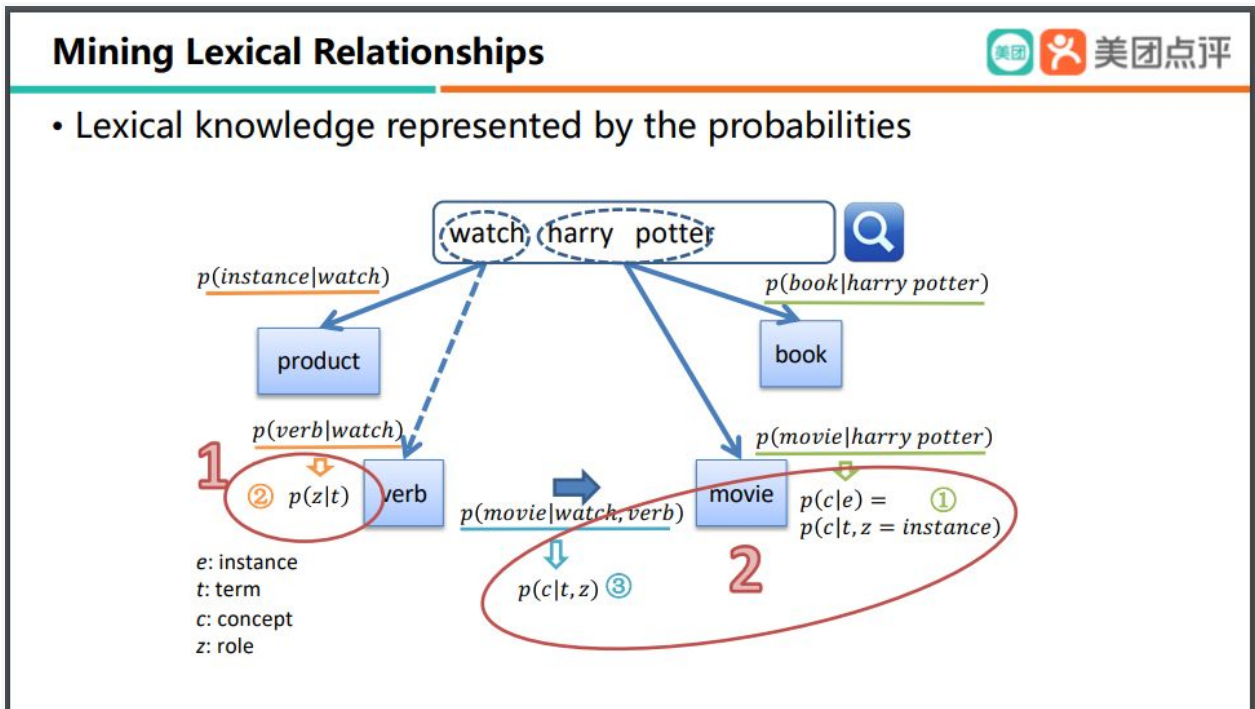
下面我们来看一下，当 Instance 有了一些 Context 之后，我们应该怎么去进行处理。我们通过一个例子，来简单地解释一下这背后最主要的思想。

比如说 iPad、Apple，其中 iPad 基本上是没有歧异的，它会映射到 Device、Product。但是对于 Apple 而言，它可能会映射到至少两类的 Concept 上，比如说 Fruit、Company。那么我们怎么用 iPad 对 Apple 做消歧呢？

方法其实也挺直观的。我们会通过大量的统计去发现像 iPad 这样的 Entity，通常会跟 Company、Product 共同出现。比如说 iPad 有可能会跟三星共同出现，有可能会跟 Google 共同出现，那么我们就发现它会经常跟 Brand、Company、Product 共同出现。于是我们就利用新挖掘出来的 Knowledge 对 Apple 做消歧，这就是背后最主要的思想。

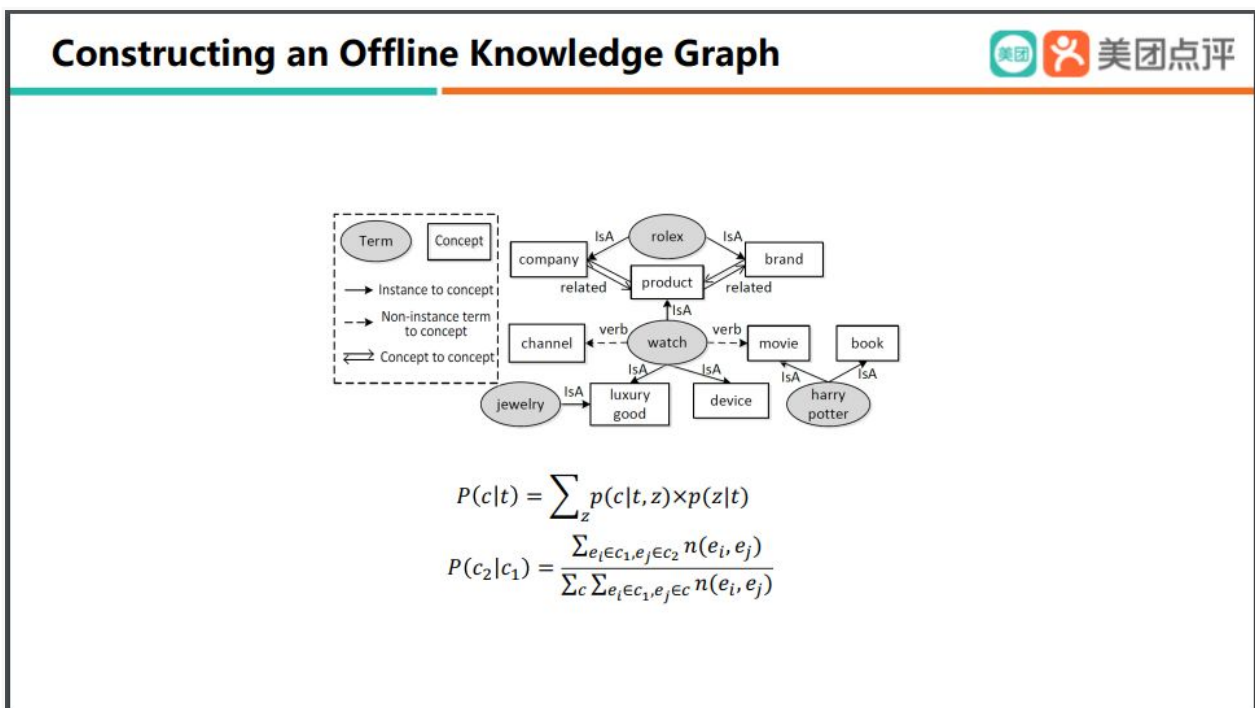


除了刚才这样一个 General Context 以外，在很多时候这些 Text 可能还会包含很多一些特殊的类型，比如说 Verb、Adjective。具体而言，我们希望在看到 Watch Harry Potter 时，能够知道 Harry Potter 是 Movie，当我们看到 Read Harry Potter 时，能够知道 Harry Potter 是 Book。同样的，Harry Potter 还有可能是一个角色名称，或者是一个游戏名称。



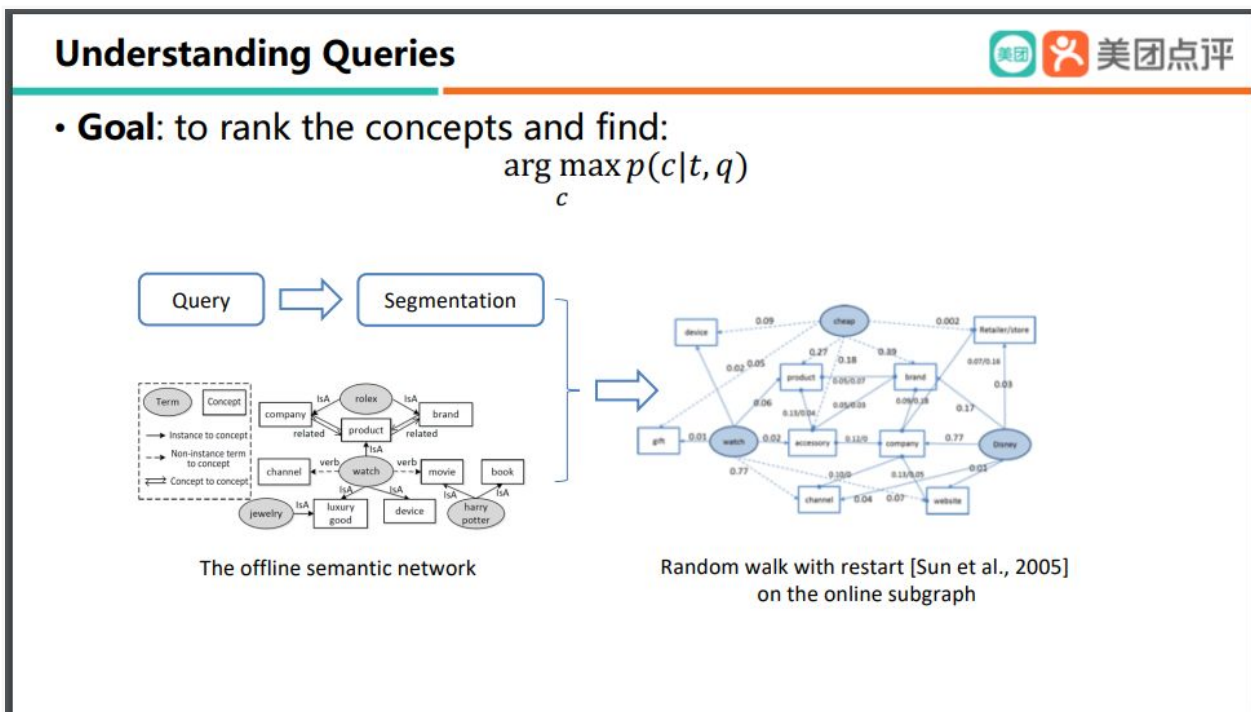
那么我们来看一看应该怎样去解决这样一件事情。当我们看到 Watch Harry Potter 时，我们首先要知道，Harry Potter 有可能是一本 Book，也有可能是一部 Movie。我们可以算出一个先验概率，这通常要通过大规模的统计。同时我们要知道，Watch 它有可能是一个名词，同时它也有可能是一个动词，并且我们还需要去挖掘，当 Watch 作为动词的时候，它和 Movie 有非常紧密的关联。

所以我们本质上是要去做一些概率上的推理，不仅要条件概率做非常细粒度的分解，最后还要做概率计算。

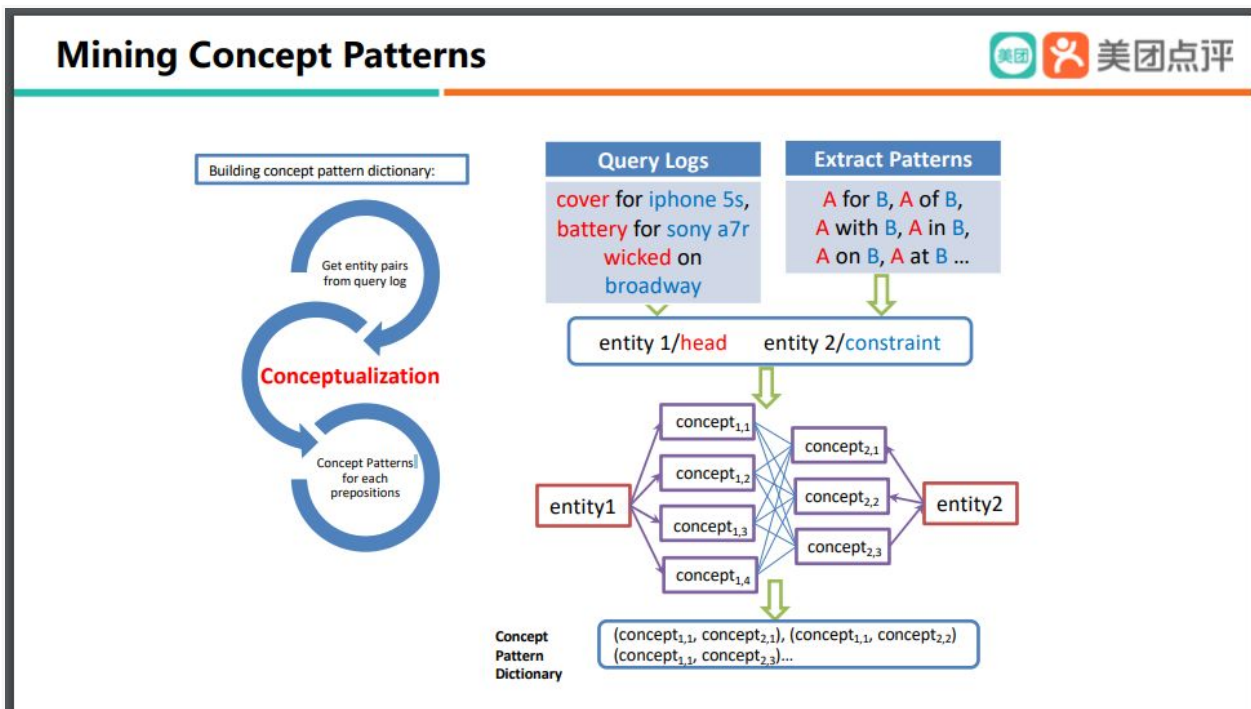




通过概率计算的方法，我们实际上就可以构建出一个非常大的离线知识图谱，那么我们在这个上面，就可以有很多的 Term，以及它们所属的一些 Type，以及不同 Term 之间的一些关联。





当我们用这样一个非常大的离线知识图谱来做 Text Understanding 的时候，我们可以首先将这个 Text 进行分割处理，在分割之后，我们实际上是可以从这个非常大的离线知识图谱中截取出它的一个子图。最后我们使用了 Random Walk With Restart 的模型，来对这样一个在线的 Subgraph 进行分类。



我们再来看一下，如果一个文本里包含了 Multiple Entities，要怎样处理？我们需要做知识挖掘，怎么做？首先我们可以得到非常多的 Query Log，然后我们也可以去预定一些 Pattern，通过这种 Pattern 的定义，可以抽取出非常多 Entity 之间 Head 和 Modifier 这样的 Relation，那么在接下来我们可以将这些 Entity 映射到 Concept 上，之后得到一个 Pattern。


## Why Concepts Can't Be Too General



 美团点评

- It may **cause too many concept pattern conflicts**: can't distinguish head and modifier for general concept pairs



	Head	Modifier
Derived Concept Pattern	device	company
Supporting Entity Pairs	iphone 4	verizon
	modem	comcast
	wireless router	comcast
	iphone 4	tmobile

	Head	Modifier
Derived Concept Pattern	company	device
Supporting Entity Pairs	amazon books	kindle
	netflix	touchpad
	skype	windows phone
	netflix	ps3



在这个过程中，我们要将 Entity 映射到 Concept 上，那么这就是前面所提到的 Conceptualization。我们希望之后的映射不能太 General，避免 Concept Pattern 冲突。

## Why Concepts Can't Be Too Specific



 美团点评

- It may generate concepts **with less representation**

...	...
device	largest desktop OS vendor
device	largest software development company
device	largest global corporation
device	latest windows and office provider
...	...

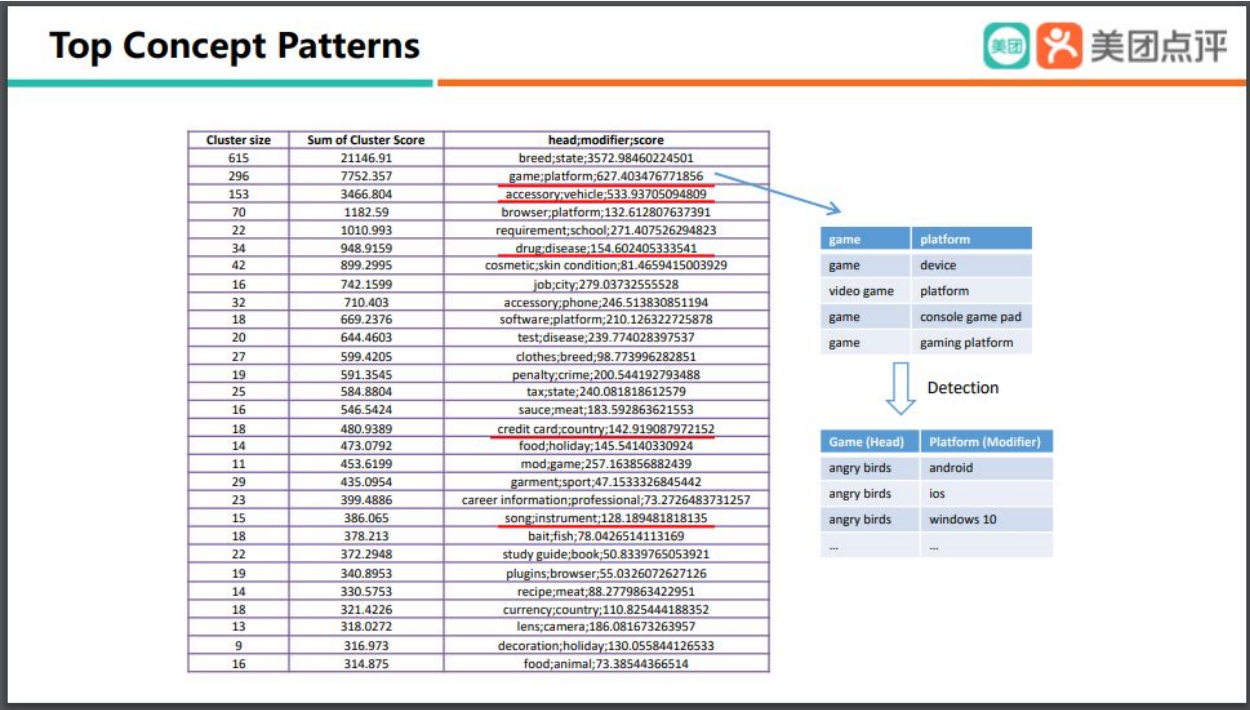
- Concept level may **regress to entity level**
  - Large storage space: up to (million \* million) patterns

We should use Basic-level Conceptualization (BLC)

但是它也不能太 Specific，因为如果太 Specific，可能就会缺少表达能力。最坏的情况，它有可能就会退化到 Entity Level，而 Entity 至少都是百万的规模，那么整个 Concept Patterns 就有可能变成百万乘以百万的级别，显然是不可用的。

所以我们就用到了前面介绍的 Basic-Level Conceptualization 的方法，将它映射到一个既不是特别 General，也不是特别 Specific 的 Concept 上。

大家可以看一下我们能够挖掘出来的一些 Top 的 Concept Patterns，比如说 Game 和 Platform，就是一个 Concept 和一个 Pattern。它有什么用？举一个具体的例子，当用户在搜 Angry Birds、iOS



的时候，我们就可以知道用户想找的是 Angry Birds 这款游戏，而 iOS 是用来限制这款游戏的一个 Platform。苹果公司每年都会推出新版本的 iOS，那么我们挖掘出这样的 Concept Pattern 之后，不管苹果出到 iOS 15或者 iOS 16，那么我们只需要将它们映射到 Platform，那么我们的 Concept Patterns 就仍然有效，这样可以很容易地进行知识扩展。

所以 Common Sense Knowledge Mining 以及 Conceptualization Modeling，可以用在很多的应用上，它可以用来算 Short Text Similarity，可以用来做 Classification、Clustering，也可以用来做广告的 Semantic Match、Q/A System、Chatbot 等等。

## 美团大脑——百科全书式知识图谱（Encyclopedia Knowledge Graph）

在介绍完 Common Sense Knowledge Graph 之后，给大家介绍一下 Encyclopedia Knowledge Graph。这是美团的知识图谱项目——美团大脑。

美团大脑是什么？美团大脑是我们正在构建中的一个全球最大的餐饮娱乐知识图谱。我们希望能够充分地挖掘关联美团点评各个业务场景里的公开数据，比如说我们有累计 40 亿的用户评价，超过 10 万条个性化标签，遍布全球的 3000 多万商户以及超过 1.4 亿的店菜，我们还定义了 20 级细粒度的情感分析。

我们希望能够充分挖掘出这些元素之间的关联，构建出一个知识的“大脑”，用它来提供更加智能的生活服务。

我们简单地介绍一下美团大脑是如何进行构建的。我们会使用 Language Model（统计语言模型）、Topic Model（主题生成模型）以及 Deep Learning Model（深度学习模型）等各种模型，希望能够做到商家标签的挖掘，菜品标签的挖掘和情感分析的挖掘等等。

为了挖掘商户标签，首先我们要让机器去阅读评论。我们使用了无监督和有监督的深度学习模型。

无监督模型我们主要用了LDA，它的特点是成本比较低，无需标注的数据。当然，它准确性会比较不可控，同时对挖掘出来的标签我们还需要进行人工的筛选。至于有监督的深度学习模型，那么我们用了



## 知识提炼：美团大脑的构建



### • 大数据：

- 累计40亿的公开评价数据
- 3,450万全球商家
- 1.4亿店菜
- 10万个个性化标签



### • NLP算法模型：

- 统计语言模型 (Language Model)
- 主题生成模型 (Topic Model)
- 深度学习模型 (Deep Learning Model)



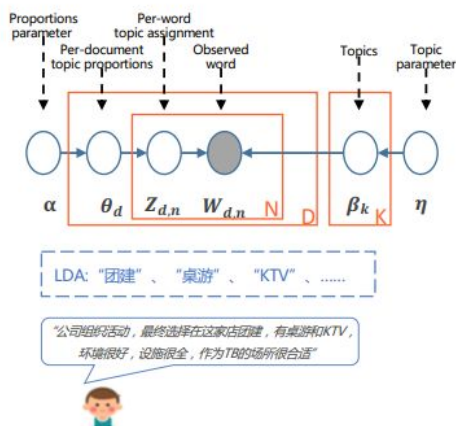
## 商户标签挖掘



### • 步骤一：用户评论内容挖掘标签 (无监督模型+有监督深度学习模型)

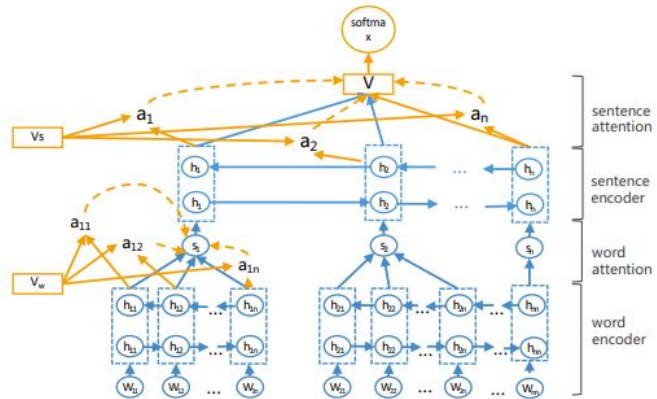
#### - 无监督模型: LDA

- 成本低，无需标注数据
- 准确性不可控，需要严格筛选分数高的标签



#### - 有监督深度学习模型：LSTM

- 成本高，需要大量标注数据
- 准确性较高



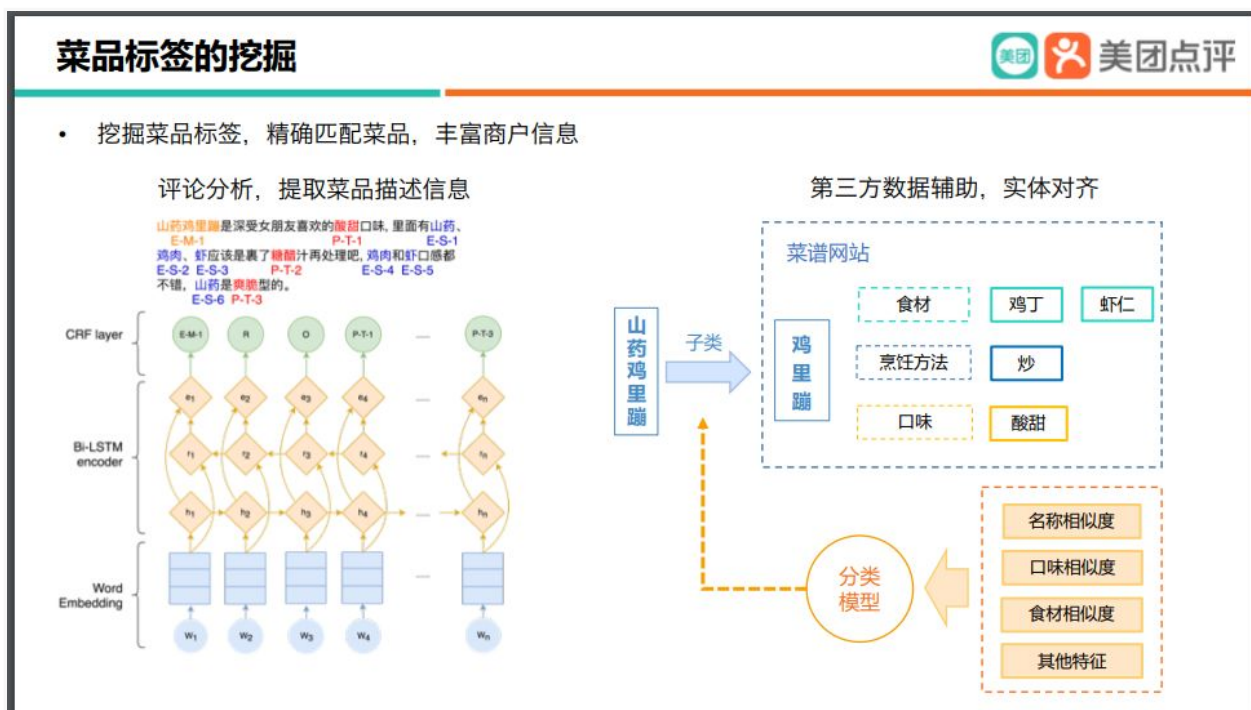
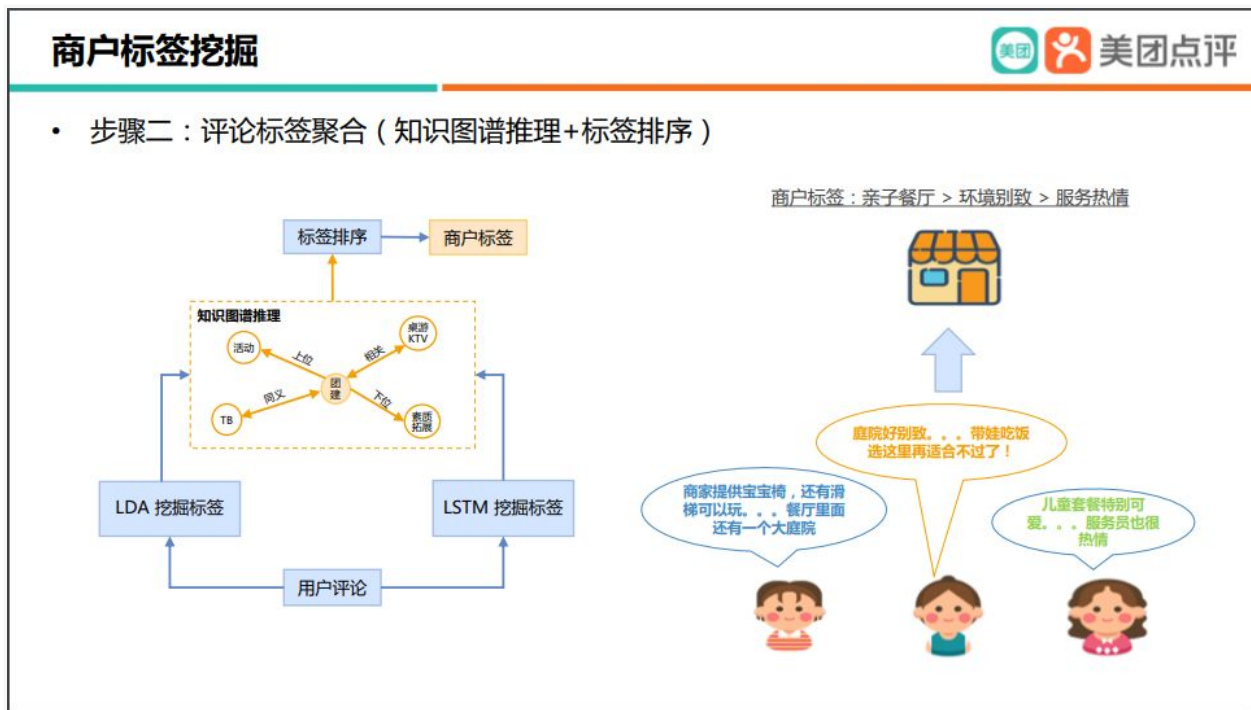
LSTM，它的特点是需要比较大量的标注数据。

通过这两种模型挖掘出来的标签，我们会再加上知识图谱里面的一些推理，最终构建出商户的标签。

如果这个商户有很多的评价，都是围绕着宝宝椅、带娃吃饭、儿童套餐等话题，那么我们就可以得出很多关于这个商户的标签。比如说我们可以知道它是一个亲子餐厅，它的环境比较别致，服务也比较热情。

下面介绍一下我们如何对菜品进行标签的挖掘？我们使用了 Bi-LSTM 以及 CRF 模型。比如说从这个评论里面我们就可以抽取出这样的 Entity，再通过与其他的一些菜谱网站做一些关联，我们就可以得到它的食材、烹饪方法、口味等信息，这样我们就为每一个店菜挖掘出了非常丰富的口味标签、食材标签等各种各样的标签。

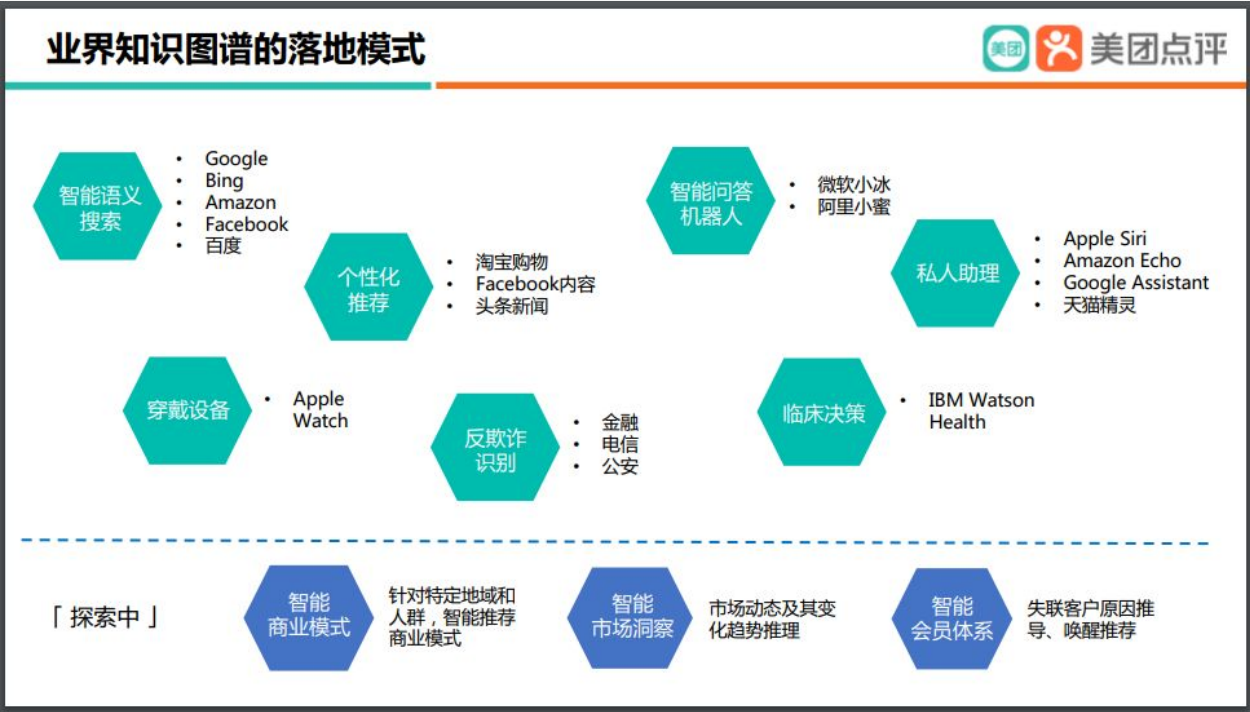
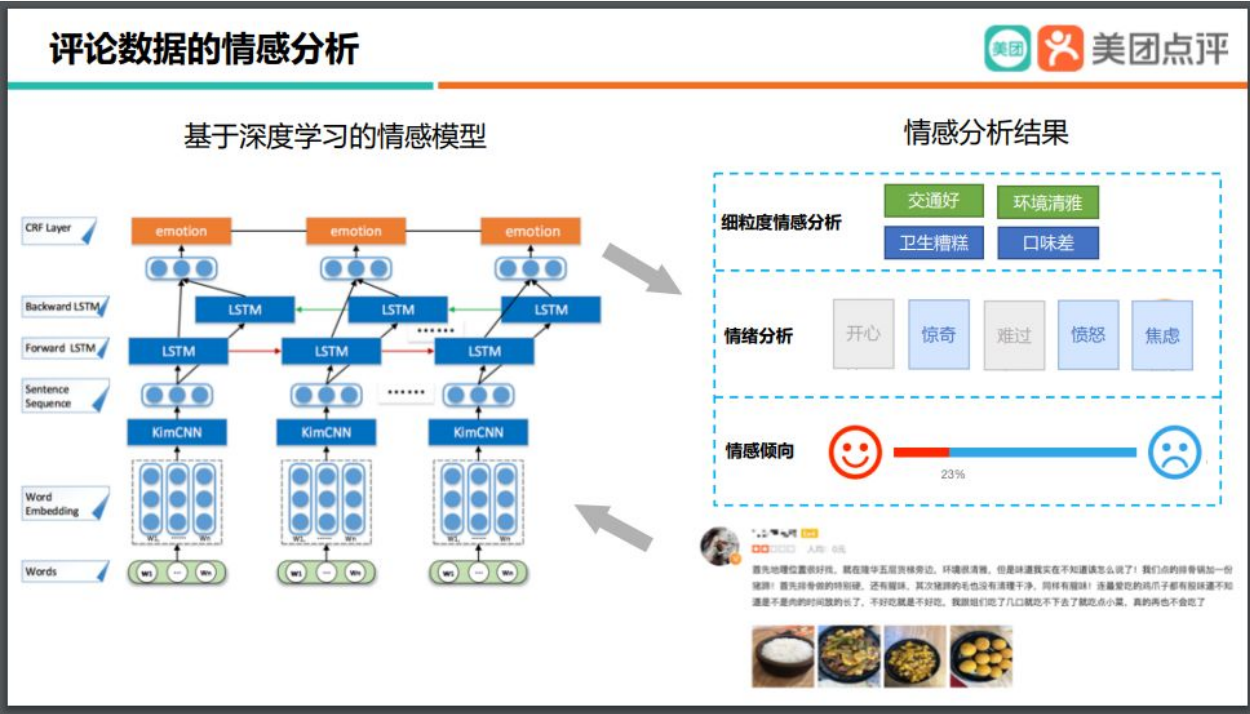




下面再简单介绍一下，我们如何进行评论数据的情感挖掘。我们用的是 CNN+LSTM 的模型，对于每一个用户的评价我们都能够分析出他的一些情感的倾向。同时我们也正在做细粒度的情感分析，我们希望能够通过用户短短的评价，分析出他在不同的维度，比如说交通、环境、卫生、菜品、口味等方面的不同的情感分析的结果。值得一提的是，这种细粒度的情感分析结果，目前在全世界范围内都没有很好的解决办法，但是美团大脑已经迈出了非常重要的一步。

下面介绍一下我们的知识图谱是如何进行落地的。目前业界知识图谱已经有非常多的成熟应用，比如搜索、推荐、问答机器人、智能助理，包括在穿戴设备、反欺诈、临床决策上都有非常好的应用。同时业界也有很多的探索，包括智能商业模式、智能市场洞察、智能会员体系等等。

如何用知识图谱来改进我们的搜索？如果大家现在打开大众点评，搜索某一个菜品时，比如说麻辣小龙虾，其实我们的机器是已经帮大家提前阅读了所有的评价，然后分析出提供这道菜品的商家，我们还会



根据用户评论的情感分析结果来改进这些搜索排序。

此外，我们也将它用在商圈的个性化推荐。当大家打开大众点评时，如果你现在位于某一个商场或者商圈，那么大家很快就能看到这个商场或者商圈的页面入口。当用户进入这个商场和商户页面时，通过知识图谱，我们就能够提供“千人千面”的个性化排序和个性化推荐。

在这背后其实使用了一个“水波”的深度学习模型，关于这个深度学习模型更详细的介绍，大家可以参见我们在 CIKM 上的一篇文章。

所有的这一切，其实还有很多的技术突破等待我们去解决。比如整个美团大脑的知识图谱在百亿的量级，这也是世界上最大的餐饮娱乐知识图谱，为了支撑这个知识图谱，我们需要去研究千亿级别的图存储和计算引擎技术。我们也正在搭建一个超大规模的 GPU 集群，来支持海量数据的深度学习算法。未



来, 当所有的这些技术都成熟之后, 我们还希望能够为用户提供“智慧餐厅”和“智能助理”的体验。

文章转载自 AI 科技大本营 (rgznai100), 部分内容有修正。

## 作者简介


- 仲远, 博士, 美团点评高级研究员、高级总监, 美团 AI 平台部 NLP 中心负责人、大众点评搜索智能中心负责人。加入美团点评前, 担任美国 Facebook 公司 Research Scientist, 负责 Facebook 产品级 NLP Service。在 Facebook 之前, 担任微软亚洲研究院的主管研究员, 负责微软研究院知识图谱项目和对话机器人项目。多年来专注于自然语言处理、知识图谱及其在文本理解方面的研究, 在国际顶级学术会议如 VLDB、ICDE、IJCAI、CIKM 等发表论文30余篇, 获得 ICDE 2015 最佳论文奖, 并是 ACL 2016 Tutorial “Understanding Short Texts”的主讲人, 出版学术专著3部, 获得美国专利5项。在 NLP 和 KG 研究领域及实际产品系统中均有丰富经验, 研究领域包括: 自然语言处理、知识图谱、深度学习、数据挖




### 美团大脑的隐性应用

基于图谱的水波深度学习模型：


- 实体连接
- 在知识图谱中扩散餐厅特性找到类似的餐厅作为推荐




#### 热门餐厅



麻辣小龙虾




老火锅




海鲜


#### 推荐餐厅



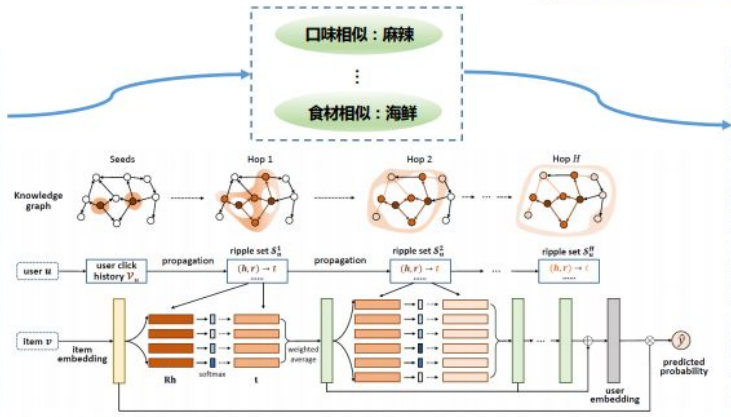
香辣蟹



干锅牛蛙





高人气川菜



The diagram illustrates the RippleNet architecture. It starts with a 'Knowledge graph' containing 'Seeds'. Propagation occurs through hops (Hop 1, Hop 2, ..., Hop  $l$ ) to form ripple sets  $S_1^i, S_2^i, \dots, S_l^i$ . These sets are processed by a neural network with layers for item embedding, softmax, and weighted average, leading to a 'predicted probability'.

Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, Minyi Guo, RippleNet: Propagating User Preferences on the Knowledge Graph for Recommender Systems, the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Turin, Italy, Oct. 2018

### 参考文献



- Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification, *IJCAI* 2017
- Efficiently Answering Technical Questions — A Knowledge Graph Approach, *AAAI* 2017
- Unsupervised Head-Modifier Detection in Search Queries, *TKDD* 2016
- Probabilistic Prototype Model for Serendipitous Property Mining, *COLING* 2016
- Understanding Short Texts, *ACL* 2016 (Tutorial)
- Understand Short Texts by Harvesting and Analyzing Semantic Knowledge, *TKDE* 2016
- Commonsense Causal Reasoning between Short Texts, *KR* 2016
- A Large Probabilistic Semantic Network based Approach to Compute Term Similarity, *TKDE* 2015
- An Inference Approach to Basic Level of Categorization, *CIKM* 2015
- Contextual Text Understanding in Distributional Semantic Space, *CIKM* 2015
- Query Understanding through Knowledge-Based Conceptualization, *IJCAI* 2015
- Short Text Understanding Through Lexical-Semantic Analysis, *ICDE* 2015 (Best Paper Award)
- Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees, *TKDE* 2015
- Concept-based Short Text Classification and Ranking, *CIKM* 2014
- Transfer Understanding from Head Queries to Tail Queries, *CIKM* 2014
- Overcoming Semantic Drift in Information Extraction, *EDBT* 2014
- Data Driven Metaphor Recognition and Explanation, *TACL* 2014
- Head, Modifier, and Constraint Detection in Short Texts, *ICDE* 2014
- Computing Term Similarity by Large Probabilistic isA Knowledge, *CIKM* 2013
- Semantic multi-dimensional scaling for open-domain sentiment analysis, in *IEEE Intelligent Systems*, 2013
- Context-Dependent Conceptualization, *IJCAI* 2013
- Automatic Extraction of Top-k Lists from the Web, *ICDE* 2013
- Attribute Extraction and Scoring: A Probabilistic Approach, *ICDE* 2013
- Identifying Users' Topical Tasks in Web Search, *WSDM* 2013
- Invited talk at UW on Probase
- Probase: A Probabilistic Taxonomy for Text Understanding, *SIGMOD* 2012
- Optimizing Index for Taxonomy Keyword Search, *SIGMOD* 2012
- Automatic Taxonomy Construction from Keywords, *KDD* 2012
- A System for Extracting Top-K Lists from the Web (demo), *KDD* 2012
- Understanding Tables on the Web, *ER* 2012
- Toward Topic Search on the Web, *ER* 2012
- Isanette: A Common and Common Sense Knowledge Base for Opinion Mining, *ICDM Workshops* 2011
- Web Scale Taxonomy Cleansing, *VLDB* 2011
- Short Text Conceptualization using a Probabilistic Knowledgebase, *IJCAI* 2011

据等。

## 招聘信息

美团点评 NLP 团队招聘各类算法人才，Base 北京上海均可。NLP 中心使命是打造世界一流的自然语言处理核心技术和服务能力，依托 NLP（自然语言处理）、Deep Learning（深度学习）、Knowledge Graph（知识图谱）等技术，处理美团点评海量文本数据，打通餐饮、旅行、休闲娱乐等各个场景数据，构建美团点评知识图谱，搭建通用 NLP Service，为美团点评各项业务提供智能的文本语义理解服务。我们的团队既注重AI技术的落地，也开展中长期的NLP及知识图谱基础研究。目前项目及业务包括美团点评知识图谱、智能客服、语音语义搜索、文章评论语义理解、美团点评智能助理等。真正助力于“帮大家吃得更好，生活更好”企业使命的实现，优化用户的生活体验，改善和提升消费者的生活品质。欢迎各位朋友推荐或自荐至 [hr.ai@meituan.com](mailto:hr.ai@meituan.com)。



算法岗： NLP算法工程师/专家/研究员 [🔗 \(https://zhaopin.meituan.com/job-detail?](https://zhaopin.meituan.com/job-detail?jobId=291801448272331465)

[jobId=291801448272331465\)](https://zhaopin.meituan.com/job-detail?jobId=291801448272331465) 、 知识图谱算法工程师/专家/研究员 [🔗](https://zhaopin.meituan.com/job-detail?jobId=291802012322333388)

[\(https://zhaopin.meituan.com/job-detail?jobId=291802012322333388\)](https://zhaopin.meituan.com/job-detail?jobId=291802012322333388)

工程岗： C++/Java研发专家/工程师 [🔗 \(https://zhaopin.meituan.com/job-detail?](https://zhaopin.meituan.com/job-detail?jobId=295725036608130037)

[jobId=295725036608130037\)](https://zhaopin.meituan.com/job-detail?jobId=295725036608130037) 、 AI平台研发工程师/专家 [🔗 \(https://zhaopin.meituan.com/job-detail?](https://zhaopin.meituan.com/job-detail?jobId=292557987555804498)

[jobId=292557987555804498\)](https://zhaopin.meituan.com/job-detail?jobId=292557987555804498)

产品岗： AI产品经理/专家 [🔗 \(https://zhaopin.meituan.com/job-detail?jobId=291803922350539431\)](https://zhaopin.meituan.com/job-detail?jobId=291803922350539431)

(NLP、数据方向)