语义计算：下一步做什么？兼谈跨语言跨媒体语义知识挖掘

厦门大学 史晓东

2014-4-18

中文信息学会第3届战略研讨会

贵阳花溪

大纲

- 动机
- 语义知识
- 跨语言语义知识挖掘
- 跨媒体语义知识挖掘
- 问题

动机


121.192.180.171/cgi-bin/jb.cgi?kw=胡锦涛
至善
输入一个或多个检索词：
胡锦涛
1494 个结果（397 毫秒）
1. 【人民网】胡锦涛发表2011年新年贺词 共同增进各国人民福祉
http://dwzy.xbmu.edu.cn/news/trans.aspx?id=4284
【 人民网 】 胡锦涛 发表 2011年 新年 贺词 共同 增进 各国 人民 福祉
2. [翻译]
http://tb.tibet.cn/zt2009/50znzl/qqgh/200903/t20090309_458826.html
没有摘要。
3. 【人民网】胡锦涛发表2011年新年贺词 共同增进各国人民福祉
http://dwzy.xbmu.edu.cn/news/wap.aspx?nid=4284&cid=26&sp=373
【 人民网 】 胡锦涛 发表 2011年 新年 贺词 共同 增进 各国 人民 福祉 【 人民网 】
4. [翻译]
http://tb.tibet.cn/2010news/xzxw/zzzj/201211/t20121116_1797041.html
没有摘要。
5. [翻译]
http://tb.tibet.cn/2010zf/flfg/201011/t20101124_774739.htm
没有摘要。

语义知识

- 什么是语义
  - 我在上一次战略研讨会中有关语义的定义想必大家还有点印象
  - 主要是2种语义
    - 外延语义：指称语义
    - 内涵语义：关联语义

语义知识（2）

- NLP处理面临着数据稀疏问题
  - 根据Zipf定律，这是永远无法解决的
  - 内因：语言是不断发展变化的
- 大数据只能解决一部分数据稀疏问题
- 剩下的数据稀疏问题要靠语义来解决

语义知识（3）

- 解决数据稀疏问题实例1
  - 枢轴语言
- 解决数据稀疏问题实例2
  - 深度学习：**learning more abstract features in successive layers (**尤其是在视觉深度学习中）

语义知识（4）

- 历年关于语义计算的重点项目(不全)：
- **NSFC**
  - **2003**：非规范知识处理的基础理论及关键技术研究
  - **2009**：网络多媒体信息的语义分析及内容监控
  - **2011**：篇章级中文语义分析理论与方法 刘挺
  - **2011:**基于本体的多策略民汉机器翻译研究 黄河燕
  - **2013:**汉语多层次语篇分析理论方法研究与应用 宗成庆
  - **2013:**跨语言社会舆情分析基础理论与关键技术研究 周国栋
  - **2013:**跨语言社会舆情分析基础理论与关键技术研究 赵小兵
  - **2014:** 面向多层次篇章语义的机器翻译方法研究与实现
- 支撑计划
  - …自动化所…北大…
- **863**
  - **2014:**互联网话语理解的认知机制与计算模型
  - **2014:**汉语认知加工机制与计算模型
- **973**
  - **2004:** 语义网格的基础理论、模型与方法研究
  - **2014:** 面向三元空间的互联网中文信息处理理论与方法 孙茂松

跨语言语义知识挖掘

- 大家熟悉的例子
  - **IBM统计机器翻译模型1：利用最简单的关联语义（共现），通过EM算法来训练**
  - 词典挖掘:
    - 利用“（）”来挖掘
    - **Wikipedia Infobox**

跨语言语义知识挖掘（2）

- 篇章挖掘
  - **Wikipedia**的不同语言对同一概念的文章（**Inter-Language Link**）
  - 在线论文库中的摘要挖掘（中国：中英；法国：法英）
  - 至善数字图书馆中同一本书的不同语言译本的对齐挖掘
- 多语话题跟踪
  - **[Bruno Pouliquen et al 2008]Story tracking: linking similar news over time and across languages**

跨媒体语义知识挖掘

- 历年关于跨媒体计算的**NSFC**重点项目(不全)：

title	**PI**	金额	单位	年代
跨媒体海量信息的综合检索与智能技术的研究	潘云鹤	**180**	浙大	**2005**
跨媒体海量信息的综合检索与智能技术研究	薛向阳	**150**	复旦	**2005**
面向互联网的跨媒体挖掘与搜索引擎	庄越挺	**280**	浙大	**2009**
跨媒体协同处理与服务的理论和应用研究	张文生	**250**	自动化所	**2011**
多媒体内容分析与搜索	徐常胜	**200**	自动化所	**2012**

跨媒体语义知识挖掘(2)

- **A picture is worth a thousand words?**
  - **[Yansong Feng & Mirella Lapata 2008]How Many Words is a Picture Worth?**
- **A sentence can be depicted by a thousand pictures!**
  - **Google: Most beautiful town in the world**

跨媒体语义知识挖掘（3）

- 所用方法：仍然是采用前面所述2种
- 关联语义：利用文字和图片的共现，来求出图片和词的概率推导关系。常见方法：
  - 图片聚类
    - 去除不正确的偶然共现
    - 可惜：聚类准确率太低
  - 主题模型：可将视觉单词和语言单词一起聚类
  - 概率图模型：关联关系弄清楚了，推理就好做了

跨媒体语义知识挖掘（4）

- 指称语义：**[Peter Young et al 2014]From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions** 采用语言表达式的视觉指称（即一组图片）来定义指称相似性度量，在某些语义推导任务中，效果好于分布式语义表示
- 苏东坡：“味摩诘之诗，诗中有画；观摩诘之画，画中有诗”

相关研究

- 机器阅读：从非结构化文本中抽取知识。
- 知识图谱：大数据（**2012年google的知识库就包含超过570亿个对象**）
- 文本蕴涵（**2004-2013 八届RTE挑战任务**）。是问答系统、机器翻译等的必要模块（不仅仅是**paraphrase**）
- 隐喻计算

问题

- 尽管取得了很多成果，对语义的丰富内涵仍然是无能为力：
  - **What is love?**
  - ***Love* is that feeling you get when you meet the right person.**
  - **Dear, you are my love.**
  - *爱是给予而不是索取*
  - *爱是*Love *爱是*Amor
    *爱是*Rarc *爱是爱心* *爱是*人类最美丽的语言
- 消歧这个概念只是面向某些任务的做法，不是语义理解的目标

问题（2）

- 语义计算需要认知
- 语义计算需要关注语用？
- 语义计算需要回到肇始于**Montague**的内涵逻辑？
- 语义计算需要涉身**(embodiment**）？
- 如何计算双关语？
  - **Fifty Shades of Grey**

谢谢大家！