谈谈语义计算：我的一点思考

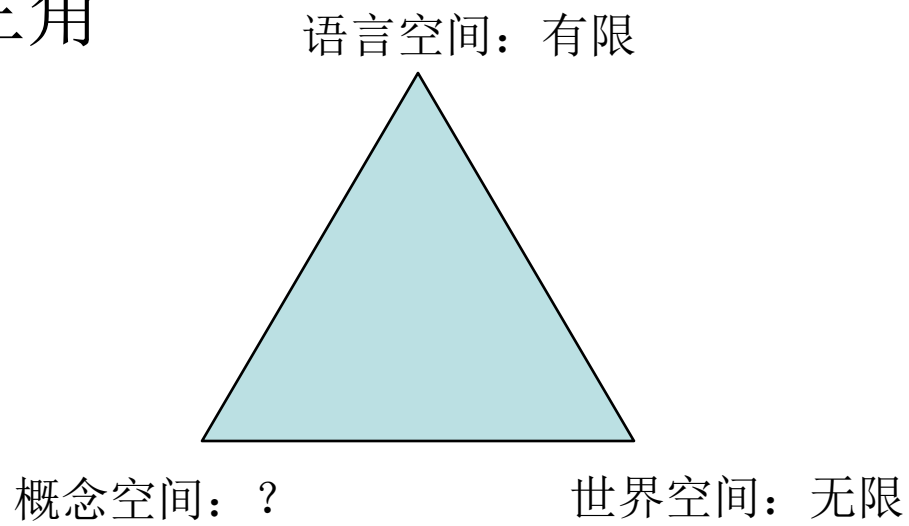
史晓东

厦门大学人工智能研究所

2012-4-13

语义计算

- 语义：
  - 内涵语义：系统内部对象之间的关系
  - 外延语义：两个系统之间的对象映射
- 语义计算的困难
  - 语义三角



语义计算

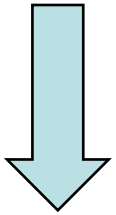
- 经典计算机科学：语言空间和概念空间基本上合二为一（没有歧义的形式语言），研究主要集中于外延语义
- 人工智能（包括NLP）：目前主要集中于语言空间和概念空间的研究
  - 传统方法：最终目的是概念空间（格语法？）
    - 好处：一揽子解决多语言 （世界语不是bytecode?）
  - 数据驱动方法：主要集中于语言空间
    - 好处：语料是可唯一直接把握的

语义计算

- 问题1：有必要研究概念空间吗？比如还要Ontology?还要知网？
  - 背景：Web is huge and available. Web media is real (微博). Data mining is almost enough (Siri?)
  - 进展：很多概念空间问题，可以被重新描述为语言空间内部问题，从而引出了新的语义计算范式。比如：话题模型改变了我对于语义的理解。

话题模型简介

- 一篇文档是一组话题的一个概率分布
- 一个话题是一组单词的一个概率分布



- 歧义问题新解：
  - 一词多义表现为同一个单词可属于不同的话题
  - 一义多词表现为不同单词的概率同分布

语义计算

- 基层干部怕互联网
  - 意见1：Web是受污染的（spam？）我要墙。
  - 证据2：Wikipedia的质量超过大英百科全书
- 问题2：
  - 如何去伪存真？(credibility computing)
  - 如何把人工语义知识网（知网，FrameNet，群体智慧等）和自动语义关系挖掘结合？你不能有个好的后天教育就不要好的先天遗传了吧？
  - 矛盾时，我信谁？

中西医问题

- 在很多具体的语义计算问题上已经取得了很大的进展（比如，关系抽取，情感分析...），越做越细（周国栋的图），就像西医一样有很多特效药，但是只能用机器学习的方法去一个个套？（然而，机器学习不是万能，特别是监督方法有很多局限，no free lunch）
- 问题3：有没有统一的语义计算理论，像中医的黄帝内经？

语义是什么

- 问题4：语义空间还是语义网络？
- 语义空间：语义就是点和点之间的关系而已。（男人：为人子，为人夫，为人父，…）点是什么？
  - 背景：Padó 2007
  - 一个实例：语义距离
    - 词距离/概念距离
    - 点距离
      - 点先变成团，避免消歧
  - 中文的意合导致的语法困难也许迎刃而解？

语义计算与机器翻译

- 训练数据：
  - 句对是不够的
  - 文档集、篇章、句群（text segments)对齐
- 可望解决目前仅靠翻译模型和语言模型（浅层词计算）的简单模型的缺陷
- 可望解决数据稀疏问题（多领域问题？适应性问题？）

总结

- 语义计算不一定非要采用传统的思路
- 语义计算需要利用Web数据和人的洞察
- 语义计算可望解决一些领域的部分瓶颈
- 语义计算也许能成为改变中文计算落后于英文计算的状况的契机