数据资源开发的回顾与展望

Historical Development and Future Directions in Data Resource Development

Workshop 出席者：Martha Palmer, Stephanie Strassel, Randee Tangi
报告撰写者：Martha Palmer, Randee Tangi, Stephanie Strassel, Christiane Fellbaum, Eduard Hovy

介绍者：北京语言大学　杨尔弘

Historically Significant Developments in Data Resource Development

- 语音转写 (Transcribed Speech)
- 电子词典（Publicly available electronic lexical resources）
- 文档集（Document Collections）
- 平行语料库（Parallel Corpora）
- 训练数据资源（Linguistic Annotation as Training Data）

1.Transcribed Speech

- TI46
  - 在1980年设计并开发的由 口语词 组成的语料库，有男女共16名讲话者，每人录入46个词
- DARPA TIMIT
  - 包括6300个句子，有630名来自美国8个不同方言的地区的讲话者，每人录入10个句子
- 其他有影响的语料库
  - 包括华尔街日报语料库、会话语音Switchboard语料库、和Hub4广播新闻语料库。

2.Publicly available electronic lexical resources

- WordNet
  - 应用最广泛的英语词汇语义词典.
  - 在WordNet中，以名词、动词、形容词和副词的同义词集合作为基本单位；词、词集间通过词汇的反映的概念语义关系相联系，这些关系包括上下位关系、整体部分关系、反义和词汇的继承关系。由此产生的网络结构使它可以唯一确定一个词的所有含义并且能够确定与其他词的相似度。
  - WordNet的公开免费使用，使之成为了语言处理系统中词汇库 的 社会标准

3.Document Collections

- 文本检索会议（TREC），作为TIPSTER Text program的一部分，由美国国家标准技术协会（NIST）和美国国防部在1992年共同主办。
- 提供大规模文档检索方法评测的数据和评测方案，极大地促进了IR和QA的技术发展。
- TREC给出一系列的测试集。每个测试集包含一组文档、主题（问题）和相关性判断（正确答案）
  - Ad hoc Test Collections, Web Test Collections, Blog Track, Confusion Track, Enterprise Track, Filtering Track, Genomics Track, HARD Track, Interactive Track, Legal Track, Novelty Track, Robust Track, Query Track, Question Answering Track, and the SPAM collection

4.Parallel Corpora

- Hansards（加拿大议会的官方记录）
  - 词对齐语料库
  - IBM的研究人员利用这个巨大的英语、法语平行语料库，开发了使得机器翻译领域发生了重大变革统计 翻译系统。

5.Linguistic Annotation as Training Data

- 宾州树库、词义标注语料库SEMCOR的应用 已经表明有限的标注数据能使复杂的自然语言处理系统得到改进。
  - 通过这些语料的训练，改进了句法分析、词义标注的系统的性能。
- 这些标注语料的成功运用，促进了越来越多的各种标注语料的发展。包括ACE标注（实体标注、名词性标注、指代、语义联系和事件）、语义标注（如情感标注）,在PropBank、NomBank 和FrameNet中的语义角色标签，以及语用标注，指代，语篇树库。

未来发展方向

- 标注科学（Science of annotation）
- 具有鲁棒性的，可扩展的标注基础架构（Robust, extensible annotation infrastructure）
- 集成、融合 出现 的技术（Closer integration of emergent technology）
- 更丰富的标注（Richer annotations）
- 语言资源包（Language resource kits）
- 可以广泛运用、经验的词汇资源（Broad coverage, empirically grounded lexical resources）

Science of annotation

- 目前对标注语料的需求很大
  - 不同类型的语言的相同标注
  - 对已有的 数据 更新、内容更丰富的标注
  - 平行语料库的平行标注
  - 合并已经标注过的语料库

Science of annotation

- 机器学习

- 可用的linguistic annotation 实际上确定了NLP的研究任务。
  - 日益丰富的标注成为复杂的nlp系统的关键因素
  - 标注中，一致性（agreement，consistency and coherence）对系统的训练具有很大影响。

Science of annotation

- 什么是需要精确标注的?
- 标注
  - 标注设计（类别，内在联系，环境）
  - 好的标注策略依赖于相关领域的固有理论
    - 语言学（语言结构）、修辞学（文本的结构和论证）、文体学（风格和主题），词典学（词法语义）
- 困难
  - 清晰理解了的语言现象并不一定能够直接的标注出来供机器学习
  - 以下因素的精确定义、理解
    - consistency,
    - Coherence
    - lack of complexity of the annotation

Science of annotation

- 成功的标注是科学，更是艺术
  - 世界范围内，只有3、4所研究机构具有进行有效标注的经验和知识
  - 他们可以将这些标注知识 编撰（ codify） 成规范，方便地应用于不同的site
  - 必须建立对有效标注详细过程的明确描述

Science of annotation

- Guildline（指导原则）中需要明确
  - 语料库
    - 什么构成一个平衡的、有代表性的适时的语料库?
  - 标注
    - 哪些语言现象是标注的目标？如何表示它们?
    - 如何编制一本标注参考手册，当标注人员面对多个选择时，能够快速、准确、一致地做出决定?
    - 如何保证并衡量作为训练数据的标注语料库实用性（usefulness）？

Science of annotation

- Guildline（指导原则）需要明确
  - 标注人员
    - 他们需要什么样的资质、培训和指导?
    - 实际标注效率如何估算?
    - 应该给他们提供哪些工具?
    - 良好的标注界面的设计原则是什么?
  - Inter-annotator一致与不一致
    - 应如何衡量一致性（kappas, F-measures, confusion matrices等）？对不同类型任务的实际期望是什么?
    - 如何发现并解决不同源头的disagreements，例如由于粗心，或模糊的说明，或模棱两可的数据造成的不一致？哪些是受到良好设计的标注界面所影响的？哪些不是?

Science of annotation

- Guideline中还需要阐明
  - 每个标注过程的详细信息以及不同类型的标注之间的相互影响
    - 比如 proposition bank规定先给出动词的框架文件，nombank则先参考对应动词的论元

- 标注方法和工具
  - 每一个不同类型的标注都需要一个稳定的、语言无关的方法和工具。
    - 比如 在不同语种上开发Propbank的方法

Science of annotation

小结

- 对“annotation science” 的原则和技术达成共识 是进行成功标注和满足日益增长的标注需求的唯一道路。
- 建立“annotation science”，解决上述的相关问题，需要进行充分的研究并有定期的技术论坛来讨论，比如acl sig为此提供了方便

- 标注语料库的重要作用：评测、定义所能够表达的层次（define the desired levels of representation）
- annotation science 的目标之一： 建构一个public domain infrastructure

Robust, extensible annotation infrastructure

- Guideline是纲领
- Infrastructure是保证措施
  - 标注工具包（自动标注作为预处理，人机交互界面……）
    - 使用相同的工具或APIs，保证每一个标注能够严格按照guideline进行标注
    - 提供标记之间的相互照应（如 不同层级标注共存的统一框架）
    - 具有对公共的、相似数据的追踪
- 目前标注技术的局限
  - 面向特定任务，不能与已有的技术、工具协同工作，

Robust, extensible annotation infrastructure

- 标注是Computing Research Infrastructure（CRI）中的重要研究领域（The CRI program at NSF now explicitly acknowledges）
- 面临的发展
  - 开发可重用的，library-based code，这些可以为特定的标注任务模块服务。
  - 开发 industrial strength toolkit
    - 对 能够很好体现 和解释 标注科学中相关内容 的 robust linguistic annotation infrastructure的持续支持

Robust, extensible annotation infrastructure

- 面临的挑战
  - 定义急需的数据，来支持机器学习技术。
  - 定义原则，来创建可以容易地与其他标注相层叠layered的模块化标注。例如：明确定义的独立（standoff） 的XML标注以及不同层级间的兼容性，使标准的查询器知道如何获取分层的标注。（查询工具可以随意访问任意层标记）
  - 开发一个基础的启动工具包（annotation infrastructure starter kit），当进行新的标注时，可以方便地将已有的所需数据和原则等资源启用。
  - 举办包括赞助者在内的定期研讨会（例如（1）中提到的ACL SIG），更公开地讨论以下数据资源的优先权：
    - New genres（电子邮件，博客，文本消息，会议等）
    - New languages for the resource kit
    - 更丰富的标注

Closer integration of emergent technology.

- 有充分的证据表明，可以通过使用准确的自动标注器对数据进行预处理，来提高人工标注的生产率
  - 机器学习的技术可以帮助选择标注实例
    - 如WSD中的主动学习
    - 对于更复杂的，新的标注内容，需要首先进行基础研究，以便可以支撑主动学习
  - 这需要标注人员和为构建IR、NLP等系统的技术人员协作，以便将 emergent 的技术用于标注的流程中

Closer integration of emergent technology.

- 需要探索的问题
  - 在标注中，何时融入pipeline technology，
  - 自动标注器的准确度需达到什么程度?
  - 需要认真研究bootstrapping自动标注器，以便从少量的已标注数据发现最少的所需的附加标注，从而使标注器达到最佳性能。
- 提高标注性能(annotation practice)是极有价值的研究领域
  - 快速地标注最有用的训练数据
  - 识别和选择有难度的/稀有的/特殊的数据，使得有限的人工标注资源被更好地利用
  - 激励机器学习的研究，最大化有限的标注数据
  - 找到 确实 需要人工调整 的数据
  - Smarter 标注，而非harder标注

Richer annotations

- 对语言的不同层级的表示的探索使NLP不断进步，可标注的层次、内容从某种程度上引导着NLP的发展方向

- 在标注研究中，重要的是找出每一个能够被成功标注的表示层次?
  - 这是对语言学、标注和机器学习方面的专业性要求很高的基础性理论研究。

Richer annotations

- 当前各个应用领域中需求迫切的标注的内容
  - MT：多语种的，完全或者部分句对齐的，且源语言和目标语言具有的一致标注的平行语料库
  - Spoken Language Understanding：包括语音信息、术语标识、语义角色和语义标签、语篇结构、方言信息等的丰富的口语的transcriptions语料库
  - Natural Language Processing：事件及其关系（因果、时间、模式等）、语篇关系等的识别；对于平行语料库，研究如何将一种语言的标注映射到另一种语言
  - Information Retrieval：有助于搜索引擎应用的语音语料库的标注

Language resource kits

- 共同的认识：要开发各语种的语言资源包，以便某种语言一旦成为焦点或关键语言时，可以很快利用资源来构建有用的处理系统
- 目前，LDC已经为开发一些小语种的语言资源，这些语言有：
  Urdu, Thai, Hungarian, Bengali, Punjabi, Tamil, and Yoruba，开发的资源包括
  - 单语言文本、平行文本、词性标注器、语言形态分析器和命名实体标注结果。
  - 这些资源开发很有限。语音识别、句法和语义结构等都还没有。

Language resource kits

- **更加完善的资源包应包含:**
  - 文本——至少拥有10万词的平行语料（新闻领域），对于每一个新闻文本，应标注名词性实体及指代情况、基本的句法注释、基本的谓词论元结构、主题以及相关性判断;
  - 语音——发音词典，最小100小时的声音，最小10-25小时的转录，50%的新闻广播，50%的谈话广播，或采访。
- **理想情况下，这样的语言包应该包含100-200种小语种语言。**

Broad coverage, empirically grounded lexical resources

- Broad coverage
  - 可以支持各种应用
- empirically grounded
  - 源自于真是语料，而且是对 perceived entailments 提供重要证据的 短语、或句子

Broad coverage, empirically grounded lexical resources

- 通常，对有指导的统计学习系统进行移植时，词典是潜在的重要资源。
  - 许多有指导的统计方法在很大程度上依赖词汇提示。
  - 自然，当出现未定义词语时，这样的方法就会产生问题
  - 一些人提出了解决这样问题的途径是 利用 通用 的词汇资源 提供的 词项之间的 相似性 链接，实际上并不是很有效
- 问题
  - 词汇资源并不是以最有效的方式组织的
  - 统计学习系统并没有设计出有效的利用方法
  - Or both above

Broad coverage, empirically grounded lexical resources

- 非常 desirable 词汇资源，如
  - name lexicon：以各种语言出现的同一个人名的所有形式，这个人名在具体文档语料中的位置链接 等
  - bilingual lexicon：包括与文本中具体用例的链接
  - 基于实例的MT系统 包含着大量与具体实例相联系的词汇资源
- Tying the “word lists” to instances in corpora helps bridge the gap between the classic lexicon and the needs of a stochastic system.

Broad coverage, empirically grounded lexical resources

- 目前在词汇资源方面存在的问题
  - 词汇资源不能以最有效的方式进行组织
  - 面向应用定制的词汇资源局限大，即使是通用的词典，也不能在ASR，NLP中共享（比如没有语音的相关信息）
- 构建词汇资源，令其能够在ASR, NLP, MT and IR中 具有一致含义地使用，是一项基础的工作

Broad coverage, empirically grounded lexical resources

- 下一步工作
  - 在单语中 建构能够支持语义相似词语的自动发现的 资源，这样词汇资源是与语料中的实例想关联的，能够解释对应的一些 paraphrases 和推理 。
  - 更进一步，建设更加丰富的词汇资源（单语或双语）来描述事件的层次，并与本体融合，而本体本身与能够支持推理的知识库关联在一起。这是我们可以从浅层语义理解层面走向更深一层的问答系统和信息抽取系统所需的理解层面的必经之路。

谢谢！

敬请批评指正