汉语事件句式规范化

邱晗

2014.4.21

Outline

- 背景介绍
  - EC，LC，PA
  - 工作目标
- 工作内容
  - 分类体系
  - EC分类
- 结果统计
  - 数量，Kappa

背景介绍

事件句式（EC）

- 小句组合蕴含了完整的事件，动作，过程，状态，关系等内容
- 能够完整表达该事件的最小结构就是一个事件句式（Event Construction）
- “刚才，这位乘客丢失了一个红色的钱包”
- “乘客-丢失-钱包”

汉语搭配关联对（LC）

- EC：事件的完整内容→汉语搭配关联对（Lexical Cohesion）：动词同各个名词性成分的搭配关系
- “ZW 乘客-丢失”，“PO 丢失-钱包”
- 动词的搭配特性，深层语义关系
- 问题：部分LC仅反映句法搭配关系，同深层语义之间存在一定的误差
  - “菜刀-切-菜”→“ZW 菜刀-切” ×

谓词论元搭配（PA）

- 谓词论元（Predicate Argument）：EC中核心谓语动词和其直接控制的名词性短语
- 控制：谓词论元之间的语义层面的联系
- 谓词：“丢失”，论元：“乘客”，“钱包”
- 谓词：“切”，论元：“菜”
- 施事和受事
  - “丢失”的语义施事为“乘客”，语义受事为“钱包”，
  - “菜”是“切”的语义受事

PA分类

- 显式PA搭配（Basic-PA，BPA）
  - =LC
  - 前一元PA，例如：“宝宝-睡觉”
  - 后一元PA，例如：“漂着-残骸”
  - 二元PA，例如：“科学家-研究-生物学”
  - 三元PA，例如：“我-借给-他-三本册子”
- 隐含PA搭配（Implicit-PA，IPA）
  - LC存在问题的部分。

假设与目标

- 目标：分析与提取IPA结构，标识EC中的深层语义关系，形成一个更加准确全面的EC描述库
- 原理：BPA和IPA都是来源于EC中→反映出完成的事件内容
- 假设：每个IPA都能够找到其对应的BPA使得两者所反映的事件内容是相同，通过一定的方法将IPA还原成为BPA
- 方法： EC分类进行规范化处理→IPA还原为描述相同事件的BPA

工作内容

分类体系建立(1/3)

基本事件句式(BEC)

- SP(B), SPO(B), SPOO(B), SPOC(B)
- PO(B)
  - 掉了-三颗牙齿
  - ≠SP(B)
- SSP(B)
  - 这次的旅行团-我-带队
- SDP(B), SDPO(B)
  - 我-同他-结婚

分类体系建立(2/3)

派生事件句式(DEC)

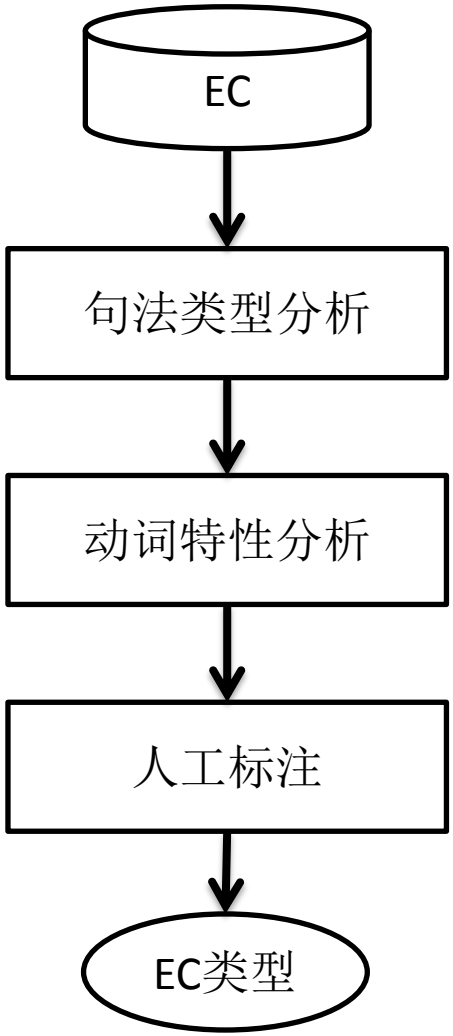
- 论元省略(O)
  - SPO(B)→SP(O)，SPO(B)→PO(O)，SDP(B)→DP(O)，SDPO(B)→DPO(O)，SPOO(B)→POO(O)，SPOC(B)→POC(O)
- 论元话题化(T)
  - 完全话题化(WT)：核电站-我国-应当大力发展
  - 动宾话题化(POT)：我-作业-早就写完了
  - 同位语话题化(CT)：王晓明-这个人-比较特殊
  - 部分话题化(PT)：敌人-我们-消灭了-三个
  - 省略后的话题化(OT)：大门-打开了
- 谓词被动化(B)
  - 大量毒品-被-销毁
- 论元pp状语化(D)
  - 研究人员-将血清-提取出来
  - 我-将钱包-递给-他

分类体系建立(3/3)

变形句式(TEC)

- 名词中心语(NH)
  - 附加中心语(EH):传统医学-具有-多样性特点-的-原因，群众-承受-的-能力
  - 论元中心语(AH)：具有-表现力-的-艺术，出让-的-土地
- 谓词中心语(PH)
  - 图书-的-借阅

EC分类处理(1/4)



EC分类处理(2/4)

- 句法类型分析
  - 作业-我-写完了→SSP
  - 菜刀-切-菜→SPO
  - 具有-表现力-的-艺术→PO_H

EC分类处理(3/4)

- 动词特性分析
  - BEC和DEC：动词搭配性质和语义限制
    - 自动判断类型：这件事-我-确实没想到（二元，满足语义限制）
    - 需要人工判断：中国-人口-众多（一元，不满足）
  - TEC：动词搭配性质
    - 自动方法缩小了所属类型的可能情况，但是仍需要人工判断
      - HPO(B)→PO_H，PO(O)→PO_H
      - HPO(B)→PO_H，POH(O)→PO_H

EC分类处理(4/4)

- 人工标注

句子情景：来自偏西北方的印欧人种

片段：来自-西北方-的-人种

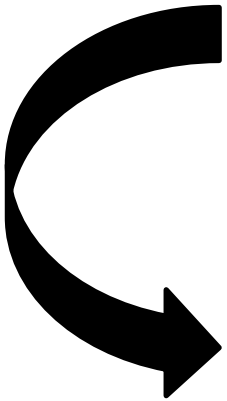
谓词特性：二元谓词

组合类型：PO_H

选项：1. 人种-来自-西北方

2. 来自-西北方

3. 以上选项均不对



序号= 6018-1

来源语料= CCGBank.ccg

完整推导树= ……

片段结构推导树= (NP C A-H (NP/NP A C-H (S\NP C H-arg-C ([S\NP]/SP H-arg leaf 来自) (SP C A-H (SP/SP A (S\NP leaf 偏)) (SP H leaf 西北方))) ([NP/NP]\[S\NP] H leaf 的)) (NP H leaf 印欧种人))

句法结构类型= PO_H

事件句式类型= ADZ

事件句式标注= [P-vp-Pred(2) 来自] [O-np-Arg2 西北方] 的 [H-np-Arg1 印欧种人]

结果统计

结果统计(1/3)

- TEC数量统计

变形方式	变形来源	数量统计	
EH	BEC	1530	5826
	DEC	4296	
AH	BEC	5365	6527
	DEC	1162	
总计		12353	

结果统计(2/3)

- 人工标注质量与难度

$$\text{Kappa} = \frac{\text{Po} - \text{Pe}}{1 - \text{Pe}}$$

Po——实际一致率

Pe——理论一致率

变形方式	变形来源	Kappa3		Kappa2	
EH	BEC	0.9341	0.8472	0.9603	0.9238
	DEC	0.8034		0.8729	
AH	BEC	0.8817	0.8338	0.9224	0.8974
	DEC	0.7962		0.8236	
总计		0.8396		0.9103	

结果统计(3/3)

- EH>AH
  - EH
    - [S-np-Arg1 轻工业] [P-vp-Pred(1) 发展] 的 [H-np 时代]
    - [S-np-Arg1 企业] [P-vp-Pred(2) 资助] [O-np-Arg2 剧团] 的 [H-np 资金]
  - AH
    - [P-vp-Pred(1) 古老] 的 [H-np-Arg1 恒星]
    - [P-vp-Pred(2) 研究] 的 [H-np-Arg2 恒星]
- BEC>DEC
  - BEC
    - [S-np-Arg1 团体] [P-vp-Pred(2) 深化] [O-np-Arg2 改革] 的 [H-np 思路]
    - [S-np-Arg1 儒家] [P-vp-Pred(2) 提倡] 的 [H-np-Arg2 礼制]
  - DEC
    - [P-vp-Pred(2) 集中] [O-np-Arg2 兵力] 的 [H-np 方法]
    - [P-vp-Pred(2) 挖好] 的 [H-np-Arg2 坑]

参考文献(1/2)

- 石定栩. 汉语句法的灵活性和句法理论. 当代语言学.第2卷2000年第1期18-26页 (2000) Adam Meyers. Annotation guidelines for NomBank - noun argument structure for PropBank. Technical report, New York University. (2007).
- Hockenmaier, J., Steedman, M.: CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. Computational Linguistics 33(3), 355--396 (2007)
- 宋彦, 黄昌宁, 揭春雨. 中文CCG树库的构建. 孙茂松、陈群秀主编《中国计算语言学研究前沿进展》（CNCCL-2011）， 221-227 (2010)
- 周强. 汉语句法树库标注体系. 中文信息学报. 18(4): 1-8. (2004)
- 袁毓林. 汉语配价语法研究[M]. 北京: 商务印书馆, 2010.
- 董振东, 董强. 知网[DB/OL]. [2003]. http://www.keenage.com/zhiwang/c_zhiwang.html. 北大计算语言学研究所. 现代汉语语法信息词典规格说明书[DB/OL]. [2000]. http://icl.pku.edu.cn/icl_groups/syntac-dictn.asp.

参考文献(2/2)

- 陈丽欧. 汉语事件内容分析系统研究与实现[D]. 清华大学: 计算机科学与技术系，2012.
- 邱晗. 汉语动词CCG范畴人工标注规范[R]. 清华大学: 信息技术研究院语音和语言技术中心, 2011. TCT到CCG bank的自动转换：设计规范Ver 3.0. 清华大学信息技术研究院语音和语言技术中心技术报告 (2011)
- Steedman M, Baldridge J. Combinatory categorial grammar[J]. Non-Transformational Syntax Oxford: Blackwell, 2011, 181-224.
- J Cohen. A coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement, 1960, 20(1): 37-46.
- S Boxwell, M White. Projecting propbank roles onto the ccgbank[J]. Proceedings of the International Conference on Language Resources and Evaluation, 2008.
- 邱晗, 周强.自动获取大规模的汉语紧密组合词汇关联对. 《清华大学学报(自然科学版)》 2011年09期

谢谢！