基于语义相似度的两岸对应词语发现方法

厦门大学自然语言处理实验室 王博立 史晓东

2015年5月9日

Outline

- 问题缘起
- 相关研究
- 我们的方法
  - 词频分析
  - 词义分析
- 实验及分析
- 下一步工作

问题缘起

大陆和台湾在语言文字的使用习惯上存在不少差异

- 字音、拼读系统、标点符号、中文排写
- 书写系统（简体字-繁体字）、词汇、语法

词汇层面的差异

- 大陆特有词汇：解放前、离休
- 台湾特有词汇：博愛座、拚輸贏
- 异名同实：奥巴马-歐巴馬、悉尼-雪梨、网络-網路
- 同名异实：土豆、窝心
- 语义上一些更细微的差异：检讨、据点

问题缘起

- 语言差异影响两岸沟通 → 简繁文本智能转换
  - 词汇层面的转换依赖于一个准确、全面的两岸对应词表
  - 两岸对应词语即“同实异名”词汇，语义应当对等
  - 但实际上不能完全对等（雪梨-悉尼/雪梨）
    - 适用于大陆到台湾文本转换的同实异名对应词表
    - 适用于台湾到大陆文本转换的同实异名对应词表
- 辅助词典编纂

相关研究

语言学方面对两岸差异词汇的研究

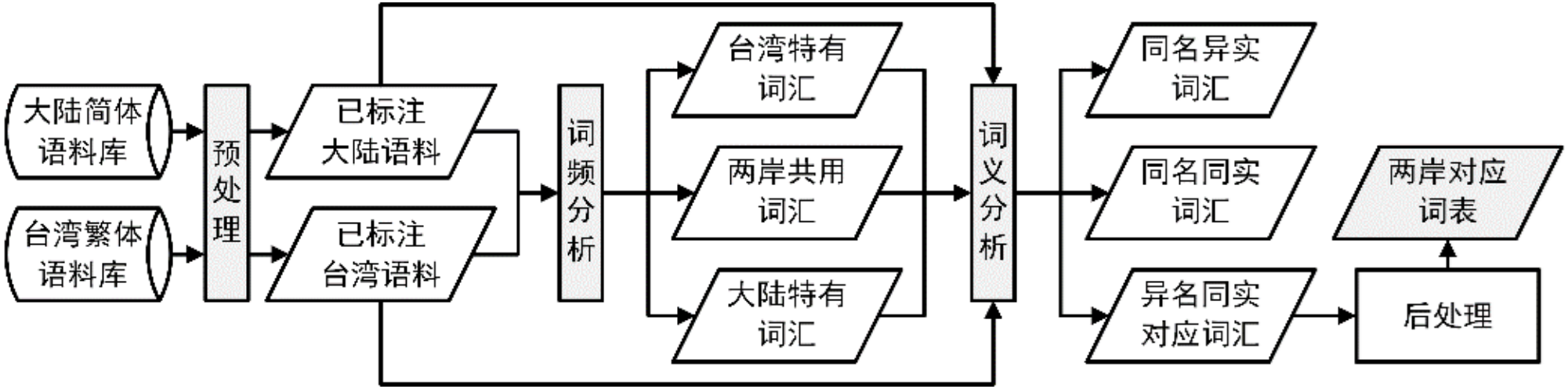
- （苏金智，1995）梳理了两岸同形异义词
- （李行健，2012）&《两岸常用词典》
  - 同中有异、同实异名、同名异实、一方特有
- 不足
  - 依靠语言学家的知识积累和手工整理，耗费大量人力
  - 难以得到全面的两岸差异词汇
  - 无法涵盖随时可能出现的新的差异词语：斯诺登-史諾登
- 解决途径：采用统计方法从大规模语料库中自动发现两岸差异词汇

相关研究

自然语言处理对语义计算的研究

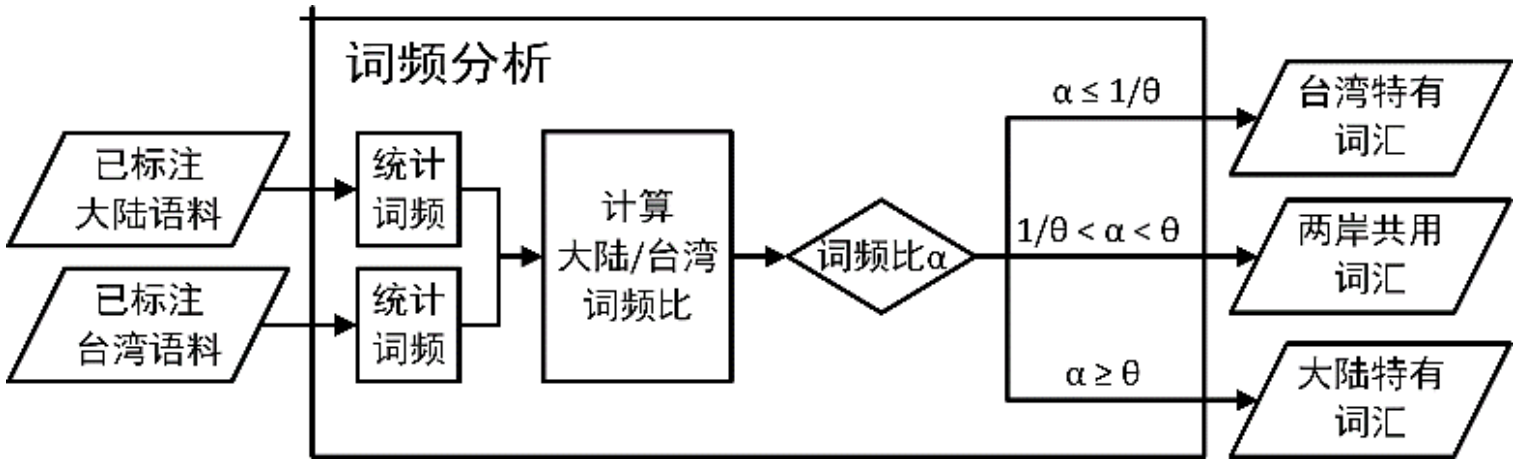
- 基于知识库的方法
  - 英文WordNet（Richardson，1994），中文知网HowNet（刘群，2002）
- 基于大规模语料库的方法
  - 基本假设：一个词的上下文提供了这个词词义的重要信息
  - One-hot Representation：上下文中词语的概率分布
    - （Chen，2011）在简繁文本自动转换中引入词汇语义一致性权重
    - （王石，2013）（石静，2013）对中文词向量的构造和相似度度量进行了大量对比实验
  - Distributed Representation：基于深度神经网络的语义表示方法
    - （Mikolov，2013）Word2Vec

我们的方法 — 总体流程



- 语料：相同年代相同领域的台湾语料和大陆语料
- 预处理
  - 分词标注：在两岸语料上使用同一套标注规范（包括切分的粒度和标注集）
  - 台湾语料繁转简

我们的方法 — 词频分析



词语的集合 $W = \{w_1, w_2, \ldots, w_M\}$

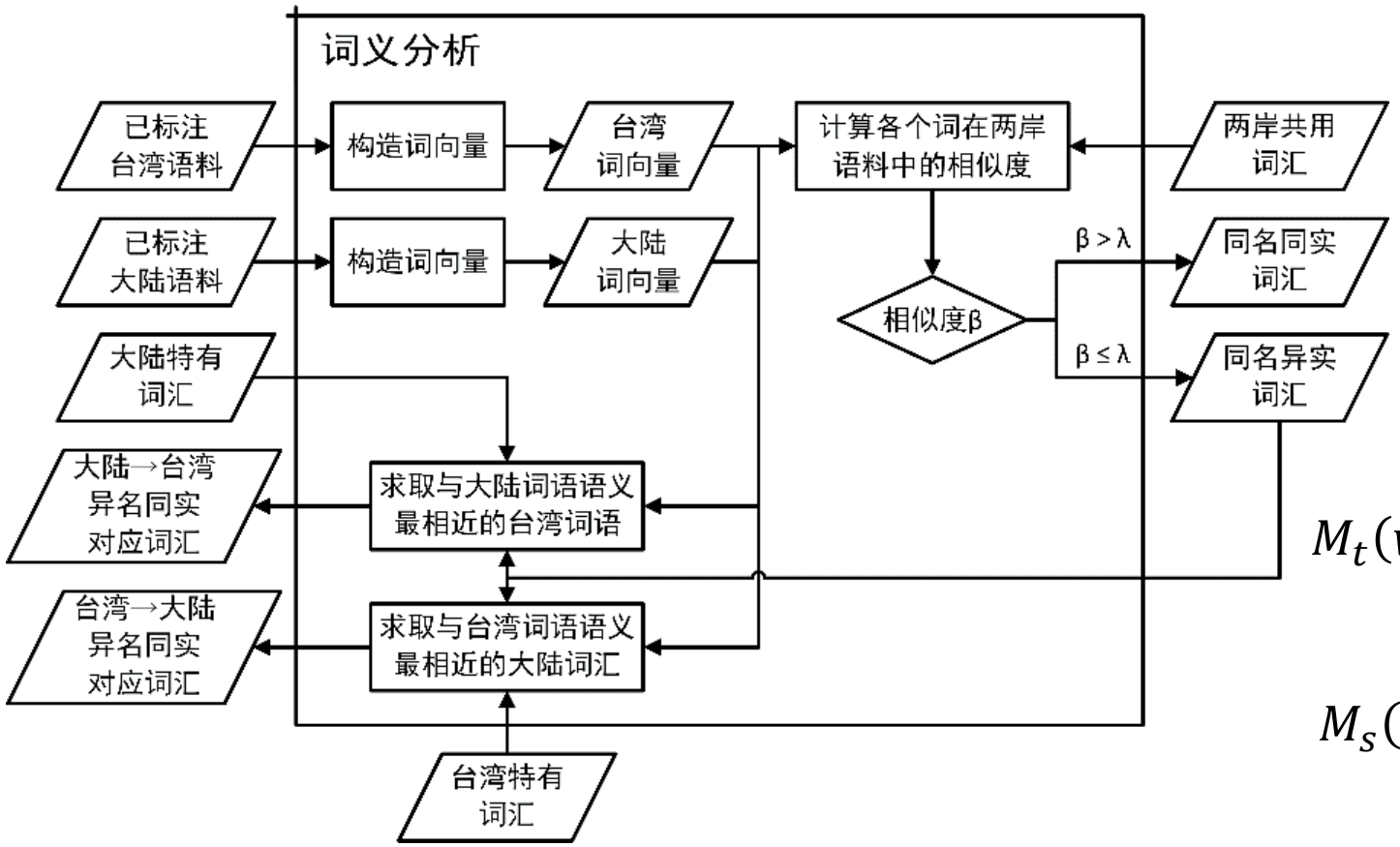
大陆语料中出现的频次 $C_s(w_i)$

台湾语料中出现的频次 $C_t(w_i)$

（数据平滑：未出现的词语频次取0.1）

词频比 $\alpha(w_i) = \dfrac{C_s(w_i)}{\sum_\omega C_s(\omega)} \Big/ \dfrac{C_t(w_i)}{\sum_\omega C_t(\omega)}$

我们的方法 — 词义分析



在两岸语境下的词义表示

$$R_s(w_i) \qquad R_t(w_i)$$

同一词语在两岸语境下的语义相似度

$$\beta(w_i) = Similarity(R_s(w_i), R_t(w_i))$$

大陆词语对应的台湾词语

$$M_t(w_i) = \underset{\omega}{argmax} Similarity(R_s(w_i), R_t(\omega))$$

台湾词语对应的大陆词语

$$M_s(w_i) = \underset{\omega}{argmax} Similarity(R_t(w_i), R_s(\omega))$$

我们的方法 — 词义表示

采用词向量方法来表示词汇的语义

采用上下文中词语的概率分布来构造词向量

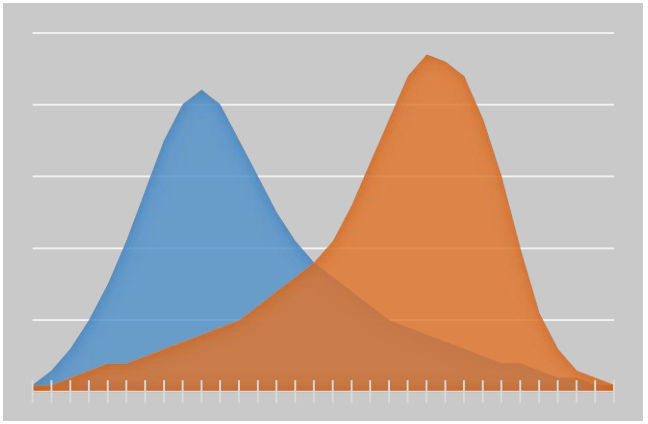
引入词序信息：左右两侧的分布分别用两个向量表示

大陆语料上的词向量 $R_s(w_i) = [R_{sl}(w_i), R_{sr}(w_i)]$

台湾语料上的词向量 $R_t(w_i) = [R_{tl}(w_i), R_{tr}(w_i)]$

我们的方法 — 词义相似度度量

- 采用直方图相交法度量词向量的相似度
- 词向量的各个维度上加权值
  - 借鉴tf-idf思想进行词频惩罚
  - 权重为该维度所代表的词语在整个语料中出现的频次的倒数



$$Similarity\left(R_s(w_i), R_t(w_i)\right) = 0.5H\left(R_{sl}(w_i), R_{tl}(w_i)\right) + 0.5H\left(R_{sr}(w_i), R_{tr}(w_i)\right)$$

$$H\left(R_{sl}(w_i), R_{tl}(w_i)\right) = \frac{\sum_k \frac{\min(R_{sl}(w_i)[k],\, R_{tl}(w_i)[k])}{C(w_k)}}{\sum_k \frac{\max(R_{sl}(w_i)[k], R_{tl}(w_i)[k])}{C(w_k)}}$$

实验

任务：在给定语料上寻找大陆词语对应的台湾词语

语料：

- 大陆：1.89亿字的新华网语料
- 台湾：1.83亿字的MSN语料

均为新闻语料，时间跨度均为2011年至2014年

实验

实验结果

- 自动发现了421个大陆词语存在对应的台湾词语
- 准确率

指标	p@1	p@3	p@5
准确率	35.15%	49.64%	53.21%

- 部分输出结果

大陆	台湾	大陆	台湾	大陆	台湾	大陆	台湾	大陆	台湾	大陆	台湾
美联储	**聯準會**	**网点**	**據點**	信息	資訊	核电站	核電廠	软件	軟體	彩票	彩券
短信	簡訊	网络	網路	概率	機率	身份证	身分證	屏幕	螢幕	**欺诈**	**詐欺**
硬件	硬體	烟花	煙火	**民警**	**員警**	厘米	公分	出租车	計程車	反复	反覆
公交	公車	**群体**	**族群**	芯片	晶片	**入市**	**進場**	智能	智慧	渠道	管道

加粗的词语为《两岸常用词典》中未收录的对应关系

分析

能够有效的自动发现两岸对应词语，具有较高的准确率

缩短了传统人工收集方法所花费的时间

能够发现一些人工收集不易发现的对应词汇

存在一部分错误的对应关系

- 存在歧义的词语：板块-類股
- 大陆一词对应台湾多词短语：残疾人-殘障/人士，平米-平方/公尺
- 不必转换或难以转换的词语：厅长、党校、纪委、信访、稀土
- 只找到语义接近的词语：高考-會考（聯考），博客-微博（部落格）

下一步工作

- 改进
  - 将搜索范围扩大到短语层面（避免分词碎片造成的遗漏）
  - 改进语义表示方法和相似度度量方法
    - 尝试采用深度神经网络训练分布式的语义表示
- 应用
  - 采用更大规模的语料，并辅之以必要的人工筛选，得到合理适用的两岸对应词表，应用于两岸文本转换系统的词转换中
- 迁移
  - 机器翻译领域：借助相关词典资源或词对齐信息，从可比语料中抽取从源语言到目标语言的词语或短语对应规则

汉字简繁文本智能转换系统

- 教育部语信司委托开发
- 2014年11月18日在京发布
- 特点
  - 字级别简转繁准确率99.991%
  - 支持面向台湾和面向古汉语的转换
  - 提供字、词、术语、标点等多层次的转换
  - 提供网站全站网页的自动转换功能
  - 向全社会免费提供在线版、word版、单机版

汉字简繁文本智能转换系统

欢迎使用：http://jf.cloudtranslation.cc/

其它相关成果

- 至善简体语料库（50亿字）
  - 新华网、搜狐博客、新浪博客、人民日报等
- 至善繁体语料库（22亿字）
  - 台湾语料：CNA、MSN等
  - 古籍和词典等
- Segtag汉语分词标注系统
  - 同时支持简/繁，古代/现代文本

汉字简繁文本智能转换系统

我们正在继续对系统进行优化、扩充和拓展，仍有许多问题需要攻坚克难。我们真诚邀请相关领域的学者（特别是台湾、香港的学者）与我们合作，共同开展更深入的研究工作。

合作内容包括但不限于以下主题：

- 两岸术语对应词表的收集整理
- 大规模古汉语繁体语料库的收集整理
- 大规模香港繁体语料库的收集整理
- 粤语相关语言资源的收集整理

谢谢！

厦门大学 王博立 史晓东