词义消歧

车万翔
社会计算与信息检索研究中心
2017年春季学期

定义

- 词义消岐 (Word Sense Disambiguation)：为一个词语从预先设定的词义项集中选择一个词义
  - 词义项集来自与词典或知识库
  - 基于知识的方法 & 监督学习的方法
- 词义区分 (Word Sense Discrimination)：在没有预定义的词义项集的情况下，将一个词语的使用划分为不同意义项
  - 无监督方法

定义

- 许多词语具有多个词义 (homonymy / polysemy)

  – **Ex: “chair” – furniture or person**
  – **Ex: “child” – young person or human offspring**

- 确定在特定句子中一个词语采用哪个词义
- 说明:
  – 通常一个词语的不同词义紧密相关

  Ex: Bank: -financial institute
  -building of the financial institute

  – 有时候几个词义能够在一个上下文中同时被激发(co-activation)

  Ex: *“This could bring competition to the trade”*
  competition: - the act of competing
  - the people who are competing

词义表示

- 词在给定上下文中的意义
- 词义表示
  - 根据词典
    - chair = a seat for one person, with a support for the back; "he put his coat over the back of the chair and sat down"
    - chair = the position of professor; "he was awarded an endowed chair in economics"
  - 根据在另一语言中的翻译
    - chair = chaise (法语)
    - chair = directeur (法语)
- 根据词出现的上下文 (discrimination)
    - “Sit on a chair”　“Take a seat on this chair”
    - “The chair of the Math Department”　“The chair of the meeting”

计算机 vs. 人

- 一词多义-很多词具有多个意义
- 计算机程序没有消歧的基础，即使对于人来说很容易
  - 计算机无先验知识
  - 大脑的工作机理？
- 歧义在人们的日常交流中并不是问题，除非在极端情况下
  - “阿隆索因车祸不幸去世”

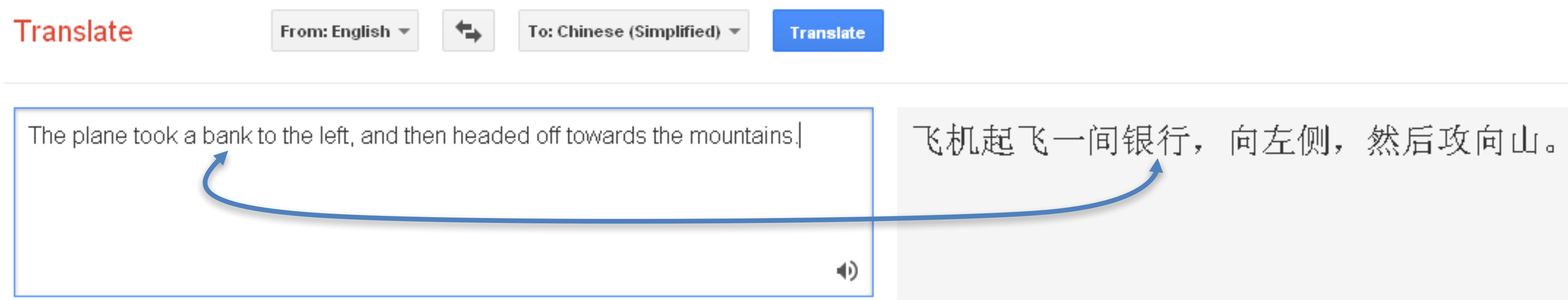
WSD 历史

- 认为是影响机器翻译的一个问题 (Weaver, 1949)
  - 一个词只有知道其特定意义才能被翻译
- 1970s - 1980s
  - 基于规则的系统
  - 依赖于人工构造的知识资源
- 1990s
  - 基于语料的方法
  - 依赖于标注好词义的文本
- 2000s
  - 混合方法
  - 利用Web数据和资源

实际应用

- 机器翻译 (Machine Translation)
  - Translate “bank” from English to Chinese
    - Is it a “银行” or a “河堤”?
- 信息检索 (Information Retrieval)
  - Find all Web Pages about “cricket” (蟋蟀/板球)
    - The sport or the insect?
- 智能问答 (Question Answering)
  - What is George Miller’s position on gun control?
    - The psychologist or US congressman?
- 知识获取 (Knowledge Acquisition)
  - Add to KB: Herb Bergson is the mayor of Duluth.
    - Minnesota or Georgia?

WSD任重而道远


Translate
From: English
To: Chinese (Simplified)
Translate
The plane took a bank to the left, and then headed off towards the mountains.
飞机起飞一间银行，向左侧，然后攻向山。


Translate
From: English
To: Chinese (Simplified)
Translate
The bank in that road is entirely too steep and is really dangerous.
银行在这条道路是完全太陡，实在是危险的。

词义消岐两类任务

- All Words Word Sense Disambiguation
  - 对文本中的所有词进行词义消岐
  - “He put his suit over the back of the chair”
- Targeted Word Sense Disambiguation
  - 对一个目标词进行词义消岐
    - “Take a seat on this chair”
    - “The chair of the Math Department”

词义消岐方法

- 基于知识的消岐
  - 使用外部词典、知识库资源
  - 使用篇章属性
- 有监督的消岐
  - 基于标注的训练数据
- 无监督的消岐
  - 基于未标注数据
    - 不使用词典、知识库资源
    - 不使用标注数据

词义消歧(WSD)之基于知识的方法

方法概述

- Knowledge-based WSD＝依赖于从词典知识库或原文本中得到的知识
- 资源
  - 使用
    - 机器可读词典
    - 原文本
  - 不使用
    - 人工标注的语料
- 不可处理所有开放词语

机器可读词典 (MRD)

- 近些年许多词典机器可读 (MRD)
  - Oxford English Dictionary
  - Collins
  - Longman Dictionary of Ordinary Contemporary English (LDOCE)
- 辞典 (Thesauruses) – 添加了同义词信息
  - Roget Thesaurus
- 语义网络 (Semantic Network) – 添加了更多的语义关系
  - WordNet

Lesk 算法

- 通过定义重叠(definition overlap)识别上下文中的词义(Michael Lesk 1986)
  1. 从MRD中获取待消岐词语的所有词义定义
  2. 确定所有词义组合的词义定义重叠程度
  3. 选择具有最高重叠度的词义组合

Example: disambiguate PINE CONE

- **PINE**
  1. **kinds of evergreen tree with needle-shaped leaves**
  2. **waste away through sorrow or illness**
- **CONE**
  1. **solid body which narrows to a point**
  2. **something of this shape whether solid or hollow**
  3. **fruit of certain evergreen trees**

Pine#1 ∩ Cone#1 = 0
Pine#2 ∩ Cone#1 = 0
Pine#1 ∩ Cone#2 = 1
Pine#2 ∩ Cone#2 = 0
Pine#1 ∩ Cone#3 = 2
Pine#2 ∩ Cone#3 = 0

利用Lesk算法对多个词(>2)进行词义消歧?

- *I saw a man who is 98 years old and can still walk and tell jokes*
  - nine open class words: see(26), man(11), year(4), old(8), can(5), still(4), walk(10), tell(8), joke(3)
- 43,929,600种词义组合! 如何找到最优的词义组合？
- 模拟退火 (Simulated annealing) [Cowie et al. 1992]
  - 定义一个函数E = 1/(1+R), R: 词义组合的冗余度(基于词出现的次数).
  - 找到最优的词义组合，最小化E

1. 初始，每个词选择其最频繁(常用)词义，计算E
2. 每次迭代中，随机选择一个词将其词义替换为另一个词义，计算E’

   如果ΔE=(E’ -E)<0, 那么保留新词义，然后进行新的随机替换

   如果ΔE=(E’ -E)>=0, 那么以一定的概率(P=exp(-ΔE/T), T为常数, 初始为1，每1000次后变为0.9T)保留新词义
3. 当词义组合不再变化，停止迭代

简化的 Lesk 算法

- 原始Lesk算法: 评估上下文中所有词语词义的重叠程度
  - 同时识别上下文中所有词语的准确词义
- 简化Lesk算法: 评估一个词的词义与当前上下文的重叠程度
  - 每次识别一个词的准确词义
- 搜索空间显著减小

简化的 Lesk 算法

- 算法步骤:
  1. 从MRD中获取待消岐词语的所有词义定义
  2. 确定每个词义与当前上下文之间的重叠度
  3. 选择具有最高重叠度的词义

Example: disambiguate PINE in

"Pine cones hanging in a tree"

- **PINE**

  1. kinds of evergreen tree with needle-shaped leaves

  2. waste away through sorrow or illness

Pine#1 ∩ Sentence = 1

Pine#2 ∩ Sentence = 0

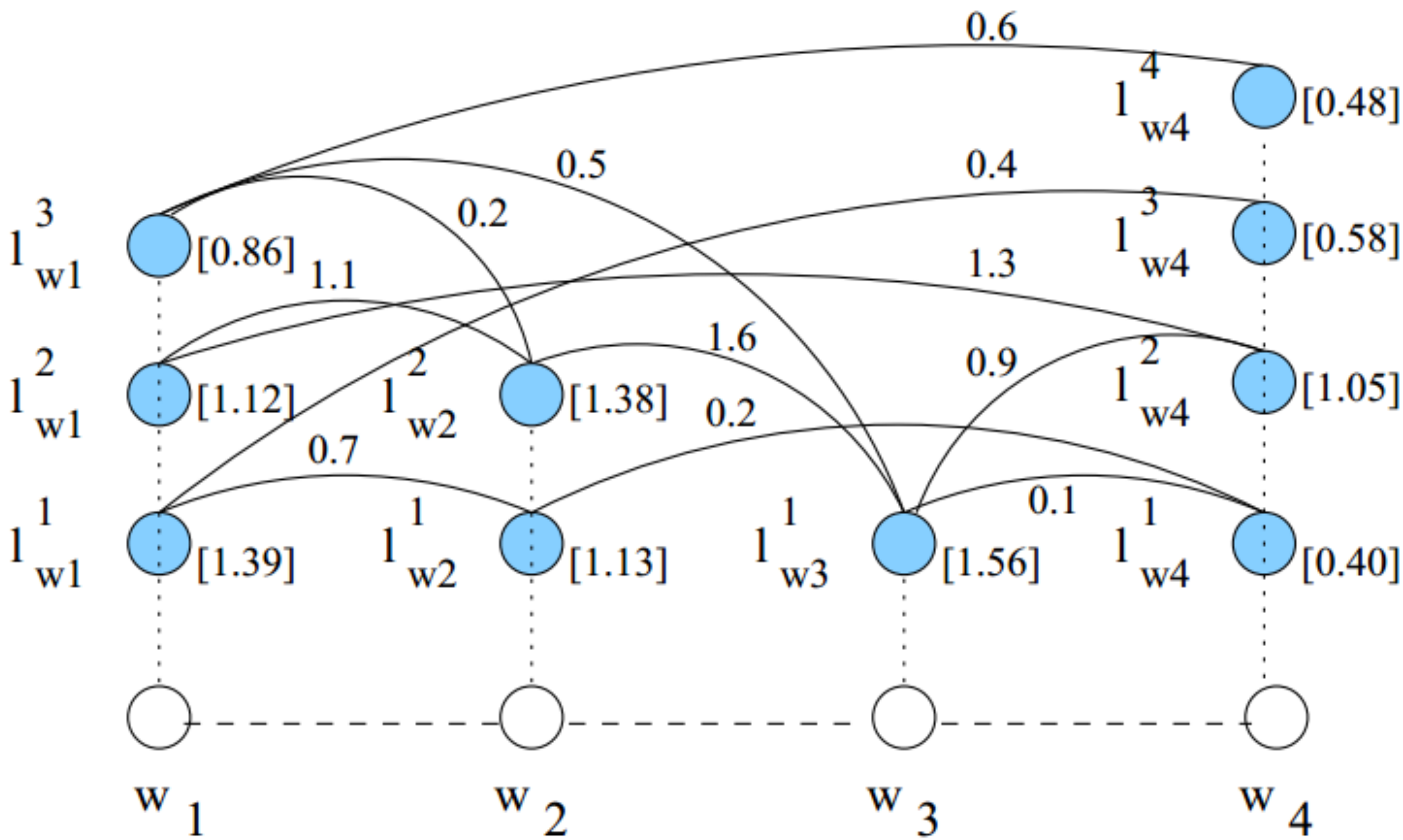
基于图排序的方法

- [Mihalcea 2005]
- 同时对所有词同时进行消岐，考虑词义之间的关联关系
- 步骤
  - 词义图的构建
    - 词的每个词义作为一个节点，词义之间的关联关系作为边(权重)
  - 基于图的排序
    - 基于Pagerank算法，一个节点的权值由跟它相连的其他节点所决定

$$P(V_a) = (1-d) + d * \sum_{V_b \in In(V_a)} \frac{P(V_b)}{|Out(V_b)|}$$

  - 词义标记选择
    - 对每个词选择权值最大的词义

基于图排序的方法



0.6
0.5
0.4
0.2
1.1
1.3
1.6
0.9
0.2
0.7
0.1
[0.48]
[0.58]
[1.05]
[0.40]
[0.86]
[1.12]
[1.39]
[1.38]
[1.13]
[1.56]
w 1
w 2
w 3
w 4

选择优先性 (Selectional Preferences)

- 一个谓语 (Predicate) 合理的参数 (比如直接宾语对象)
- 一种约束给定上下文中词语可能意义的方法
  - 如： “Wash a dish” vs. “Cook a dish”
  - WASH-OBJECT vs. COOK-FOOD

- 进而获取语义类之间的可能关系
- 获取方法：文本分析统计、机器学习
- Alternative terminology
  - Selectional Restrictions
  - Selectional Preferences
  - Selectional Constraints

Verb	Plaus./Implaus.
see	friend/method
read	article/fashion
find	label/fever
hear	story/issue
write	letter/market
urge	daughter/contrast
warn	driver/engine
judge	contest/climate
teach	language/distance
show	sample/travel
expect	visit/mouth
answer	request/tragedy
recognize	author/pocket
repeat	comment/journal
understand	concept/session
remember	reply/smoke

使用选择优先性进行WSD

- Algorithm:
  1. 对于给定的句法关系R学习到选择优先集合
  2. 给定由关系R相连的词对W1– W2
  3. 找到所有的合适的选择优先项 W1– C （word-to-class) or C1– C2 (class-to-class)
  4. 基于选择的语义类选择词W1 and W2的词义

Example: disambiguate *coffee* in “drink *coffee*”

1. (beverage) a beverage consisting of an infusion of ground coffee beans
2. (tree) any of several small trees native to the tropical Old World
3. (color) a medium to dark brown color

词汇链 (Lexical Chain)

- 词汇链是一个语义上相关联的词的序列, 表达了一个篇章/段落的连贯性以及意义的连续性
- 识别词汇链算法:
  1. 选择文本中的候选词：候选词之间应该能计算语义相似性，通常具有相同的词性
  2. 对于每个候选词的每个词义，计算该词义与已有词汇链中概念的语义相关度，计算找到一个跟该词义相关的词汇链，找到接收该候选词义的词链
  3. 如果找到这样的词汇链，那么将这个词插入该词汇链中，否则，创建一个新的词汇链

词汇链 (Lexical Chain)

A very long **train traveling** along the rails with a constant **velocity** in a certain **direction** …

train
- #1: public transport
- #2: order set of things
- #3: piece of cloth

travel
- #1 change location
- #2: undergo transportation

rail
- #1: a barrier
- # 2: a bar of steel for trains
- #3: a small bird

基于词汇链进行 WSD

- 识别文本中的词汇链
  - 通常每次针对一类词性
- 基于词语所属词汇链，识别词语的词义

基于每个篇章段落一种意义

- 在一个篇章段落中，一个词的所有出现都倾向于表达同一个意义

如：**The ambiguous word PLANT occurs 10 times in a discourse**
all instances of “plant” carry the same meaning

基于每个词语搭配一种意义

- 词语搭配(collocation)：经常共同出现，强相关的词对
- 一个词在同样的搭配使用中倾向于表达同样的意义
  - 相邻搭配中更加明显
  - 词语距离增大则减弱

The ambiguous word PLANT preserves its meaning in all its occurrences within the collocation "industrial plant", regardless of the context where this collocation occurs

词义消歧(WSD)之基于有监督学习的方法

方法概述

- 有监督的WSD: 从人工标注词义的文本上学习到分类器

- 将WSD问题看作一个分类问题
  - 基于目标词的上下文为目标词从给定词义选项中选择最准确的词义

标注词义的文本

Bonnie and Clyde are two really famous criminals, I think they were **bank/1** robbers
My **bank/1** charges too much for an overdraft.
I went to the **bank/1** to deposit my check and get a new ATM card.
The University of Minnesota has an East and a West **Bank/2** campus right on the Mississippi River.
My grandfather planted his pole in the **bank/2** and got a great big catfish!
The **bank/2** is pretty muddy, I can’t walk there.

词义的词袋模型表示 (基于在上下文窗口中词的共现)

FINANCIAL_BANK_BAG: a an and are ATM Bonnie card charges check Clyde criminals deposit famous for get I much My new overdraft really robbers the they think to too two went were
RIVER_BANK_BAG: a an and big campus cant catfish East got grandfather great has his I in is Minnesota Mississippi muddy My of on planted pole pretty right River The the there University walk West

简单的有监督WSD方法

给定包含“bank”的句子S；

对于S中每个词W_i ：

如果W_i 属于FINANCIAL_BANK_BAG，那么

Sense_1 = Sense_1 + 1;

如果W_i 属于RIVER_BANK_BAG 那么

Sense_2 = Sense_2 + 1;

如果Sense_1 > Sense_2 ，那么选择词义“Financial”

否则如果 Sense_2 > Sense_1，那么选择词义 “River”

否则，打印 “Cannot Decide”；

有监督方法框架

- 训练数据获取: 构建训练数据，每个目标词人工从预定义词义集合中标注词义
- 特征选择: 选择特征集合，表示上下文
  - co-occurrences, collocations, POS tags, verb-obj relations, etc…
- 训练集特征向量构建: 将标注词义的训练样例转换为特征向量
- 分类器学习: 使用一种机器学习算法学习一个分类器
- 测试集特征向量构建: 将单独的测试样例转换成特征向量
  - 正确的词义标签已知，但不使用
- 分类器测试: 使用分类器为测试样例赋予词义标签

从文本到特征向量

- My/pronoun grandfather/noun used/verb to/prep fish/verb along/adv the/det banks/SHORE of/prep the/det Mississippi/noun River/noun. (S1)
- The/det bank/FINANCE issued/verb a/det check/noun for/prep the/det amount/noun of/prep interest/noun. (S2)

	P-2	P-1	P+1	P+2	fish	check	river	interest	SENSE TAG
S1	adv	det	prep	det	Y	N	Y	N	SHORE
S2		det	verb	det	N	Y	N	Y	FINANCE

有监督学习算法

- 机器学习领域提供了很多这样的算法，许多算法都在WSD上取得好结果
  - Support Vector Machines
  - Nearest Neighbor Classifiers
  - Decision Trees
  - Decision Lists
  - Naïve Bayesian Classifiers
  - Perceptrons
  - Neural Networks
  - Graphical Models
  - Log Linear Models

使用单分类器的有监督WSD

- 大多数有监督机器学习能够有效进行WSD
- 不同的方法一般在所采用的特征上有所区别
- 有效的特征包括:
  - Co-occurrences or keywords (global)
  - Collocations (local)
  - Bigrams (local and global)
  - Part of speech (local)
  - Predicate-argument relations
    - Verb-object, subject-verb,
  - Heads of Noun and Verb Phrases

分类器集成 (Ensemble)

- 将不同性质的分类器集成起来通常能够提高总体效果
  - 不同的学习算法
  - 不同角度/视角的特征表示
  - 对训练集的不同采样(sampling)
- Bagging, Stacking, Boosting, …
- 怎样融合分类器结果?
  - Simple Majority Voting
  - Averaging of probabilities across multiple classifier output
- 许多WSD系统都采用了集成方法

词义消歧(WSD)之基于半监督学习的方法

方法概述

- 有监督 (Supervised) WSD = 从足量标注数据中学习词义分类器
- 半监督 (Semi-supervised) WSD = 从少量标注数据与大量未标注数据中学习词义分类器

自举方法 (Bootstrapping)

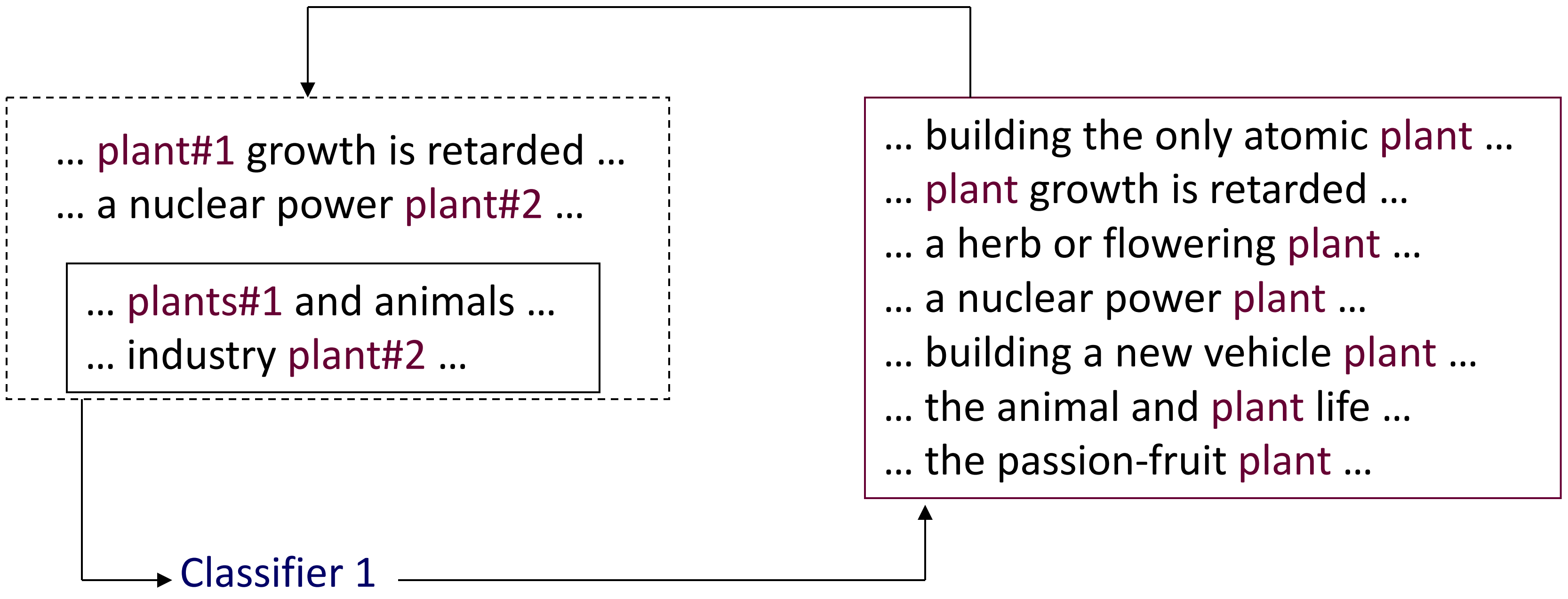
- 基于少量训练数据构建词义分类器
  - 扩展分类器的适用性

- 自举方法
  - Co-training
  - Self-training

自举方法的部件

- 输入
  - 少量标注数据
  - 大量未标注数据
  - 基本的分类器

- 输出
  - 比基本分类器具有更好效果的分类器


… plant#1 growth is retarded …
… a nuclear power plant#2 …
… plants#1 and animals …
… industry plant#2 …
Classifier 1
… building the only atomic plant …
… plant growth is retarded …
… a herb or flowering plant …
… a nuclear power plant …
… building a new vehicle plant …
… the animal and plant life …
… the passion-fruit plant …

通用自举过程

- 已标注的训练集L
- 未标注集合U
- 基本分类器C

- 1. 创建一个样例池U'
  - 从U中随机选择P个样例
- 2. 循环I次：
  - 基于L训练C，并用C标注U'
  - 从U ' 中选择G个最可信的样例添加到L
    - 保持L中的分布
  - 从U中选择样例重填U'
    - 保持U'的大小为P

通用自举过程

– 已标注的训练集L
– 未标注集合U
– 基本分类器C

- 1. 创建一个样例池U'
  - 从U中随机选择P个样例
- 2. 循环I次：
  - 基于L训练C，并用C标注U'
  - 从U ' 中选择G个最可信的样例添加到L
    - 保持L中的分布
  - 从U中选择样例重填U'
    - 保持U'的大小为P

Pool Size

Iteration Number

Growth Size

主要不足：最优参数值的选择比较困难

Self-training

- 单个分类器
- 基于自己的输出重新训练
- Self-training for NLP
  - Part of speech tagging
  - Co-reference resolution
  - Sentiment analysis

协同学习(Co-training)

- 两个分类器
  - 两种相互独立的视角
  - [独立性要求可放宽]
- Co-training in NLP
  - Statistical parsing
  - Co-reference resolution
  - Part of speech tagging
  - Sentiment analysis
  - ...

协同学习(Co-training)

- 已标注的训练集L, 每个样例两种视角表示
- 末标注集合U，每个样例两种视角表示
- 基本分类器C

- 1. 创建一个样例池U'
  - 从U中随机选择P个样例
- 2. 循环I次：
  - 基于L和视角一训练C_1，并用C_1标注U '，从U '中选择G个最可信的样例；
  - 基于L和视角二训练C_2，并用C_2标注U '，从U '中选择G个最可信的样例；
  - 将选择的样例添加到L中；
  - 从U中选择样例重填U'
    - 保持U'的大小为P

词义消歧(WSD)之基于无监督学习的方法

方法概述

- 无监督的词义区分 (Word Sense Discrimination): 基于上下文相似性将词进行聚类
- 假设
  - 具有相似意义的词倾向于出现在相似的上下文中
- 仅使用原文本中的信息，不使用外部知识库或人工标注
- 没有词义列表/目录的知识，因此聚类没有词义标签

方法概述

- 资源
  - 大量的原始语料
- 范畴
  - 每个上下文中的目标词汇需要进行词义区分
  - 计算上下文的相似程度
  - 特征可以通过单独的数据来确定
  - 对于词义类簇并不赋词义标签
- Word Sense Discrimination看作是发现那些出现在相似上下文中的目标词，并将它们聚集成一个类簇的问题

聚类方法

- 特征选择

  E.g. (Pedersen and Bruce, 1997) explore discrimination with a small number (approx 30) of features near target word.

  - Morphological form of target word (1)
  - Part of Speech two words to left and right of target word (4)
  - Co-occurrences (3) most frequent content words in context
  - Unrestricted collocations (19) most frequent words located one position to left or right of target, OR
  - Content collocations (19) most frequent content words located one position to left or right of target
- 相似度计算
- 聚类算法
  - 层次式聚类，EM算法、基于图切割的聚类等

分析

- 无监督方法不能发现与通过有监督学习得到的相同的词义类簇
- 基于已有词义类别/标签对无监督学习结果进行评价过于苛刻.
  - 可考虑人工评价

利用隐含语义分析

- Adapted by (Schütze, 1998) to word sense discrimination
- 数据表示为词语共现矩阵 (co-occurrence matrix)
- 对共现矩阵进行SVD (Singular Value Decomposition) 分解降维
  - 重要的维度跟语义概念关联
- 目标词汇的特征表示为其上下文中所有词汇特征向量的平均值（二阶表示）
- 通过余弦测度计算特征向量的相似度，然后进行聚类

分析

- 基于直接/一阶 (first order) 特征的聚类方法需要大量数据来获取有效特征
- 二阶表示 (Second order representations) 可以很好地利用少量数据获得丰富的非稀疏的上下文表示
- http://senseclusters.sourceforge.net 包括了SVD的完整无监督词义区分的系统

词义标注数据

- Senseval/Semeval评测数据
  - http://www.senseval.org
- Data for lexical sample
  - English (with respect to Hector, WordNet, Wordsmyth)
  - Basque, Catalan, Chinese, Czech, Romanian, Spanish, …
  - Data produced within Open Mind Word Expert project http://teach-computers.org
- Data for all words
  - English, Italian, Czech (Senseval-2 and Senseval-3)
  - SemCor (200,000 running words) http://www.cs.unt.edu/~rada/downloads.html
- Pointers to additional data available from
  - http://www.senseval.org/data.html

WSD Software – Lexical Sample

- Duluth Senseval-2 systems
  - Lexical decision tree systems that participated in Senseval-2 and 3
  - http://www.d.umn.edu/~tpederse/senseval2.html
- SyntaLex
  - Enhance Duluth Senseval-2 with syntactic features, participated in Senseval-3
  - http://www.d.umn.edu/~tpederse/syntalex.html
- WSDShell
  - Shell for running Weka experiments with wide range of options
  - http://www.d.umn.edu/~tpederse/wsdshell.html
- SenseTools
  - For easy implementation of supervised WSD, used by the above 3 systems
  - Transforms Senseval-formatted data into the files required by Weka
  - http://www.d.umn.edu/~tpederse/sensetools.html
- SenseRelate::TargetWord
  - Identifies the sense of a word based on the semantic relation with its neighbors
  - http://search.cpan.org/dist/WordNet-SenseRelate-TargetWord
  - Uses WordNet::Similarity – measures of similarity based on WordNet
    - http://search.cpan.org/dist/WordNet-Similarity

WSD Software – All Words

- SenseLearner
  - A minimally supervised approach for all open class words
  - Extension of a system participating in Senseval-3
  - http://lit.csci.unt.edu/~senselearner
- SenseRelate::AllWords
  - Identifies the sense of a word based on the semantic relation with its neighbors
  - http://search.cpan.org/dist/WordNet-SenseRelate-AllWords

WSD Software – Unsupervised

- Clustering by Committee
  - http://www.cs.ualberta.ca/~lindek/demos/wordcluster.htm
- InfoMap
  - Represent the meanings of words in vector space
  - http://infomap-nlp.sourceforge.net
- SenseClusters
  - Finds clusters of words that occur in similar context
  - http://senseclusters.sourceforge.net

互联网与 WSD

- 互联网已成为 NLP 的一个重要数据来源，包括 WSD
- 通过搜索能找到目标词汇的大量实例
- 搜索引擎能够选择和验证词语搭配 (collocations) 与其他的关联 (association)
  - “strong tea” : 13,000 hits
  - “powerful tea” : 428 hits
  - “sparkling tea” : 376 hits

互联网与WSD

- 维基百科提供了大量的词义列表/目录，包含新词.

Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Jordan is a country in the Middle East.

Jordan or **Jordán** may also refer to:

Geographical

Middle East

- The Jordan River
- Jordan, Tehran, Iran, an avenue and a surrounding district

United States

See also: Jordan Township (disambiguation)

- Jordan, Indiana (disambiguation), several places
- Jordan, Iowa
- Jordan, Minnesota, a city in Scott County
- Jordan, Minneapolis, a neighborhood of Minneapolis, Minnesota
- Jordan, Montana
- Jordan, New York
- Jordan, North Carolina
- Jordan, Oregon
- Jordan, Wisconsin, a town
- Jordan, Portage County, Wisconsin, an unincorporated community

Elsewhere

- Germán Jordán Province, Bolivia
- Jordan, Guimaras, Philippines
- Jordan, Hong Kong
- Jordan (Neumark), Poland
- Jordán Pond, pond in Tábor, Czech Republic
- Jordan River, New Zealand
- Jordan, Ontario, Canada
- Jordanhill, Glasgow, UK

Music

- "Jordan", a hymn tune by composer William Billings
- "Jordan" a 1998 song from Megaherz's *Kopfschuss*
- "Jordan" (song), a Buckethead song
- "Jordan", a 2006 song from Bellowhead's *Burlesque*
- "Jordan, Minnesota", a 1986 song from Big Black's *Atomizer*

Mathematics

- Gauss–Jordan elimination, version of Gaussian elimination
- Jordan algebra, a non-associative algebra over a field
- Jordan curve theorem in topology
- Jordan decomposition (disambiguation), several measures
- Jordan measure or Jordan content, an early form of measure
- Jordan normal form or Jordan canonical form of a matrix
- Jordan's lemma in complex analysis
- Jordan's theorem (multiply transitive groups)
- Jordan–Schönflies theorem in geometric topology
- Jordan–Hölder theorem in group theory
- Jordan's theorem in economics

People

- Jordan (name), list of people with this surname or given name

People adopting name Jordan

- Jordan (Katie Price), English former glamour model
- Jordan (Pamela Rooke), model and actress related to the punk movement

Other

- Jordan almonds, a type of candy
- Jordan Grand Prix, which competed in Formula 1 from 1991-2005
- Jordan Motor Company, an automobile manufacturer of the 1920s
- Jordan College (disambiguation), several colleges both real and fictional
- Jordan, archaic slang for a chamber pot

互联网与WSD

- 但是，互联网存在如下不足
  - 互联网上存在大量的垃圾内容，需要过滤
  - 搜索引擎返回的结果页面数只是估计值，并且不断在变化
  - 搜索引擎可能关闭API，阻止访问
  - 访问互联网获取数据通常比较慢