自然语言处理应用
信息抽取

孙承杰
计算机科学与技术学院

主要内容

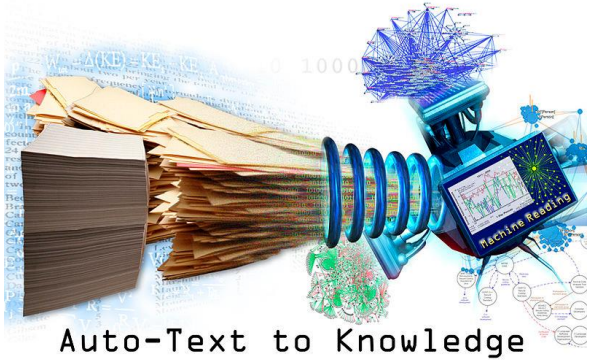
- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

主要内容

- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

信息抽取（IE）

- The goal of IE is to automatically extract structured information, i.e. categorized and contextually and semantically well-defined data from a certain domain, from unstructured machine-readable documents.
- 无结构数据结构化



信息抽取中的主要任务

- 命名实体识别
  - 识别和分类文本中出现的“实体提及”

昨天下午，市政协、市委统战部联合举办北京市全国政协委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。市政协主席吉林参加。

昨天下午，<ORG>市政协</ORG>、<ORG>市委统战部</ORG>联合举办<LOC>北京市</LOC> <ORG>全国政协</ORG>委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。<ORG>市政协</ORG>主席<PER>吉林</PER>参加。

信息抽取中的主要任务

- 实体链接
  - 将“实体提及”链接到知识库中对应的实体

昨天下午，<ORG>市政协</ORG>、<ORG>市委统战部</ORG>联合举办<LOC>北京市</LOC> <ORG>全国政协</ORG>委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。<ORG>市政协</ORG>主席<PER>吉林</PER>参加。

<PER>吉林</PER> ---> **吉林 （北京市政协党组书记、主席）**

信息抽取中的主要任务

- 关系抽取
  - 找到句子中有关系的两个实体，并识别出他们之间的关系类型

昨天下午，<ORG>市政协</ORG>、<ORG>市委统战部</ORG>联合举办<LOC>北京市</LOC> <ORG>全国政协</ORG>委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。<ORG>市政协</ORG>主席<PER>吉林</PER>参加。

R = （市政协，主席，吉林）

信息抽取中的主要任务

- 事件抽取
  - 事件的元素包括: 触发词、事件类型、论元及论元角色。
  - 事件抽取就要是找到一个事件对应的元素。

昨天下午，<ORG>市政协</ORG>、<ORG>市委统战部</ORG>联合举办<LOC>北京市</LOC> <ORG>全国政协</ORG>委员视察考察活动，围绕历史文化街区改造和疏解整治促提升专项行动进展等视察并座谈。<ORG>市政协</ORG>主席<PER>吉林</PER>参加。

信息抽取的发展

- MUC (Message Understanding for Comprehension)
- MET (Multilingual Entity Task Evaluation)
- ACE (Automatic Content Extraction)
- DUC (Document Understanding Conferences)
- TAC (Text Analysis Conference)

MUC (1)

- 美国政府支持的一个专门致力于真实新闻文本理解的例会。
  - 1991--1997
  - DARPA (Defense Advanced Research Projects Agency)
  - 组织对来自世界各地不同单位的消息理解系统进行系列化的评测活动。
- 主要的评测项目是从新闻报道中提取特定的信息，填入某种数据库中。
  - 评测语料大都出自各大通讯社发布的新闻。
  - 对每一条消息，由专业人员人工给出标准答案，然后将参测系统的输出结果与标准答案比较，按一定的评价指标给出所有系统的评测结果，其中最主要的指标是准确率、查全率等。
- MUC定义的概念、模型和技术规范对整个信息抽取领域起到了引领作用。

MUC (2)

Conference	Year	Text Source	Topic (Domain)
MUC-1	1987	Mil. reports	Fleet Operations
MUC-2	1989	Mil. reports	Fleet Operations
MUC-3	1991	News reports	Terrorist activities in Latin America
MUC-4	1992	News reports	Terrorist activities in Latin America
MUC-5	1993	News reports	Corporate Joint Ventures, Microelectronic production
MUC-6	1995	News reports	Negotiation of Labor Disputes and Corporate Management Succession
MUC-7	1997	News reports	Airplane crashes, and Rocket/Missile Launches

MUC (3)

- 5个典型的提取阶段：(MUC-7 IE Task Definition Version 5.1)
  - NE (Named Entities)
  - ER (Entity Relations)
  - Template Scenario (Event Structures)
  - Coreference (Identity descriptions)
  - Template Merger
- 具体提取哪些 NE，ER，Events 以及做哪些Coref, Merger 是任务相关的(每次MUC独立定义)。

MUC (4)

- 5个典型的提取阶段：

  - NE (Named Entities):提取文本中相关的命名实体，包括人名、机构/公司名称的识别

    国家财政部/Org 部长 项怀诚/Person

- ER (Entity Relations):提取命名实体之间的各种关系（事实）

  Post_of(部长,项怀诚)， *employee_of*(国家财政部,项怀诚)

  - Template Scenario (Event Structures)：事件

    召开会议(Time<…>, Spot<…>, Convener<…>, Topic<…>)

  - Coreference (Identity descriptions) :
    代词、名词共指

  - Template Merger : 相同事件的合并

MET

- MET: Multilingual Entity Task Evaluation
- 也是DARPA发起的一个测评项目。
- MET主要是对日语、汉语以及西班牙语等多语种新闻文献进行命名实体抽取。
- MET-1和MET-2测试分别于1996年和1998年进行。

ACE(1)

- ACE (Automatic Content Extraction)
  - A research program for developing advanced information extraction technologies convened by the NIST from 1999 to 2008。
- 关注三种信息的自动化内容抽取：
  - 网络上的在线新闻
  - 通过ASR（自动语音识别的）得到的广播新闻
  - 以及通过OCR（光学字符识别）得到的报纸新闻
- 两个目的：
  - 希望在自动化内容抽取基础之上，为数据挖掘、链接分析、自动摘要等打下基础
  - 通过将相应的信息提供给相应的分析师，以提高信息分析的能力。

ACE(2)

- 项目为期10年
  - ACE Phase-1(1999.7-2000.12)优先发展的是实体探测及追踪(EDT, Entity Detection and Tracking) 。
  - ACE Phase2(2001-2008)被称为EDT + RDC。其中RDC为Relation Detection and Characterization。
  - ACE第二阶段希望在第一阶段实体探测的基础之上，引入对实体关系的评测，需要能够将标识出的实体之间的关系揭示出来。

DUC

- Sponsored by the Advanced Research and Development Activity (ARDA), the conference series is run by the National Institute of Standards and Technology (NIST) to further progress in summarization and enable researchers to participate in large-scale experiments.
- 2000—2007
- In 2008, DUC became a Summarization track in the Text Analysis Conference (TAC)

TAC

- From 2008
- Grew out of NIST's Document Understanding Conference (DUC) and the Question Answering Track of TREC.
- A series of workshops that provides the infrastructure for large-scale evaluation of Natural Language Processing technology
- TAC's primary purpose is *not* competitive benchmarking; the emphasis is on advancing the state of the art through evaluation results

Goals of TAC

- to promote research in NLP based on large common test collections;
- to improve evaluation methodologies and measures for NLP;
- to build a series of test collections that evolve to anticipate the evaluation needs of modern NLP systems;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in NLP methodologies on real-world problems.

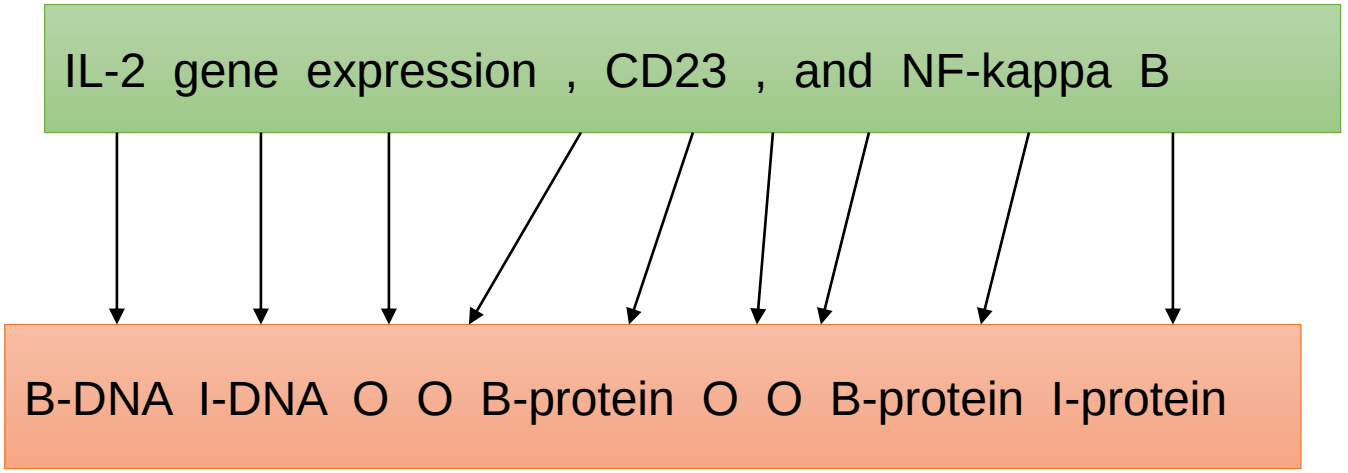
主要内容

- 信息抽取的定义、任务及发展
- 命名实体识别
- 实体链接
- 关系抽取

命名实体识别

- 定义
- 难点
- 应用
- 主要方法
- 评价

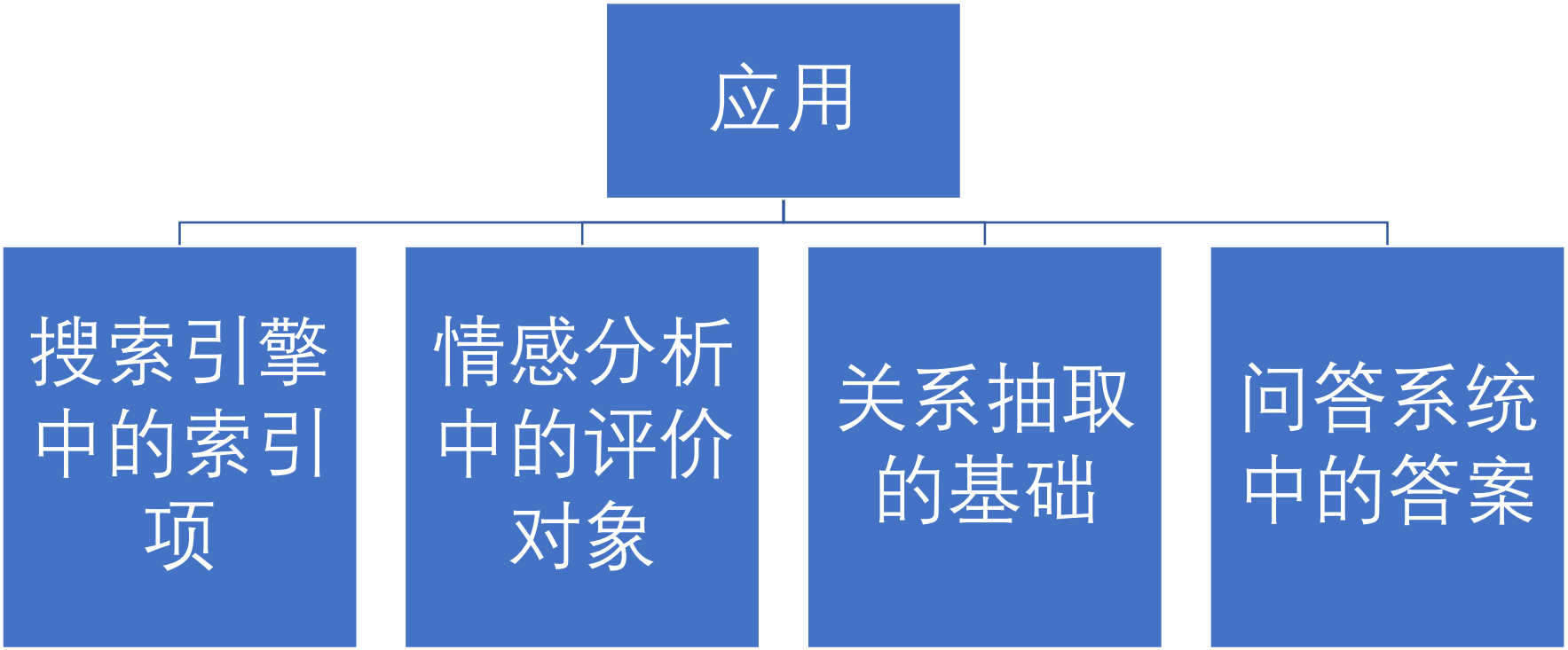
命名实体识别定义


IL-2 gene expression , CD23 , and NF-kappa B
B-DNA I-DNA O O B-protein O O B-protein I-protein

命名实体识别

- 难点
  - 种类繁多，命名方式灵活多样
  - 同一实体对应很多变体
  - 相同的词或者短语可以表示不同类别的实体
  - 存在嵌套
  - 语言不断进化，新的挑战不断出现

命名实体识别



命名实体识别

- 主要方法
  - 基于规则的方法
  - 基于词典的方法
  - 机器学习方法
    - 最大熵
    - 条件随机场
    - 深度学习

基于规则的方法

Rule: TheGazOrganization
Priority: 50
// Matches “The <in list of company names>”
({Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})
→ Organization

Rule: LocOrganization
Priority: 50
// Matches “London Police”
({DictionaryLookup = location | DictionaryLookup = country} {DictionaryLookup
= organization} {DictionaryLookup = organization}?) → Organization

- 规则来自GATE (General Arch. For Text Engineering), http://gate.ac.uk/
- 规则需要利用词性和词典

基于最大熵模型的方法

- Maximum Entropy (Maxent) Classifier
  - Maximum Entropy was first introduced to NLP area by Berger, et al (1996) and Della Pietra, et al. 1997.

$$\mathrm{P}(c|d) = \frac{1}{Z(d)} exp \sum_{i=1}^{K} \lambda_i f_i(c,d)$$

$$Z(d) = \sum_{c=1}^{C} exp \sum_{i=1}^{K} \lambda_i f_i(c,d)$$

Classify result

$$c^* = \underset{c \in \{1,2,\ldots,C\}}{\operatorname{argmax}} P(c|d)$$

最大熵模型

- Given this model form, we will choose parameters $\{\lambda_i\}$ that *maximize the conditional likelihood* of the training data according to this model.
- We construct not only classifications, but probability distributions over classifications.
  - There are other (good!) ways of discriminating classes – SVMs, boosting, even perceptrons – but these methods are not as trivial to interpret as distributions over classes.

最大熵模型

- 最大熵模型是一种分类模型，适用于很多自然语言处理任务:
  - Sentence boundary detection (Mikheev 2000)
    - Is a period end of sentence or abbreviation?
  - Sentiment analysis (Pang and Lee 2002)
    - Word unigrams, bigrams, POS counts, …
  - PP attachment (Ratnaparkhi 1998)
    - Attach to verb or noun? Features of head noun, preposition, etc.
  - Parsing decisions in general (Ratnaparkhi 1997; Johnson et al. 1999, etc.)

最大熵模型中的特征

- *features f* are elementary pieces of evidence that link aspects of what we observe *d* with a category *c* that we want to predict
- A feature is a function with a bounded real value
- Models will assign to each feature a *weight:*
  - A positive weight votes that this configuration is likely correct
  - A negative weight votes that this configuration is likely incorrect

特征举例

LOCATION *in Arcadia* LOCATION *in Québec* DRUG *taking Zantac* PERSON *saw Sue*

- $f_1(c, d) \equiv$

$[c = \text{LOCATION} \wedge w_{-1} = \text{“in”} \wedge \text{isCapitalized}(w)]$

- $f_2(c, d) \equiv$

$[c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$

- $f_3(c, d) \equiv$

$[c = \text{DRUG} \wedge \text{ends}(w, \text{“c”})]$

如何利用特征进行分类？

- Make a probabilistic model from the linear combination $\Sigma\lambda_i f_i(c,d)$

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

Makes votes positive

Normalizes votes

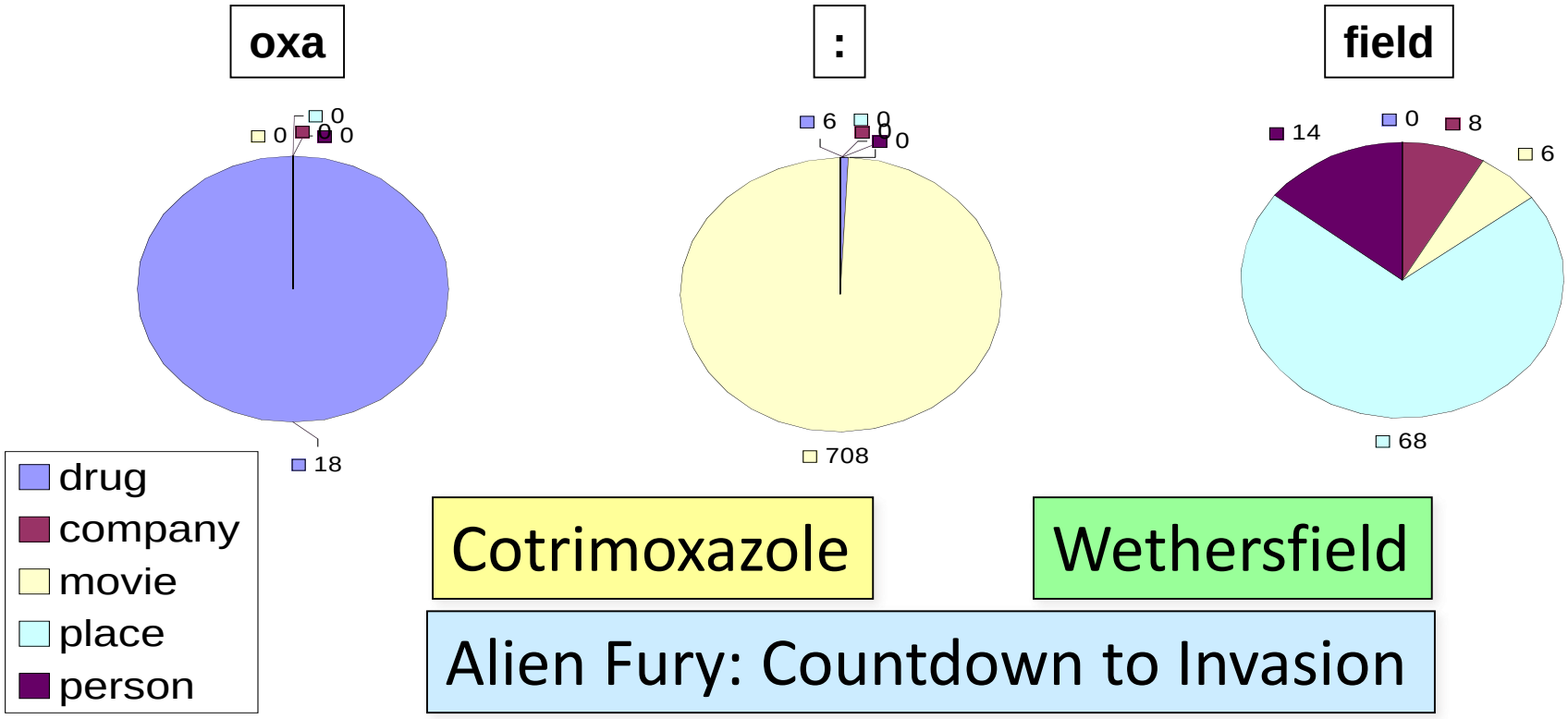
  - P(LOCATION|*in Québec*) = $e^{1.8}e^{-0.6}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.586$
  - P(DRUG|*in Québec*) = $e^{0.3}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.238$
  - P(PERSON|*in Québec*) = $e^{0}/(e^{1.8}e^{-0.6} + e^{0.3} + e^{0}) = 0.176$

- The weights are the parameters of the model, combined via a “soft max” function

最大熵模型里常用的特征类型

- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)
  - Substring of words
  - Word Shape
- Other kinds of inferred linguistic classification
  - Part-of-speech tags
- Label context
  - Previous (and perhaps next) label

Features: Word substrings



Features: Word shapes

- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

逻辑回归与最大熵模型

- Maxent models in NLP are essentially the same as multiclass logistic regression models in statistics (or machine learning)
  - If you haven't seen these before, don't worry, this presentation is self-contained!
  - If you have seen these before you might think about:
    - The parameterization is slightly different in a way that is advantageous for NLP-style models with tons of sparse features (but statistically inelegant)
    - The key role of feature functions in NLP

如何构建最大熵模型

- Define features (indicator functions) over data points
  - Features represent sets of data points which are distinctive enough to deserve model parameters.
    - Words, but also “word contains number”, “word ends with *ing*”, etc.
- Encode each Φ feature as a unique String
  - A datum will give rise to a set of Strings: the active Φ features
  - Each feature $f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$ gets a real number weight
- We concentrate on Φ features but the math uses i indices of f_i

如何构建最大熵模型

- Features are often added during model development to target errors
  - Often, the easiest thing to think of are features that mark bad combinations
- Then, for any given feature weights, we want to be able to calculate:
  - Data conditional likelihood
  - Derivative of the likelihood wrt each feature weight
    - Uses expectations of each feature according to the model
- We can then find the optimum feature weights (discussed later).

如何构建最大熵模型

- 指数模型似然度
- Maximum (Conditional) Likelihood Models :
  - Given a model form, choose values of parameters to maximize the (conditional) likelihood of the data.

$$\log P(C \mid D, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c \mid d, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

如何构建最大熵模型

- The (log) conditional likelihood of iid data (*C*,*D*) according to maxent model is a function of the data and the parameters λ:

$$\log P(C \mid D, \lambda) = \log \prod_{(c,d) \in (C,D)} P(c \mid d, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c \mid d, \lambda)$$

- If there aren't many values of *c*, it's easy to calculate:

$$\log P(C \mid D, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

如何构建最大熵模型

- 似然度的计算
- We can separate this into two components:

$$\log P(C \mid D, \lambda) = \sum_{(c,d) \in (C,D)} \log \exp \sum_i \lambda_i f_i(c,d) \; - \sum_{(c,d) \in (C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)$$

$$\log P(C \mid D, \lambda) = N(\lambda) - M(\lambda)$$

- The derivative is the difference between the derivatives of each component

The Derivative I: Numerator

$$\frac{\partial N(\lambda)}{\partial \lambda_i} = \frac{\partial \sum_{(c,d)\in(C,D)} \log \exp \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i}$$

$$= \frac{\partial \sum_{(c,d)\in(C,D)} \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i}$$

$$= \sum_{(c,d)\in(C,D)} \frac{\partial \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} \qquad = \sum_{(c,d)\in(C,D)} f_i(c,d)$$

Derivative of the numerator is: the empirical count (f_i, c)

The Derivative II: Denominator

$$\frac{\partial M(\lambda)}{\partial \lambda_i} = \frac{\partial \sum_{(c,d)\in(C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i}$$

$$= \sum_{(c,d)\in(C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \frac{\partial \sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i}$$

$$= \sum_{(c,d)\in(C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c',d)}{1} \frac{\partial \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i}$$

$$= \sum_{(c,d)\in(C,D)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c',d)}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'',d)} \frac{\partial \sum_i \lambda_i f_i(c',d)}{\partial \lambda_i}$$

$$= \sum_{(c,d)\in(C,D)} \sum_{c'} P(c'|d,\lambda)\, f_i(c',d)$$

= predicted count (f_i, λ)

The Derivative III

$$\frac{\partial \log P(C \mid D, \lambda)}{\partial \lambda_i} = \text{actual count}(f_i, C) - \text{predicted count}(f_i, \lambda)$$

- The optimum parameters are the ones for which each feature's predicted expectation equals its empirical expectation. The optimum distribution is:
  - Always unique (but parameters may not be unique)
  - Always exists (if feature counts are from actual data).
- These models are called maximum entropy models because we find the model having maximum entropy and satisfying the constraints:

$$E_p(f_j) = E_{\tilde{p}}(f_j), \forall j$$

参数优化

- We want to choose parameters $\lambda_1, \lambda_2, \lambda_3, \ldots$ that maximize the conditional log-likelihood of the training data

$$CLogLik(D) = \sum_{i=1}^{n} \log P(c_i \mid d_i)$$

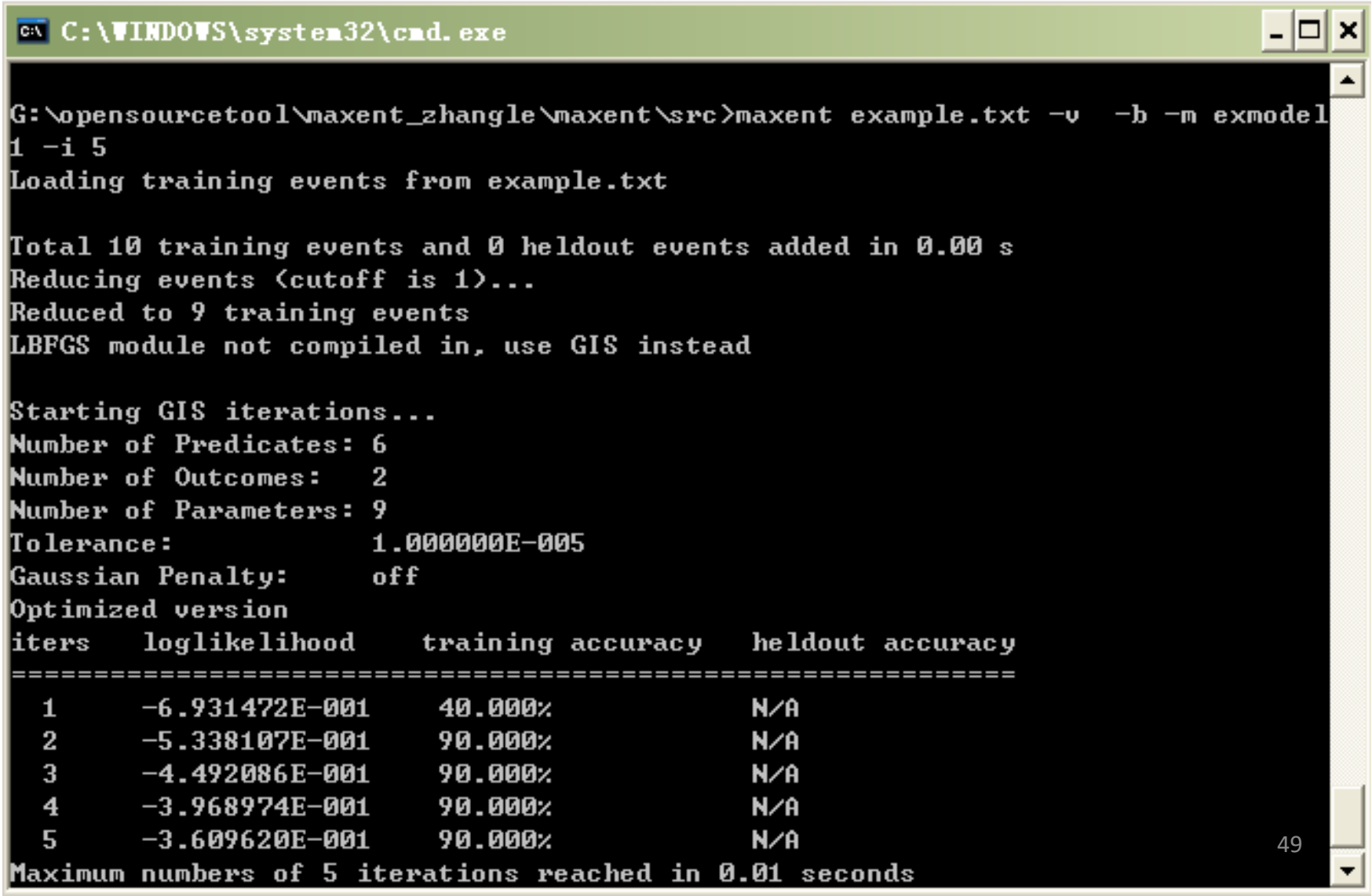
参数优化方法

- Limited-memory BFGS Method
  - Most effective [Mal 02]
  - Popular in late 1990s
- Iterative Scaling Methods
  - Improved Iterative Scaling
  - Generalized Iterative Scaling
  - Correction-Free GIS (SCGIS)
    - A faster GIS variant [Goodman 02] [Curran 03]
    - Needn't the constant C in GIS

最大熵模型使用举例

- Example data
  1. No_Umbrella Warm Dry
  2. No_Umbrella Cold Dry
  3. Umbrella Cold Rainy
  4. Umbrella Cold Dry
  5. No_Umbrella Warm Dry
  6. Umbrella Cold Dry Early
  7. Umbrella Cold Rainy Early
  8. No_Umbrella Cold Dry Late
  9. No_Umbrella Warm Rainy Late
  10. No_Umbrella Warm Dry Late

使用Zhangle的工具包进行训练



输出的模型长什么样子？

- Predicate(6):
  - Warm, Dry, Cold, Rainy, Early， Late
- Outcomes(2):
  - Umbrella, No_Umbrella
- Parameters(9):
  - 1 0 #warm only for label 0(No_Umbrella)
  - 2 0 1 #Dry both for label 0 and label 1
  - 2 0 1 #Cold both for label 0 and label 1
  - 2 0 1 #Rainy both for label 0 and label 1
  - 1 1 #Early only for label 1
  - 1 0 #Late only for label 0

```
#txt,maxent
6
Warm
Dry
Cold
Rainy
Early
Late
2
No_Umbrella
Umbrella
1 0  #warm only for label 0(No_Umbrella)
2 0 1 #Dry both for label 0 and label 1
2 0 1 #Cold both for label 0 and label 1
2 0 1 #Rainy both for label 0 and label 1
1 1 #Early only for label 1
1 0 #Late only for label 0
9
0.67742682914650987
0.31066551494595002
-0.5653652638130936
-0.4869650257145761
0.316244897311394
-0.31190677177012249
0.19747190915845117
0.74127573146151871
0.7414182802707221
```

CRFs [Lafferty, Pereira, and McCallum 2001]

- Another sequence model: Conditional Random Fields (CRFs)
- A whole-sequence conditional model rather than a chaining of local models.

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

- The space of c's is now the space of sequences
  - But if the features f_i remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming
- Training is slower, but CRFs avoid causal-competition biases
- These (or a variant using a max margin criterion) are seen as the state-of-the-art these days … but in practice usually work much the same as MEMMs.

命名实体识别的评价

	actual class（标准答案）	
predicted class（预测结果）	**tp** (true positive) Correct result	**fp** (false positive) Unexpected result
	fn (false negative) Missing result	**tn** (true negative) Correct absence of result

命名实体识别的评价

- 准确率 Precision(P)

$$P = \frac{tp}{tp + fp}$$

- 召回率 Recall(R)

$$R = \frac{tp}{tp + fn}$$

- F1度量 F1-measure(F1)

$$F1 = \frac{2PR}{P + R}$$

Useful Toolkits

- Maximum Entropy
  - http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html
- Conditional Random Fields
  - Mallet http://mallet.cs.umass.edu/fst.php
  - CRF++ https://taku910.github.io/crfpp/
  - LSTM-CRF https://github.com/abhyudaynj/LSTM-CRF-models