7.1 Noisy channel models

噪声信道模型（Noisy channel models）试图通过带噪声的输出信号恢复输入信号，因此常常从概率的角度进行描述，其图示及形式化定义如图 7.1.1 所示。

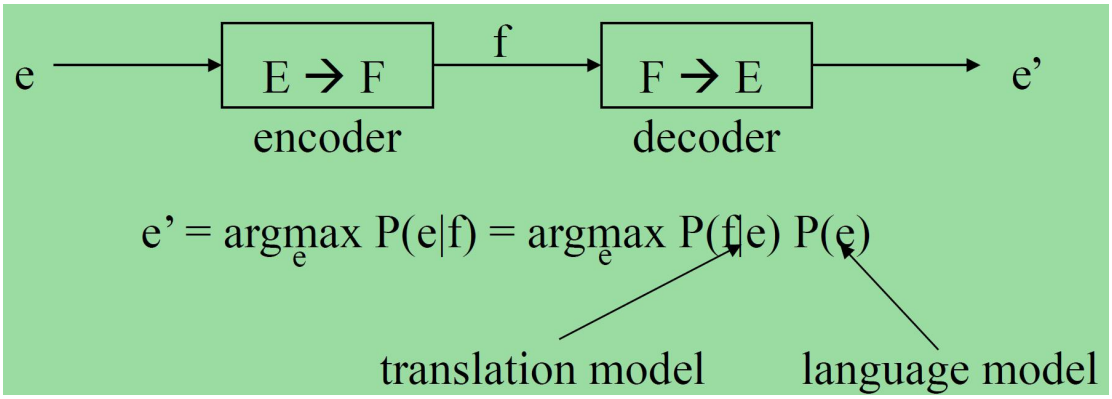

图 7.1.1

其中，语言模型（language model）在上一章已经有较为系统的介绍，而转移模型（translation model）则是这一章需要关注的。显然，较好的恢复结果 e'，对应的语言概率和转移概率都应当是较大的。

本节对 NCM 在机器翻译（Machine translation）和拼写纠错（Spelling correction）的应用各举一例，直观说明了合理与不合理的结果对应的 P(e|f)和 P(e)有明显差异。NCM 在 Handwriting recognition、Text generation、Text summarization 中也应用广泛。

7.2 Part of speech tagging

词性可以分为两类，即 open class 和 closed class，对于前者 you can add new words any time，对于后者 it's very difficult to invent new instance。

Open class 如:

- nouns, non-modal verbs, adjectives, adverbs

Closed class 如:

- prepositions, modal verbs, conjunctions, particles, determiners, pronouns

接下来列出了使用的词性标记，如图 7.2.1。

Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'oeuvre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*

Tag	Description	Example
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

图 7.2.1

我们需要对词性有一些直观的认识。统计表明，多词性的单词出现频率更高（11% of all types but 40% of all tokens in the Brown corpus are ambiguous）。同一单词的不同词性可能发音有所差异。词性标记作为一项基本的技术是很多应用的基础，常用的方法可分为三类：

- rule-based
- machine learning (e.g., conditional random fields, maximum entropy Markov models)
- transformation-based

在介绍具体方法之前，先讨论两个问题。第一个是，直观来看我们可以利用哪些信息进行词性标记。一类是目标单词（individual word）本身的信息，如词义（lexical information）、拼写方式（如-or 后缀的名词概率大）、大小写（capitalization，大写的可能是专有名词）。另一类是相邻单词（neighboring words）的信息，如它们的词性等等。

第二个问题是如何评价词性标记的效果。这似乎是显然的，因为我们可以明确判定算法给出的标记结果是否正确。然而一个问题是，POS 问题的 baseline 很高——利用两个简单的策略：tag each word with its most likely tag, and tag each OOV word as a noun，可以达到约 90%的正确率。目前，英语的 POS 可以达到 97%，略低于 98% 的 human performance（一些情况下可能有不止一种合理的解释，人们无法也没有必要对词性完全达成一致）。可以借鉴 4.5.2 节给出的评价标准$\kappa=(P(A)-P(E))/(1-P(E))$，或者更准确的$\kappa=(P(A)-P(E))/(P(M)-P(E))$，其中 P(E)是基准正确率，P(M)是 human performance。

接下来介绍了 rule-based POS method。这种方法很直观，对给定语句首先列出每个单词可能的词性，然后利用一组规则对词性组合空间剪枝，如果有必要再用后处理得到最终结果。规则集可能是庞大的（例如数百条），规则可能是“冠词后面不能跟动词”等。这种方法对全新的句子也能适用，但构建过程十分复杂，而且对不同语言需要较多调整。

其他方法将在后续内容中介绍。

7.3-7.4 Hidden Markov Models

本节介绍马尔科夫模型，我们需要回顾一下这部分内容。

首先从可见马尔科夫模型（Visible Markov Model）开始，可以用一个三元组来描述它：

- Q = sequence of states
- A = state transition probabilities
- Π = initial state probabilities

整个转移过程的参数是可见的，一阶马尔科夫过程的概率可以表示为：

$P(X_1,\cdots X_T) = P(X_1)\ P(X_2|X_1)\ P(X_3|X_2)\ \cdots\ P(X_T|X_{T-1})$

也常用状态转移图来表示，如图 7.3.1。

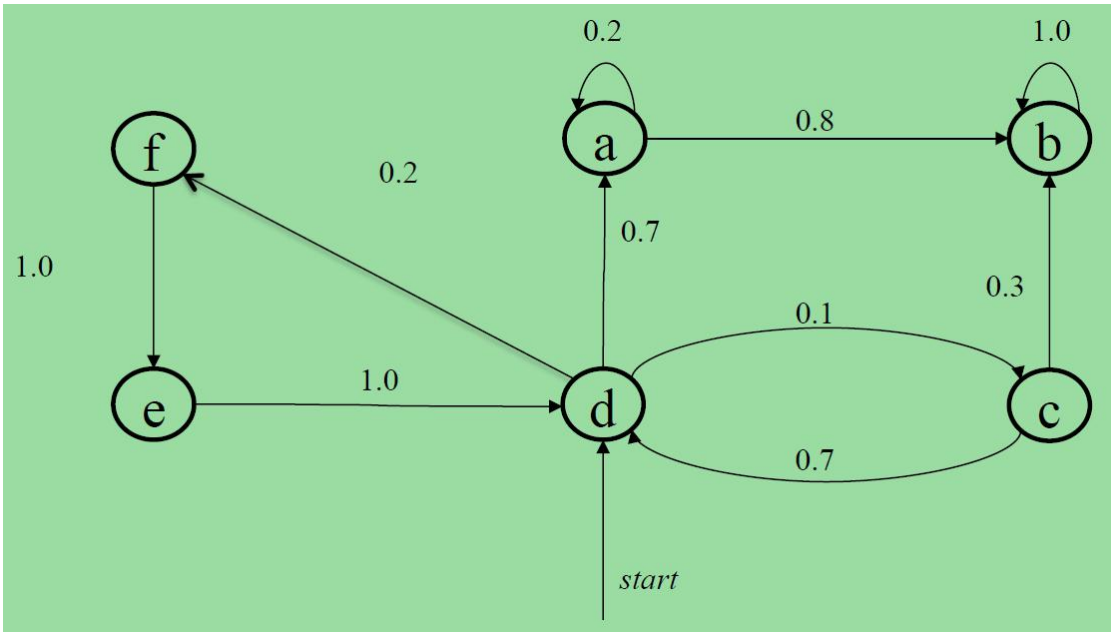

图 7.3.1

隐马尔科夫模型（Hidden Markov Model）描述的情形是，我们无法直接观察状态的变化（因此称为隐含），只能观察到间接的观测量，这个观测量的值是依赖于隐含的状态的。它可以用五元组来表示：

- Q = sequence of states
- O = sequence of observations, drawn from a vocabulary
- A = state transition probabilities
- B = symbol emission probabilities
- Π = initial state probabilities

这里可以记μ= (A,B,Π)，表示全体概率参数。图 7.3.2 是 HMM 的简单示意图。

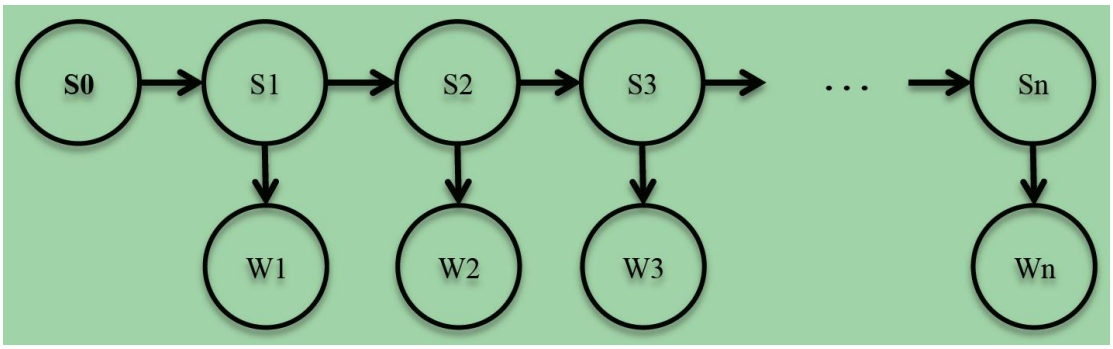

图 7.3.2

HMM 比 VMM 复杂很多，尤其是利用极大似然估计发射概率（emission probabilities）时面临数据稀疏的问题。基本的解决方法仍然是平滑估计，以及利用启发信息（如单词拼写等）。

HMM 常解决的三种问题：

– Given μ = (A,B,Π), find P(O|μ). 这是所谓的 Observation likelihood 问题，可用于对观测序列进行分类，即判断观测序列是哪个 HMM 生成的；

– Given O, μ, what is (X_1,…X_{T+1}). 所谓的解码（decoding）问题，得到最可能的状态序列；

– Given O and a space of all possible μ, find model that best describes the observations. 参数学习问题，也是最为困难的一种。

上面的三种问题中，第一种最为容易，但也不是分别计算每种状态组合对应的概率再求和，而是利用 Forward Algorithm，每次增加一个时刻迭代地覆盖所有的组合，这种方式相对显然不再赘述（课程里面并没有详细介绍）。

第二种问题的经典解法是 Viterbi Algorithm。对于一阶马尔科夫模型，长度为 n 的观测对应的最可能状态序列，是在第 n-1 个状态的基础上又做了一步转移，因此我们需要对第 n-1 个状态取遍所有可能值，并分别记录到达该状态的最可能路径。因此，这也是一个迭代的过程，思路是直观的不再赘述。

第三种问题的求解方法可以分为 supervised、unsupervised、semi-supervised 三种。Supervised 最常见的就是极大似然估计+平滑。

Unsupervised 是另一种思路，需要事先明确隐含状态的数目（对于词性标记而言没任何问题），以此来确定全体参数的规模和结构。另外，unsupervised 最终还是需要人工确定每个隐含状态到底实际对应何种含义（如哪种词性）。参数求解过程常用 EM 算法，即随机初

始化参数后，（E step）计算每个观测序列被各种状态组合生成的概率，（M step）根据这些概率重新估计参数。这一过程也有变种，但一种基本的方式是，在 E step 中直接认为观测序列是由其最似然的状态序列生成的，相当于利用当前参数对观测序列完成了自动标记，再在 M step 中重新估计参数。该方法得到的结果一般是局部最优解。

Semi-supervised 是上述两者的结合，它拥有少量的人工标记样本。

7.5 Statistical POS Tagging

首先结合具体例子简单介绍了怎样将 HMM 用于 POS，包括形式化的说明，转移概率、发射概率的计算示例。原理在之前已经明确，这里只是加强直观认识。

随后简要介绍了 Transformation-Based Learning，这是 Eric Brill 在 90 年代中期提出的方法。这里的 transformation 理解为变化、变通可能更好，因为该方法的基本思路是对每个单词都有默认的标记方式，但同时也训练了一些规则，当满足这些规则时默认的标记方式将会被更改。这些规则可能是考虑相邻单词的词性，考虑当前单词的拼写特点，等等。如何训练这些规则并没有介绍，应当去读论文。

这里给出了一些关于 POS 的认识。POS 的一个难题是模型在 new domains 上的泛化能力较弱，这样的情况如训练集与测试集不是同领域，或者遇到全新（novel）的单词。这些情况是正确率降低的主要原因。一种有价值的思路是利用 distributional clustering，即将语义相近的单词聚类再用于参数估计（combine statistics about semantically related words），有效的类别如 names of companies, days of the week, animals 等等。

最后给出了关于 HMM 的链接。

7.6 Information Extraction

信息提取（Information Extraction）一般针对非结构化（unstructured）或半结构化（semi-structured）的数据，这样才具有实际的价值。信息提取一般针对 entities, events, times, relations 等。其中命名实体（Named Entities）范围很广，包括 people、locations、organizations（team、newspapers、companies）、geo-political entities 等。

信息提取一般也作为序列标记（Sequence Labeling）问题，即对每个实体标记其实体类

型。该标记过程的一种方法是利用特征（上下文、目标单词），首先对实体进行定位或称为分割（Segmentation），对其进行分类（Classification）。

信息提取也有领域化的应用，例如在生物学文献中提取基因、蛋白质名称等。

7.7 Relation Extraction

本节开始简单介绍了关系提取（Relation Extraction）是提取实体之间的联系（links between entities），如某人工作于某公司（Works-for）、某厂制造某物（Manufactures）、某物位于某地（Located-at）等。相关的比赛如 MUC（Message Understanding Conference），评价标准是 precision、recall、F-measure。

本节后半部分又回到了关系提取（这个结构安排也是莫名），我们记录在这里。关系提取是很重要的任务，因为如果我们能够很好地提取实体及其关系，就可以自动构建知识数据库，就是如此振奋人心。而相关的方法可大概分为三类：using patterns、supervised learning、semi-supervised learning。

Using patterns 在这里比较狭义，其实是指特定模式构成的规则。一个例子是 Marti Hearst 在 90 年代的论文，文中将常见关系的模式列举出来，例如 X IS-A Y 关系的模式有：X and other Y、X or other Y、Y such as X、Y, including X、Y, especially X 等，通过正则匹配等方式即可提取关系。

Supervised learning 则类似于一般的序列标记方法，只是这里标记的数据是两个实体及其关系，因此用于构建分类器等模型的特征（尤其是上下文特征）可能有其特点。需要注意尽管两实体之间的内容是最重要的，但也常有关系关键词并不在两者中间。

Semi-supervised learning 中标记样本较少，按一定方式扩展标记规模。例如，已知两实体在某语句中的关系（如 **Beethoven** was born in December **1770**），可以认为其他语句中的这两个实体也是同种关系，这样就获得了描述这种关系的多种上下文（birth 等）。再根据这些上下文，去识别具有这种关系的其他实体对（如 **Laoyan**,**1994**）。概括起来，就是利用相同的实体、相同的上下文交替地扩展标记。

这里 Radev 提到了模板填充（Filling the Templates）的观点，并指出填充内容可能来自文本内容、可能是多值的、可能是预先定义好的候选值（如性别）。这样看来，不同的应用背景下信息提取过程可以从不同的角度来看待。某些通用的信息提取可能是序列标记的过

程，但某些特定领域的信息提取可能看做模板填充更为合适（如会议通告等）。注意，如果看做模板填充，评价结果时可能会考虑正确提取的数目和提取的完整性这两个不同层次。

既然可以看做序列标记的过程，那么处理方法一般就是特征提取＋诸如 HMM 等序列模型。这里顺便介绍了正则表达式（regular expressions），正如我们之前的认识，这是一种强大的匹配和提取特定串模式的工具。这里主要是介绍了常用符号并举例说明，不再赘述。

随后介绍了一种良好的广泛应用于序列标记的表达格式，称为 IOB format。这里 IOB 不是简称，而是代表格式中用到的三类标记：O 代表 other，标记我们不关心的内容；B 开头的标记（形如 B-Label），代表这个单词是该 label 的起始单词；I 开头的标记（形如 I-Label），代表这个单词是该 label 的一部分但不是起始单词。图 7.7.1 给出了一个例子。

```
file_id sent_id word_id iob_inner pos word
 0002  1   0 B-PER   NNP   Rudolph
 0002  1   1 I-PER   NNP   Agnew
 0002  1   2 O       COMMA COMMA
 0002  1   3 B-NP    CD    55
 0002  1   4 I-NP    NNS   years
 0002  1   5 B-ADJP  JJ    old
 0002  1   6 O       CC    and
 0002  1   7 B-NP    JJ    former
 0002  1   8 I-NP    NN    chairman
 0002  1   9 B-PP    IN    of
 0002  1 10 B-ORG   NNP   Consolidated
 0002  1 11 I-ORG   NNP   Gold
 0002  1 12 I-ORG   NNP   Fields
 0002  1 13 I-ORG   NNP   PLC
 0002  1 14 O       COMMA COMMA
 0002  1 15 B-VP    VBD   was
 0002  1 16 I-VP    VBN   named
 0002  1 17 B-NP    DT    a
 0002  1 18 I-NP    JJ    nonexecutive
 0002  1 19 I-NP    NN    director
 0002  1 20 B-PP    IN    of
 0002  1 21 B-NP    DT    this
 0002  1 22 I-NP    JJ    British
 0002  1 23 I-NP    JJ    industrial
 0002  1 24 I-NP    NN    conglomerate
 0002  1 25 O       .     .
```

图 7.7.1