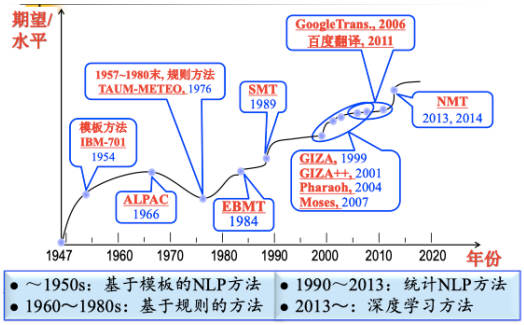
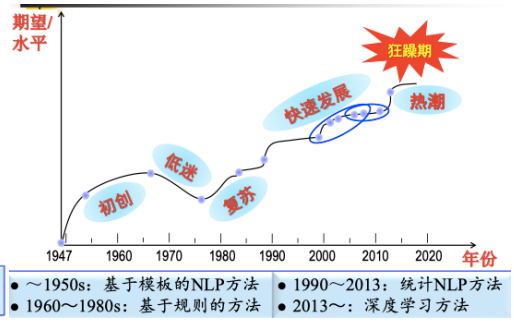
- NLU、CL、NLP 基本概念，学科的产生与发展

  NLU: 自然语言理解是探索人类自身语言能力和语言思维活动的本质，研究模仿人类语言认知过程的自然语言处理方法和实现技术的一门学科。它是人工智能早期研究的领域之一，是一门在语言学、计算机科学、 认知科学、信息论和数学等多学科基础上形成的交叉学科。

  CL(Computational Linguistics): 通过建立形式化的计算模型来分析、理解和生成自然语言的学科，是人工智能和语言学的分支学科。 计算语言学是典型的交叉学科，其研究常常涉及计算机科学、语言学、数学等多个学科的知识。与内容接近的学科自然语言处理相比较，计算语言学更加侧重基础理论和方法的研究。

  NLP: 自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。





- 形式文法与自动机之间的对应关系

  若 G 是一个正则文法，则存在一个有限自动机 M，使得：T(M) = L(G)。

  ◆由 G 构造 M 的一般步骤：

  (1) 令 $\Sigma = V_T$, $Q = V_N \cup \{T\}$, $q_0 = S$，其中，T 是一个新增加的非终结符。

  (2) 如果在 P 中有产生式 $S \to \varepsilon$，则 $F = \{S, T\}$，否则 $F = \{T\}$。

  (3) 如果在 P 中有产生式 $B \to a$，$B \in V_N$，$a \in V_T$，则 $T \in \delta(B, a)$。

  (4) 如果在 P 中有产生式 $B \to aC$，$B, C \in V_N$，$a \in V_T$，则 $C \in \delta(B, a)$。

  (5) 对于每一个 $a \in V_T$，有 $\delta(T, a) = \phi$。

  由 M 构造 G 的一般步骤：

  (1) 令 $V_N = Q$，$V_T = \Sigma$，$S = q_0$；

  (2) 如果 $C \in \delta(B, a)$，$B, C \in Q$，$a \in \Sigma$，则在P 中有产生式 $B \to aC$；

  (3) 如果 $C \in \delta(B, a)$，$C \in F$，则在P 中有产生式 $B \to a$。

- 编辑距离计算方法

  设 X 为拼写错误的字符串，其长度为 m, Y 为 X 对应的正确的单词(答案)，其长度为 n。则 X 和 Y 的编辑距离 ed(X[m], Y[n]) 定义为：从字符串 X 转换到 Y 需要的插入、删除、替换和交换两个相邻的基本单位(字符)的最小个数。如：

  ed (recoginze, recognize) = 1

  ed (sailn, failing) = 3

- 汉语自动分词中面临的主要问题、分词性能评价
  - 汉语自动分词中的主要问题：
    - ✓ 汉语分词规范问题（汉语中什么是词？）
    - ✓ 歧义切分字段处理

- ✧ 交集型歧义
  e.g. 结合成分子
- ✧ 组合型歧义
  e.g. 门把手弄坏了

✓ 未登录词的识别

➢ 分词与词性标注结果评价：

$$P=\frac{n}{N}\times 100\% \quad R=\frac{n}{M}\times 100\% \quad F-measure=\frac{(\beta^2+1)\times P\times R}{\beta^2\times P+R}\times 100\%$$

➢ 汉语自动分词基本算法

1. 最大匹配法
   a) 正向最大匹配算法 (Forward MM, FMM)
   b) 逆向最大匹配算法 (Backward MM, BMM)
   c) 双向最大匹配算法 (Bi-directional MM)
2. 最少分词法
3. 基于语言模型的分词方法
4. 基于 HMM 的分词方法

- 常见命名实体的类型及实体消歧方法

常见命名实体的类型：人名、机构名、地名、时间、 日期、货币和百分比。

实体消歧方法：

1、 基于聚类的实体消歧

同一指称项具有近似的上下文。核心问题：选取何种特征对于指称项进行表示

- ➢ 词袋模型（利用待消歧实体周边的词来构造向量）
- ➢ 语义特征（利用词袋和浅层语义特征，共同来表示指称项，利用余弦 相似度来计算两个指称项的相似度）
- ➢ 社会化网络（不同的人具有不同的社会关系）
- ➢ Wikipedia
- ➢ 多源异构知识

2、 基于实体链接的实体消歧

给定实体指称项和它所在的文本，将其链接到给定知识库中的相应实体上。

目前实体链接方法主要是如何更有效挖掘实体指称项信息， 如何更准确地计算实体指称项和实体概念之间的相似度。

- 词性标注面临的主要问题及其策略
  - ➢ 词性标注面临的主要问题：
    消除词性兼类歧义（形同音不同\同形、同音，但意义毫不相干\具有典型意义的兼类词\上述情况的组合）
  - ➢ 标注集的一般确定原则：标准性、兼容性、可拓展性
  - ➢ 词性标注方法：
    基于规则：手工编写消歧规则、根据词语的结构建立词性标注规则
    基于统计模型
    规则和统计方法相结合
    基于有限状态变换机
    基于神经网络

- 复杂特征集、合一运算及句子、词汇等描述方法
  - ➢ 功能合一文法 (Function Unification Grammar, FUG)

采用复杂特征集来描述词、句法规则、语义信息，以及句子的结构功能。采用合一运算对复杂特征集进行运算。

- ✓ 复杂特征集

  设α为一个功能描述 FD (Functional Description)， 当且仅当α可以表示为：

$$\begin{pmatrix} f_1 = v_1 \\ f_2 = v_2 \\ \cdots \\ f_n = v_n \end{pmatrix} \quad n \geq 1$$

  其中，f_i 表示特征名， v_i 表示特征值，且满足以下两个条件：

  (1) 特征名 fi 为原子，特征值 vi 为原子或另一个功能描述；

  (2) α(fi) = vi (i =1, ···, n)，读作：复杂特征集α中，特征 fi 的值等于 vi。

- ✓ 合一运算
  - ✧ 复杂特征集相容的定义

    若α、β均为复杂特征集，则α、β是相容的，当且仅当：

    (1) 如果α(f)=a,β(f)=b,且 a、b 都是原子,那么当且仅当 a = b 时α、β是相容的;

    (2) 如果α(f)、β(f)均为复杂特征集,α、β是相容的，当且仅当α(f)、β(f)相容。
  - ✧ 合一运算的递归定义

    (1) 在 a、b都是原子的情况下，如果 $a=b$，则 $a \cup b = a$，否则 $a \cup b = \varnothing$；

    (2) 如果α、β 均为复杂特征集， 则

    (a) 若 $\alpha(f) = v$，但 $\beta(f)$ 的值未经定义，则 $f = v$ 属于 $\alpha \cup \beta$；

    (b) 若 $\beta(f) = v$，但 $\alpha(f)$ 的值未经定义，则 $f = v$ 属于 $\alpha \cup \beta$；

    (c) 若 $\alpha(f) = v_1$，但 $\beta(f) = v_2$，且 v_1 与 v_2 相容(不相抵触)，则 $f = (v_1 \cup v_2)$ 属于 $\alpha \cup \beta$，否则，$\alpha \cup \beta = \varnothing$。
  - ✧ 合一运算的作用

    (1) 合并原有的特征信息，构造新的特征结构；

    (2) 检查特征的相容性和规则执行的前提条件是否满足，如果参与合一的特征相冲突，就立即宣布合一失败。

- ➢ 词汇功能语法（Lexical Functional Grammar, LFG）

  基本观点：句子由两个相对独立的层次来描述：

  (1) 成分结构层次（Constitute structure, c-结构）

  句法规则：用上下无关文法表示

  词法规则：由词典信息提供，它带语法功能结构的预示信息

  (2) 功能结构层次（Functional structure, f-结构）

  用以表示句子的功能关系。包含语法信息，也包含语义信息

  由 c-结构构造 f-结构：

  Step-1: 从 c-结构求出功能描述式(functional descriptions，简称 f-描述)；

  Step-2: 从 f-描述构造 f-结构。

- 语言模型的定义及其主要的数据平滑方法
  - ➢ n 元语法（n 元文法）实际上就是 n 个临近词构成的一个 n 元词组，或称词序列。

不失一般性，对于 n>2 的 n-gram，p(s) 可以分解为：

$$p(s) = \prod_{i=1}^{m+1} p(w_i \mid w_{i-n+1}^{i-1})$$

  - ➢ 数据平滑
    - ✓ 基本思想：调整最大似然估计的概率值,使零概率增值，使非零概率下调," 劫富济贫" ，消除零概率，改进模型的整体正确率。
    - ✓ 基本目标：测试样本的语言模型困惑度越小越好。

✓ 基本约束： $\sum_{w_i} p(w_i \mid w_1, w_2, \cdots, w_{i-1}) = 1$

✓ 数据平滑方法：

✧ 加 1 法(Additive smoothing)

基本思想: 每一种情况出现的次数加 1。

✧ 减值法/折扣法(Discounting)

基本思想 :修改训练样本中事件的实际计数，使样本中(实际出现的)不同事件的概率之和小于 1，剩余的概率量分配给未见概率。

1. Good-Turing 估计

$$\text{由于，} N = \sum_{r=0}^{\infty} n_r r^* = \sum_{r=0}^{\infty} (r+1) n_{r+1} \quad \text{所以，} r^* = (r+1)\frac{n_{r+1}}{n_r} \quad \sum_{r>0} n_r \times p_r = 1 - \frac{n_1}{N} < 1$$

对非 0 事件按公式削减出现的次数，节留出来的概率均分给 0 概率事件。

2. Back-off (后备/后退)方法（Katz 后退法）

基本思想：当某一事件在样本中出现的频率大于阈值 K (通常取 K 为 0 或 1)时，运用最大似然估计的减值法来估计其概率，否则，使用低阶的，即 (n-1)gram 的概率替代 n-gram 概率，而这种替代需受归一化因子α的作用。

$$p_{\text{katz}}(w_i \mid w_{i-1}) = \begin{cases} d_r \dfrac{C(w_{i-1}w_i)}{C(w_{i-1})} & \text{if } C(w_{i-1}w_i) = r > 0 \\ \alpha(w_{i-1}) p_{\text{ML}}(w_i) & \text{if } C(w_{i-1}w_i) = 0 \end{cases}$$

对非 0 事件按 Good-Turing 法计算减值，节留出来的概率按低阶分布分给 0 概率事件。

3. 绝对减值法 (Absolute discounting)

基本思想：从每个计数 r 中减去同样的量，剩余的概率量由未见事件均分。

$$p_r = \begin{cases} \dfrac{r-b}{N} & \text{当 } r > 0 \\ \dfrac{b(R-n_0)}{Nn_0} & \text{当 } r = 0 \end{cases}$$

对非 0 事件无条件削减某一固定的出现次数值，节留出来的概率均分给 0 概率事件。

4. 线性减值法 (Linear discounting)

基本思想：从每个计数 r 中减去与该计数成正比的量(减值函数为线性的)，剩余概率量α被 n 0 个未见事件均分。

$$p_r = \begin{cases} \dfrac{(1-\alpha)r}{N} & \text{当 } r > 0 \\ \dfrac{\alpha}{n_0} & \text{当 } r = 0 \end{cases}$$

绝对减值法产生的 n-gram 通常优于线性减值法。

对非 0 事件根据出现次数按比例削减次数值，节留出来的概率均分给 0 概率事件。

✧ 删除插值法 (Deleted interpolation)

基本思想：用低阶语法估计高阶语法，即当 3-gram 的值不能从训练数据中准确估计时，用 2-gram 来替代， 同样，当 2-gram 的值不能从训练语料中准确估计时， 可以用 1-gram 的值来代替。

$$p(w_3 \mid w_1 w_2) = \lambda_3 p'(w_3 \mid w_1 w_2) + \lambda_2 p'(w_3 \mid w_2) + \lambda_1 p'(w_3) \quad \cdots (5$$

其中， $\lambda_1 + \lambda_2 + \lambda_3 = 1$

- HMM、ME、CRFs 等概率图模型的构成及其之间的区别

HMM 是一种生成式模型，利用转移矩阵和生成矩阵建模相邻状态的转移概率和状态到观察的生成概率。无法利用复杂特征（只有两个矩阵建模）

ME 是一种判别式模型，可以使用任意的复杂特征（特征函 数），但是 ME 只能建模观察序列和某一状态的关系，状态 之间的关系无法得到充分利用。
CRF 是一种判别式模型，可以使用任意的复杂特征（特征函 数）（虽然每个位置有一个矩阵，但是矩阵的元素是由特 征函数和权重计算得到，我们可以任意地定义特征函数从 而考虑各种特征），可以建模观察序列和多个状态的关系 ，考虑了状态之间的关系。

- 语料库的类型及其典型语料库和语言知识库
  - 语料库的类型：
    - ✓ 按内容构成和目的划分
      - 异质的 (heterogeneous)
      - 同质的(homogeneous)
        e.g. 美国的 TIPSTER 项目只收集军事方面的文本
      - 系统的(systematic)
      - 专用的(specialized)
        e.g. 北美的人文科学语料库
    - ✓ 按语言种类划分
      - 单语的
      - 双语的或多语的
        e.g. 口语语料库：BTEC
    - ✓ 是否标注
      - 具有词性标注
      - 句法结构信息标注(树库)
        e.g. 宾夕法尼亚大学(UPenn)树库(Tree Bank)
      - 语义信息标注
    - ✓ 共时语料库与历时语料库
    - ✓ 平衡语料库
      e.g. 台湾中研院平衡语料库（Sinica Corpus）
    - ✓ 平行语料库
      e.g. 国际英语语料库
  - 词汇知识库
    - ✓ WordNet（普林斯顿大学）
      解决词典中同义信息的组织问题。
    - ✓ 知网（HowNet)
      它所着力要反映的是概念的共性和个性，同时知网还着力要反映概念之间和概念的属性之间的各种关系。

- 词义消歧的基本思路和实现方法
  - 基本思路
    每个词表达不同的含意时其上下文（语境）往往不同，也就是说，不同的词义对应不同的上下文，因此，如果能够将多义词的上下文区别开，其词义自然就明确了。
  - 实现方法
    - ✓ 基于上下文分类的消歧方法
      1. 基于贝叶斯分类器
         a) 对于词典中所有的词 v k 利用训练语料计算
         $$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$
         b) 对于 w 的每个语义 s i 计算：

$$p(s_i) = \frac{N(s_i)}{N(w)}$$

c) w 的每个语义 s i 计算 p(s i)，并根据上下文中的每个词 v k 计算 p(w|s i)，选择：

$$\hat{s}_i = \arg\max_{s_i}\left[p(s_i)\prod_{v_k \in C} p(v_k \mid s_i)\right]$$

2. 基于最大熵的消歧方法

基本思想：在已知部分知识的前提 下，关于未知分布最合理的推断应该是符合已知知识 最不确定或最大随机的推断。

$$p^*(a \mid b) = \frac{1}{Z(b)}\exp(\sum_{j=1}^{l}\lambda_j \cdot f_j(a,b))$$

$$Z(b) = \sum_a \exp(\sum_{j=1}^{l}\lambda_j \cdot f_j(a,b))$$

a) 确定特征函数

b) 获取λ参数（GIS 算法）

3. 基于词典的词义消歧
   a) 基于语义定义的消歧
   b) 基于义类辞典(thesaurus) 的消歧
   c) 基于双语词典的消歧
   d) Yarowsky 消歧算法
4. 无监督的词义消歧方法

- 短语结构分析器基本方法（Chart Parser / CYK Parser）
  - Chart Parser
    - ✓ 自底向上的 chart 分析算法
      1. 给定一组 CFG 规则：XP->α1…αn (n>=1)
      2. 给定一个句子的词性序列：S=W1W2…Wn
      3. 构造一个线图
      4. 建立一个二维表：记录每一条边的起始位置和终止位置
    - ✓ 数据结构
      - ✧ 线图(Chart)：保存分析过程中已经建立的成分(包括终结符和非终结符)、位置(包括起点和终点)。通常以 n×n 的数组表示(n 为句子包含的词数)。
      - ✧ 代理表(待处理表)(Agenda)：记录刚刚得到的一些重写规则所代表的成分，这些重写规则的右端符号串与输入词性串(或短语标志串)中的一段完全匹配， 通常以栈或线性队列表示。
      - ✧ 活动边集(ActiveArc)：记录那些右端符号串与输入串的某一段相匹配，但还未完全匹配的重写规则， 通常以数组或列表存储。
  - CYK Parser
    - ✓ 算法：
      - ✧ 对 Chomsky 文法进行范式化：

        $A \rightarrow w$ 或 $A \rightarrow BC$

        $A, B, C \in V_N$， $w \in V_T$， $G=(V_N, V_T, P, S)$
      - ✧ 自下而上的分析方法
      - ✧ 构造 (n+1)×(n+1) 识别矩阵，n 为输入句子长度。假设输入句子 x=w1w2…wn，wi 为构成句子的单词，n=|x|。
    - ✓ 识别矩阵的构成
      - ✧ 方阵对角线以下全部为 0
      - ✧ 主对角线以上的元素由文法 G 的非终结符构成

✧ 主对角线上的元素由输入句子的终结符号(单词) 构成

✓ 识别矩阵构造步骤

1. 首先构造主对角线，令 t 0,0 =0，然后，从 t 1,1 到 t n,n 在主对角线的位置上依次放入输入句子 x 的单词 wi。
2. 构造主对角线以上紧靠主对角线的元素 t i, i+1，其中，i = 0, 1, 2 , …, n-1。对于输入句子 x = w1 w2 … wn，从 w1 开始分析。如果有 A->w i+1，则 t i, i+1=A。
3. 按平行于主对角线的方向，一层一层地向上填写矩阵的各个元素 t i,j，其中，i= 0,1, …,n-d, j=d+i, d=2, 3, …, n。如果存在一个正整数 k, i+1<=k <=j-1，在文法 G 的规则集中有产生式 A->BC, 并且，B∈t i,k , C∈t k,j， 那么，将 A 写到矩阵 t i,j 位置上。

判断句子 x 由文法 G 所产生的充要条件是：t 0,n=S。

- 依存句法理论的基本思想

在依存语法理论中，“依存”就是指词与词之间支配与被支配的关系，这种关系不是对等的，而是有方向的。处于支配地位的成分称为支配者 (governor, regent, head)，而处于被支配地位的成分称为从属者(modifier, subordinate, dependency)。

两个有向图用带有方向的弧(或称边，edge)来表 示两个成分之间的依存关系，支配者在有向弧的发出端，被支配者在箭头端，我们通常说被支配者依存于支配者。

- 依存关系分析与短语结构分析结果之间的关系

短语结构可转换为依存结构

➢ 实现方法：

(1) 定义中心词抽取规则，产生中心词表;

(2) 根据中心词表，为句法树中每个节点选择中心子节点;

(3) 将非中心子节点的中心词依存到中心子节点的中心词上，得到相应的依存结构。

- 句法分析结果评价方法

➢ 无标记依存正确率(unlabeled attachment score, UA)：
所有词中找到其正确支配词的词所占的百分比， 没有找到支配词的词(即根结点)也算在内。

➢ 带标记依存正确率(labeled attachment score, LA)：
所有词中找到其正确支配词并且依存关系类型也标注正确的词所占的百分比，根结点也算在内。

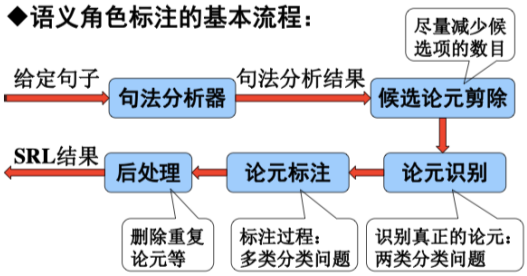
➢ 依存正确率(dependency accuracy, DA)：
所有非根结点词中找到其正确支配词的词所占的百分比。

➢ 根正确率(root accuracy, RA)：
所有句子中找到正确根节点的句子所占的百分比。

➢ 完全匹配率(complete match, CM)：所有句子中无标记依存结构完全正确的句子所占的百分比。
所有句子中无标记依存结构完全正确的句子所占的百分比。

- 语义角色标注基本方法

基本任务：以句子为分析单位，以句子中的谓词为核心， 分析句子中的其他成分与谓词之间的关系。

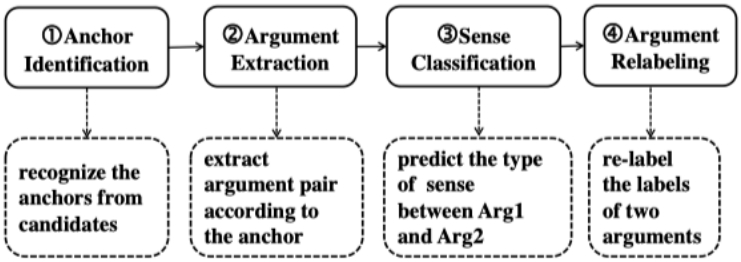


1. 基于短语结构句法分析的 SRL 方法

一个论元被表示为连续的几个词（短语）和一个语义角色标签

2. 基于依存关系的 SRL 方法
   一 个论元被表示为一个中心词和一个语义角色标签
3. 基于语块分析的 SRL 方法
   将语义角色标注作为一个序列标注问题来解决。一般采用 BIO(分别表示：开始、属于、不属于 ）的方式来定义序列标注的标签集，将不同的语块赋予不同的标签。不需要剪除候选论元，论元识别和标注同时进行。

- 篇章理论
  篇章关系分析框架



- 机器翻译的基本方法（原理）
  - 直接转换法
    从源语言句子的表层出发，将单词、短语或句子 直接置换成目标语言译文，必要时进行简单的词序调整。对原文句子的分析仅满足于特定译文生成的需要。 这类翻译系统一般针对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。
  - 基于规则的翻译方法
    1. 对源语言句子进行词法分析
    2. 对源语言句子进行词法分析
    3. 源语言句子结构到译文结构的转换
    4. 译文句法结构生成
    5. 源语言词汇到译文词汇的转换
    6. 译文词法选择与生成
  - 基于中间语言的翻译方法
    方法：输入语句 -> 中间语言 -> 翻译结果
    代表系统：JANUS (CMU) 早期版本
  - 基于事例的翻译方法
    方法：输入语句 -> 与事例相似度比较 -> 翻译结果
    资源：大规模事例库
  - 统计机器翻译（生成式模型）

$$T' = \arg\max_{T} P(T) \times P(S|T)$$

语言模型 Language model, LM　　翻译模型 Translation model, TM

引入对位模型计算翻译概率 P(S|T)

$$P(S|T) = \sum_{A} P(S, A|T)$$

$$P(S, A|T) = p(m|T) \times P(A|T, m) \times P(S|T, A, m)$$

对位模型　　词汇翻译模型

  - 基于短语的翻译模型（判别式模型）
    - ✓ 基于最大熵的方法
      解决方法：将语言模型概率、正向和反向翻译概率都视为一个特征，采用最大熵方法建模

$$T' = \underset{T}{\operatorname{argmax}} P(T|S) = \underset{T}{\operatorname{argmax}} \frac{exp\{\sum_1^M \lambda_m h_m(T,S)\}}{\sum_{T^*} exp\{\sum_1^M \lambda_m h_m(T^*,S)\}}$$
$$= \underset{T}{\operatorname{argmax}} \left\{ \sum_1^M \lambda_m h_m(T,S) \right\}$$

✓ 基于短语的翻译模型（短语：连续的词串）

$$T' = \underset{T}{\operatorname{argmax}} P(T|S)$$
$$= \underset{T,S_1^K}{\operatorname{argmax}} P(T, S_1^K|S)$$
$$= \underset{T,S_1^K,T_1^K,T_1^{K'}}{\operatorname{argmax}} \underbrace{P(S_1^K|S)}_{\text{短语划分模型}} \underbrace{P(T_1^K|S_1^K,S)}_{\text{短语翻译模型}} \underbrace{P(T_1^{K'}|T_1^K,S_1^K,S)}_{\text{短语调序模型}} \underbrace{P(T|T_1^{K'},T_1^K,S_1^K,S)}_{\text{目标语言模型}}$$

✧ 短语翻译规则抽取

根据源语言中任一短语，根据词语对齐找到目标语言句子中的对齐片段，若满足对齐一致性，则为一条短语翻译规则。

✧ 短语翻译概率估计：四个翻译概率（最大似然）

1. 正向、逆向短语翻译概率 $p(t|s)$, $p(s|t)$
2. 正向、逆向词汇化翻译概率 $p$lex(t|s),plex(s|t)

✧ 短语调序模型两种常用方法：

1. 距离跳转模型：d=next_begin – last_end - 1
2. 分类模型: Monotone (M) Swap (S) Discontinuous (D)

- 机器翻译译文评价方法
  - 主观评测：（1）流畅度 （2）充分性 （3）语义保持性
  - 客观评测

1. 句子错误率：译文与参考答案不完全为错误句子。错误句子占全部译文的比率。
2. 单词错误率(Multiple Word Error Rate on Multiple Reference，记作 mWER)：分别计算译文与每个参考译文的编辑距离，以最短的为评分依据，进行归一化处理
3. 与位置无关的单词错误率 (Position independent mWER，记作 mPER)：不考虑单词在句子中的顺序
4. METEOR 评测方法：对候选译文与参考译文进行词对齐，计算词汇完全匹配、词干匹配、同义词匹配等各种情况的准确率 (P)、召回率(R)和 F 平均值

$$F = \frac{10PR}{R+9P} \qquad Score = F \times (1 - Penalty)$$
$$Penalty = 0.5 \times \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

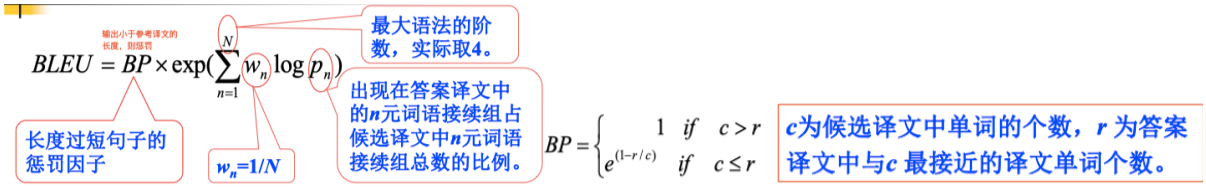
5. BLEU 评价方法

基本思想：将机器翻译产生的候选译文与人翻译的多个参考译文相比较，越接近，候选译文的正确率越高。

实现方法：统计同时出现在系统译文和参考译文中的 n 元词的个数，最后把匹配到的 n 元词的数目除以系统译文的 n 元词数目，得到评测结果。

修正的计算一元语法精确度的方法：针对某个待评测的系统译文句子，首先统计每个单词在所有参考译文中出现次数的最大值 Max_Ref_Count，然后，统计该单词在系统译文中出现的总次数 Count，取 Count 和 Max_Ref_Count 两者中小的一个，即

Count_clip=min(Count, Max_Ref_Count)



BLEU 分值范围：0 ~ 1，分值越高表示译文质量越好，分值越小，译文质量越差。

6. NIST 评测方法

   因为 n 值较大的统计单元出现的概率较低，因此用 n-gram 同现概率的算术平均值取代几何平均值。如果一个 n 元词在参考译文中出现的次数越少，表明它所包含的信息量越大，那么，它对于该 n 元词就赋予更高的权重。

- 自动文摘类型及基本方法
  - 文本摘要分类
    1. 文档数目：单文档摘要、多文档摘要
    2. 输入语言与输出语言的关系：单语摘要、跨语言摘要、多语言摘要
    3. 是否有用户输入：通用摘要、用户查询摘要
    4. 摘要方法：抽取式摘要、压缩式摘要、理解式摘要
    5. 摘要长度：标题式摘要、短摘要、长摘要
  - 文本摘要方法
    - ✓ 抽取式摘要
      直接从原文中抽取已有的句子组成摘要
      启发式规则/机器学习方法/图模型方法
    - ✓ 压缩式摘要
      抽取并简化原文中的重要句子构成文摘
      可视为树结构的精简问题/可视为 01 序列标注任务
    - ✓ 理解式摘要
      改写或重新组织原文内容形成最终文摘
      基于谓词论元结构的理解式摘要
    - ✓ 端到端摘要方法

- 文本分类的基本原理

  文本分类系统基本框架：文本 -> 文本表示 -> 特征选择 -> 分类算法
  - 文本表示
    - ✓ 向量空间模型（VSM）也称词袋模型（BOW）
    - ✓ 词频（Term Frequency, TF）
    - ✓ 布尔变量（是否出现）
    - ✓ 逆文档频率（Inverse Document Frequency,IDF）
    - ✓ TF-IDF

      $$\omega_{ki} = tf_{ki} \cdot \log \frac{N}{df_i}$$
  - 特征选择
    - ✓ 文档频率（Document Frequency, DF）
      根据训练语料中的文档频率，对所有特征进行排序
    - ✓ 互信息（Mutual Information, MI）
      互信息是关于两个随机变量互相依赖程度的一种度量

$$I(X,Y)=H(X)-H(X\mid Y)=\sum_{y}\sum_{x}P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

$$MI(t_i,c_j)=\log\frac{P(t_i,c_j)}{P(t_i)P(c_j)}\approx\log\frac{A_{ij}N_{all}}{(A_{ij}+C_{ij})(A_{ij}+B_{ij})}$$

■ A 关于特征 t_i 与类别 c_j的统计表

特征 \ 类别	c_j	$\bar{c}_j$
t_i	A_{ij}	B_{ij}
$\bar{t}_i$	C_{ij}	D_{ij}

$$P(c_j)\approx(A_{ij}+C_{ij})/N_{all}$$
$$P(t_i)\approx(A_{ij}+B_{ij})/N_{all}$$
$$P(\bar{t}_i)\approx(C_{ij}+D_{ij})/N_{all}$$
$$P(c_j\mid t_i)\approx\frac{A_{ij}+1}{A_{ij}+B_{ij}+C}$$
$$P(c_j\mid \bar{t}_i)\approx\frac{C_{ij}+1}{C_{ij}+D_{ij}+C}$$ +1平滑

✓ 信息增益（Information Gain, IG）

IG 衡量特征能够为分类系统带来多少信息

$$IG(t_i)=\{-\sum_{j=1}^{C}P(c_j)\log P(c_j)\}+\{P(t_i)[\sum_{i=1}^{C}P(c_j\mid t_i)\log P(c_j\mid t_i)]+P(\bar{t}_i)[\sum_{j=1}^{C}P(c_j\mid \bar{t}_i)\log P(c_j\mid \bar{t}_i)]\}$$

✓ Chi-Square 统计（Chi-Square Statistics ， CHI）

➢ 分类算法

✓ 生成式模型

学习算法：最大似然、最大后验

✧ 朴素贝叶斯

$$P(c_j\mid \mathbf{x})=\frac{P(\mathbf{x},c_j)}{P(\mathbf{x})}\propto P(\mathbf{x},c_j)=P(c_j)\prod_{i=1}^{M}P(w_i\mid c_j)^{N(w_i)}$$

最大似然估计：

$$P(c_j)\approx\frac{1+N(c_j)}{C+N_{all}}\qquad P(w_i\mid c_j)\approx\frac{1+N(w_i,c_j)}{M+\sum_{i'=1}^{M}N(w_{i'},c_j)}$$

✓ 判别式模型

学习算法：梯度下降、牛顿法

✧ 线性判别函数（Linear Discriminate Function）

$$g_j(\mathbf{x})=p(c_j\mid \mathbf{x})=\sum_{l=1}^{2}w_{jl}x_l+w_{j0}=\sum_{l=0}^{2}w_{jl}x_l,\ j=0,1,2$$

✧ 支持向量机（Support Vector Machine）

$$\min\frac{1}{2}\|\mathbf{w}\|^2\quad s.t.\ y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1,i=1,\ldots,n$$

✧ 最大熵模型（Maximum Entropy）

- 知识图谱概念

知识图谱以结构化三元组的形式存储现实世界中的实体及其关系，表示为 $\mathcal{G}=(\mathcal{E},\mathcal{R},\mathcal{S})$ ，三元组通常描述了一个特定领域中的事实，由头实体、尾实体和描述这两个实体之间的关系组成。关系有时也称为属性，尾实体被称为属性值。从图结构的角度看，实体是知识图谱中的节点，关系是连接两个节点的有向边。它是一个知识系统，以一种统一的方式表示知识框架和知识 实例两个层面的知识内容，各个具体实例数据只有在满足系统约定的 “框架” 约束下运用才能体现 “知识” 。知识图谱中的知识定义和实例数据及其相关的配套标准、技术、

应用系统共同构成广义的知识图谱。

- 问答系统基本原理及实现方法

  输入：自然语言的问句，而非关键词的组合

  输出：直接答案，而非文档集合

  - 基于知识推理的问答系统

  主要特点：答案或者从知识库中检索得到，或者在知识库上经过推理得到

  - 问答式检索系统

    信息检索 + 信息抽取

    信息检索 + 模式匹配

    信息检索 + 自然语言处理技术

    基于统计翻译模型的问答技术
  - 社区问答系统

    指用户之间通过提出和回答问题的方式共享和积累知识，从而提供知识交流与信息服务的社会化系统
    - ✓ 预处理（问题分类、作弊检测）
    - ✓ 与回答新提交问题相关的研究（相似问题检索、答案质量评估）

      主要方法：采用统计机器学习方法，比如分类或回归等
    - ✓ 与用户体验相关的研究（用户满意度预测）
  - 知识库问答系统

    自然语言问句到结构化查询语句（SPRAQL）的映射问题
    - ✓ 基于符号逻辑的技术

    系统框架：问题预处理（问题类型、去除无用词等）->短语检测&资源映射&特征提取->MLN 联合消歧->构造查询图->生成查询
    - ✓ 基于数值运算的技术

    对问句和答案（对应的子图）在同一个空间中 进行联合表示学习，学习问句中词以及知识库 中概念和关系的表示和匹配。
  - 阅读理解式问答系统
    - ✓ 基于传统特征工程的方法
    - ✓ 基于神经网络的方法
  - 对话系统

    基本架构：语音识别、自然语言理解、对话管理、自然语言生成、语音合成

    对话管理（状态跟踪和学习）的三类方法：
    - ✓ 有限状态机
    - ✓ 基于框架的方法
    - ✓ 统计方法（利用统计框架从大量的对话语料中自动学习对话管理模型）

    聊天机器人：

    基于检索（预定义一个 post-response 库）

    基于生成（不依赖于预定义的反应库，重新开始生成新的反应）