Lecture12 Information from parts of words: Subword Models

- 语音学 音素
  分类感知——语言：发音时间不同
- parts of word
  语素是最小的语义单位
  深度学习:形态学研究少
  Wickelphones (Rumelhart& McClelland 1986)提出了一个如何用英语建立过去式形式的模型，使用字符三元组学习英语动词的过去式。

单词的表达：
有些语言不会在单词之间放置空格，例如中文“
美国关岛国际机场及其办公室均接获”
不少语言的代词、介词、插入语使用都会有些不同，英语或日耳曼语言中也有复合名词

要建立词级的模型，在处理单词时就遇到不少问题：

- 大量的词汇，丰富的表达
- 例如 姓名在翻译时基本是音译，在重写时根据不同的发音写法也会不同。
- 单词的使用不是词典上的规范词（缩写等）

Character-Level Models

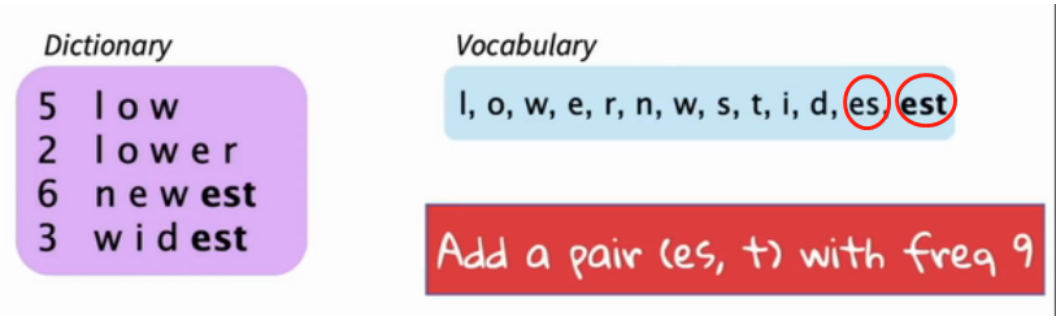
- 词嵌入可以由字符嵌入构成
  - 为未知单词生成嵌入
  - 相似的拼写共享相似的嵌入
  - 解决OOV问题
- 连续的语言作为字符处理，不考虑词级——purely character-Level Models

问题： 经过LSTM后序列变得很长，字符中没有很多信息，必须做反向传播，模型运行很慢。
以捷克语为例：使用word-level model遇到<unk>,需要再进行处理；而character-level model在处理人名类似音译也能处理的很好
近期的研究：

- Jason Lee, KyunghyunCho, Thomas Hoffmann. 2017. 编码使用字符嵌入，使用四个卷积，通过最大池化再经过多层卷积，编码器：char-level GRU。
- Revisiting Character-Based Neural Machine Translation with Capacity and Compression. 2018. Cherry, Foster, Bapna, Firat, Macherey, Google AI. LSTM序列来比较单词和基于字符的模型。使用双向LSTM编码器和单向LSTM解码器

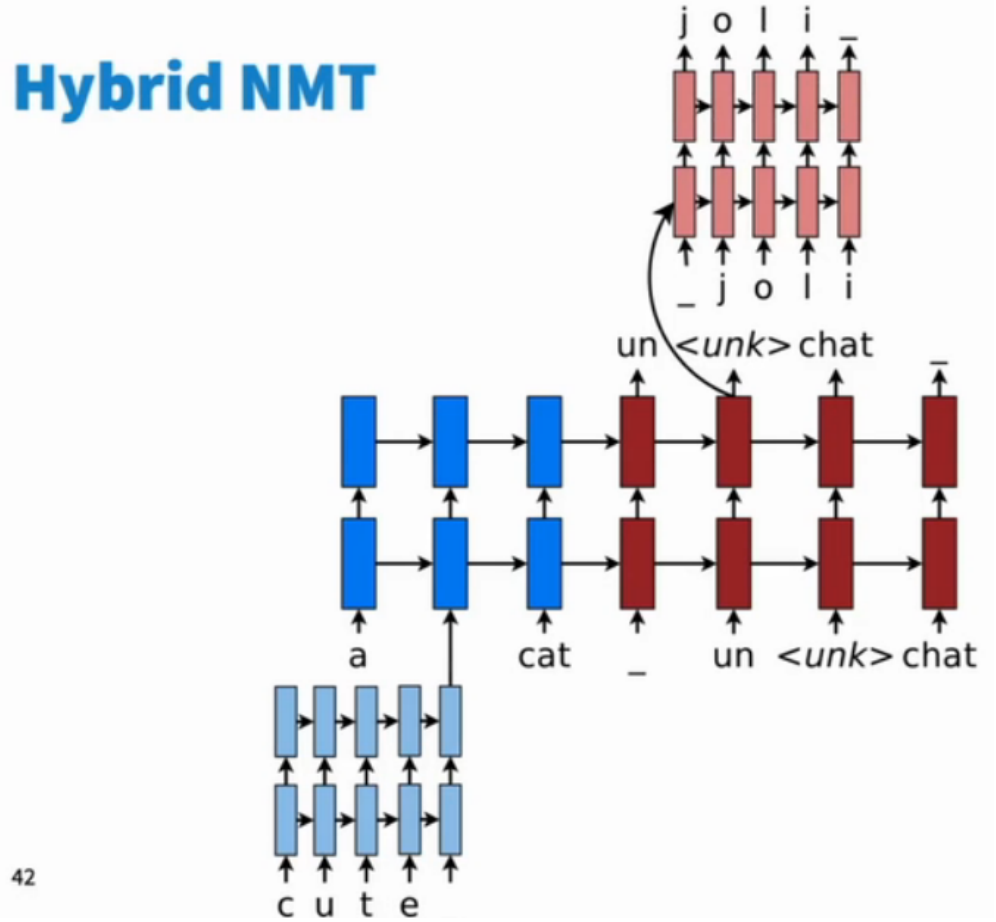
Sub-word models: two trends

1. word piece model。Sennrich, Haddow, Birch, ACL16a, Chung, Cho, Bengio, ACL16。与word model相同的结构，至少是word pieces。构建单独的单词表示
    - **BPE**（byte pair encoding）与DL无关。实际上使用有限的词汇表，但是有无限的词汇量。
        - 压缩算法：将出现自频繁的字节对构成新的字节元素，就能缩短序列长度。将这种思想用到字符和字符n-gram上
        - 分词算法：从unigram词汇开始，将最常见的ngram添加到词汇表。直到词汇量达到要求。



    - GNMT
    - BERT。使用大量的词汇量，未知单词使用word pieces
2. 混合结构。Costa-Jussà& Fonollosa, ACL16, Luong & Manning, ACL16主要是用单词模型，但是通过字符或者更低接别的模型来处理未知单词
    - Learning Character-level Representations for Part-ofSpeech Tagging (Dos Santos and Zadrozny2014)。对字符进行卷积以生成单词嵌入
    - 也可以使用LSTM来代替卷积
    - Yoon Kim, Yacine Jernite, David Sontag, Alexander M. Rush. 2015. 从字符开始，利用一些相关的sub-word和罕见的词.

- Hybrid NMT. 对罕见词使用字符级模型



Chars for word embeddings

- Cao and Rei 2016.使用w2v模型，从字符序列开始运行双向LSTM来计算单词表示
- FastText embeddings. 对于某些n-gram大小，将其表示为一组n-gram。
  where = <wh, whe, her, ere, ex>,<where>
  将word表示为这些表示的和。上下文单词得分为

$$s(w,c) = \sum_{g \in G(w)} \mathbf{z}_g^{\mathrm{T}} \mathbf{v}_c$$