3.1.1 Semantic Similarity: Synonymy and other Semantic Relations

首先介绍一些语义之间的关系。

同义词 synonym：具有相似含义的不同单词 Different words (and also word compounds) can have similar meanings。一些语境下同义词之间可以互换，但不少情况下某一个可能是习惯的用法，因此不能互换，例如 big leagues 是常见用法，large leagues 则不是。

另一个相关的概念是一词多义 polysemy：the property of words to have multiple senses。关于一词多义注意以下几点：

1. 同一单词也可具有不同的词性；

2. 不同含义的使用频率常常是不同的 don't have to be equally frequent；

3. 一些含义之间是有交叉的 some of the senses may overlap，这也是不同字典可能会对列出不同含义列表的原因；

4. 某些单词具有很多含义 some words can be highly polysemous，例如 get 具有 30 多种。

其他一些关系如：

- 反义词 antonymy，例如 raise-lower；

- 上义词 hypernymy，下义词 hyponymy，例如 a deer is a hypernym for elk（麋鹿），an elk is a hypernym for deer。

- 整体-部分关系 meronymy，可分为成员 membership meronymy 和部件两类 part meronymy。前者如 a flock(兽群，鸟群) includes birds，后者如 a table has legs。

应当注意的是，语义关系存在于词的含义之间，而不是词汇之间（Semantic relations hold between word senses, not between words）。例如，hot 的反义词可以是 mild（温柔的。不浓的）当 hot 取含义 spicy（辛辣的）时，也可以是 cold 当 hot 取含义 warm（热的）时。因此, A word may have different antonyms depending on the sense of it。

鉴于此，我们使用术语 synset（同义词集合），它代表一个单词的同义词集合（group together all synonyms of the same word）。具有多义性的单词将对应多个 synset。

3.1.2 Wordnet

这是一个 George 和 Christiane 推进的项目，它主要针对英语，包括词汇（主要是名词和动词，也有形容词和副词）的数据库（database）以及词汇之间的关系。

Wordnet 主要描述上义关系（hypernymy），因此总体而言它是一种森林结构。例如，bar 具有多种含义，从 bar 向上可以有 bar->room->area->...->entity，也可以有 bar->counter->table->furniture（家具）->...->entity，等等。

此外，Wordnet 还可以描述词的其他性质（property），主要如常用性（familiarity）和多义性（polysemy，这里其实是记录词义的个数）。例如从 Wordnet 可以得到 board used as a noun is familiar (polysemy count = 9)，serendipity used as a noun is very rare (polysemy count = 1)。常用性可分为以下几个程度：familiar, common, uncommon, rare, very, rare。

最后介绍了一些其他描述词汇及其之间关系的网络。

3.2 Thesaurus-based Word Similarity Methods

本节介绍基于词典的词相似性计算方法。这种方法都是以某种包含词义关系的词典（如上节的 Wordnet）为基础，定义词汇之间相似性的度量。这里介绍了四种定义方式：

1. Sim (v,w) = - pathlength (v,w)，pathlength 计算从 v 到 w 路径的长度；
2. Sim (v,w) = - log pathlength (v,w)；
3. Sim (v,w) = - log P(LCS(v,w))，这里 LCS（lowest common subsumer）获得能够包含 v 和 w 最小词层，P 计算该词层出现的概率，显然词层越高包含范围越广，概率越大；
4. Sim (v,w) = 2 log P(LCS(v,w)) / (log P(v) + log P(w))。

上述度量 3 和 4 做出的改进，主要是考虑到基于上义（hypernym）关系建立的词典其路径长短不总能代表相似性。当然，基于上义关系的词典总会有其局限性，无论怎样定义度量这种局限总会体现。

此外，这种基于词典的方式最大的缺陷在于某些语言根本没有相应的词典，以及某些新兴或特定领域的词汇不在词典当中。显然，建立和维护这种词典是困难的。

3.3 The Vector Space Model

向量空间模型的关键在于将单词、句子、段落或文章表示为向量。构造向量的方式不是唯一的，一种经典的方式是利用上下文信息构造词的向量，可以度量分布相似性（Distributional Similarity）。其基本的出发点是出现在相似上下文的词汇可能在含义上是相关联的（Two words that appear in similar contexts are likely to be semantically related）。正如 J.R. Firth 的名言“You will know a word by the company that it keeps”。

词所处的上下文（context）的含义可以十分广泛，至少包括以下方面：

1. 前一单词 The word before the target word；
2. 后一单词 The word after the target word；
3. 前后 n 个单词 Any word within n words of the target word；
4. 与目标词在句法上相关的词 Any word within a specific syntactic relationship with the target word (e.g., the head of the dependency or the subject of the sentence)；
5. 同一句中的词 Any word within the same sentence；
6. 同一文档中的词 Any word within the same document。

3.4 Dimensionality Reduction

之前给出的词向量用于计算相似性存在问题，包括相似性小于向量夹角（如一词多义导致的）、相似性大于向量夹角（如同义词之间）。有效的降维有助于发现隐藏的相似性（hidden similarity），而且可以有效解决向量过于稀疏的问题。

对于方阵 A 而言，可以展开为 UBU^{-1}，其中 U 为 A 的特征向量矩阵，B 为特征值构成的对角阵。对于非方阵，利用 SVD（Singular Value Decomposition），可以分解为 $A=UBV^T$，其中 U 是矩阵 AA^T 的特征向量，B 是 A^TA 的特征值构成的对角阵（注意不是方阵了），V 是矩阵 A^TA 的特征向量。

利用 UBV^T 可以完全重构原矩阵 A，但通过舍弃较小的特征值获得 B*，再 UB^*V^T 可以近似重构 A。尤其是，UB*实现了对 A 的行向量的降维，B^*V^T 实现了对 A 的列向量的降维。

为什么会这样呢？以 A 是文档-关键词矩阵（a document to term matrix，列是文档、行是关键词）为例，可以看到 AA^T 其实是描述每对关键词之间相似度的矩阵，A^TA 是文档相似

度矩阵。我们通过保留这两个矩阵较大的特征值保留了方差较大的维度，进而分别对文档向量和关键词向量实现了降维。

这一过程也称为 LSI（latent semantic indexing 或 latent semantic analysis，隐含语义分析）。

3.5.1 NLP Tasks 1

本节介绍各种 NLP 任务。

1. 词性标注（Part of Speech Tagging），需要应对各种歧义；

2. 语法分析（Parsing, produce a syntactic representation for a sentence），包括上下文无关语法和上下文相关语法。

前者常使用 a constituent structure, often a phrase structure grammar，包括两部分，即句子成分之间的规则（也可以说 rules between nonterminals），和终结符号的规则（rules of lexicon, terminals or words）。常常表示为树结构，这个和形式语言课程里面学的是一回事。这儿举了一个 Garden path sentences 的例子，来说明相似句子形式对应不同语法结构，进而含义差别很大，Garden path sentences 满足下面三个标准：

- The part before // should be a complete sentence
- The full sentence has a different meaning than the part before //
- The part before // should not already be ambiguous

后者常使用 a dependency grammar，它最关注的不是句子成分或短语，而是词之间的关系（the relationships between the words in the sentence, without any explicit constituent structure）。它也常常表示为树结构，其中子节点是父节点的参数（arguments），或者子节点修饰（modify）父节点。除了树结构，也可以用一系列二元组来描述语法分析结果，其中每个二元组记录了一对父子节点，显然根节点只会在所有的二元组的父节点位置出现一次，以此为基础就可以生成整个树结构。

3.5.2 NLP Tasks 2

1. 信息提取（Information Extraction, reading a sentence and named entities and relationships between that），所关注的信息可以是实体、事件、规律等，因具体应用而异；

2. 语义分析（semantic analysis），例如推理过程（类似人工智能课程里面学习的谓词逻辑）中的一阶逻辑∀x,y: Mother(x,y) ⇒ Parent (x,y)；

3. 阅读理解（Reading Comprehension），理解段落或文章并回答相关问题；

4. 词汇去歧义(Word Sense Disambiguation, take a word in the sentence, look at its context, and determine which of the senses in WordNet or in a dictionary was meant)，作为基础环节对其他 NLP 任务是很重要的（例如机器翻译 machine translation）；

5. 命名实体识别（Named Entity Recognition），定位出命名实体并标注实体类别，如人名、人名缩写、地点等；

6. 语义角色标记（Semantic Role Labeling），常见以动词为出发点，动词会有各类的参数，其中最重要的是执行者和接受者；

7. 指代消解（或共指消解，Coreference Resolution, understand when two phrases are meant to refer to the same person or entity），可以分为两种形式，即指前照应（实体在前代指在后，Anaphoric, the mention of the entity happens first and then another expression is used to refer back to an entity that has been introduced before)，和指后照应(Cataphor, the pronoun or the reference is introduced first, before the entity itself is introduced)；

8. 省略(ellipsis, a certain word is missing from a sentence because it's implied and it can be understood from the context)，如利用平行结构（parallelism）。

3.5.2 NLP Tasks 3

1. 问答（Question Answering），例如知识问答系统；

2. 情感分析（Sentiment Analysis, to recognize the object that is being discussed and understand the sentiment, whether it has positive or negative in it.）Some of the challenges here are that some of the comments may not be about the object itself but they can be about some property of the object.

3. 机器翻译（Machine Translation），Generate sentences satisfy two criteria. They have to be both grammatical in English and also faithful to the foreign sentence. Here faithful means that the words in English sentences have to be somehow related to the words in the foreign translation.

4. 文本概括（Text Summarization），two different forms. Have a single document and want to produce a short version of that document. Or multi-document summarization, have a series of connected documents, then summary should contain all the information that appears in all of them as consensus, and focus on the differences between the input document.

5. 文本转语音（Text to Speech），例如将文本生成不同性别、年龄、感情色彩的语音；

6. 对话系统（Dialogue Systems）。

7. 等等。