Text Processing 第三章

Notebook:	Text Processing		
Created:	1/28/2020 9:24 AM	**Updated:**	1/28/2020 6:17 PM
Author:	Alvin		

第三部分 NLG

1.NLG定义：

从非文本，非语义的输入，得到文本和语义的输出。
类似输入比如，数字，RDF triples(一种数据格式)
输出文件，报告等，都是文本数据。
需要：语言知识，领域知识

- Input: numerical weather predictions
  - From supercomputer running a numerical weather simulation
- Output: textual weather forecast
  - Users prefer some NLG texts over human texts!
  - More consistent, better word choice

一个简单例子，通过超算预测出的天气预报数据，来生成人类语言的文本。

2.NLG过程：

三个步骤

Document Planning： decide on contnent and strutuce of text

1.Content Selection: 这个步骤中要决定的是文本中什么是重要的，什么可以构成很好的语言叙述，什么更容易表述。

Content Selection的三种方法：

1.Theoretical approach: 基于用户、任务、上下文的深层知识的深层推理

- 减少用户需要了解的信息。
- 依赖比较深度的信息储备，比如用户，领域，世界
- 使用AI reasoning engine，就是在信息中用逻辑规则来减少新的信息。
- 实际中不实用，现实中对用户和内容的知识是很缺乏的。

2.Pragmatic approach：编写能够模仿人类语言程序schemas

- 分析文献文本，人工标定内容和结构规则
- 通常基于模仿人类书写的文本中的模版
- 明确规定结构和内容

Creating Schemas:

一般用Java或者其他编程语言写，并没有严谨的方法论支撑。

- 文本可能有时候并不完全一致，并非出自一人之手
- 其次文献并不包含所有的情况。

3.Statistical approach：使用学习技术来学习从文献中学习行文规则

- 统计学习技术（包括深度学习），使用机器学习来学习如何选择内容/规则/程序
- 如果数据量很大的话，可以考虑使用。

2.Structure: 叙述的顺序，措辞的结构。

- CONCESSION (although, despite)

- CONTRAST (but, however)
- ELABORATION (usually no cue)
- EXAMPLE (for example, for instance)
- REASON (because, since)
- SEQUENCE (and, also)

总结：

- Content determination是NLG的第一部也是最重要的一步。
- 大多基于模仿人类书写的文本
- 同样决定结构，比如：Tree structure, rhetorical relations

Microplanning: decide how to linguistically express text

Microplanning是NLG的第二个步骤。

1.Lexical/syntactic choice: 使用什么样的语句，语言结构来表达信息内容。

影响Lexical choice的问题：

- Frequency (affects readability)
  - lie vs prevarication
- Formality:
  - Error vs howler
- Focus, expectations
  - not many, few, a few, only a few [students failed the exam]
- Technical terms
  - (statistics) standard error, not
  - standard mistake
- Convention
  - Temperature falls, Wind speed eases

Statistics-Based Lexical Choice for NLG from quantitative information:

- 我们的目标是建立一个统计算法，检测数据维度和单词的关系。
- 不依赖人工的规则。
- 同时预测什么样的词在什么时候应该被使用。
- 一个词可以涉及多个维度。

例子：

$$P(\text{“muggy”} \mid ws{=}20, temp{=}35, humid{=}97, \ldots)$$

具体实例参考一下NLG第三个PDF里面的例子，了解一下实现过程，个人觉得应该考试不会让手算这东西（因为他PPT里都没写细写），但是了解一下过程还是可以。

2.Aggregation：有用信息如何在句子和段落中分布。

把不同的独立的表达合成一个单独，或者更简洁的表达形式。

建议：越简洁越好，如何合成取决于词句之间有多(similar)相似，同时取决于文献的(genre)类型。

3.Reference：文本信息如何和具体描述对象和事件对应。

Reference的类别：

- Pronoun – it, them, him, you,...
- Name – Dr Adam Smith, Adam Smith, Adam, Dr Smith
- Definite NP – the big black dog, the big dog, the black dog, the dog

建议：使用名词，名称，definite noun phrase（有定名词词组），且只使用文献中出现的形式。

Realisation:

NLG的第三个步骤

从有结构的input里生成线性的文本，保证(syntactic)句法的正确性。

- Grammar: 语言的语法和不同的行文手法，比如媒体播报和论文写作
- Structure：HTML,RTF,或者输出其他需要的文件格式。

Problem：语言的部分细节实在太多以至于开发NLG系统的人不想去关心，于是用realiser自动处理这些。

Syntax:

- 语句必须服从英语语法
- 语法的很多方面很奇怪。
- 只要告诉realiser，动词，时态，是与否，它会自动找到对应的动词组。
- 同样自动化其他比较模糊(obscure)的信息 encodings of information

Morphology:

- Variations of a root form of a word, e.g., prefixes, suffixes
- Inflectional morphology - same core meaning
  - plurals, past tense, superlatives, e.g., dog, dogs
  - part of speech unchanged
- Derivational morphology - change meaning
  - prefix *re* means do again: reheat, resit
  - suffix *er* means one who: teacher, baker
  - part of speech changed

Realiser:

- 自动计算词汇的不同形态
- 自动插入合适的标点(punctutaion)来形成结构

- 很多输出文件格式：TXT,HTML和MS WORD。

后面介绍了一些NLG systems，比如simpleNLG,KPML,openCCG

总结：用Realiser来自动处理一些过于吹毛求疵(finicky)的语言语法细节，是NLG的一大优点

Advanced: User-Adaptation：

- Texts should depend on
  - User's personality
  - User's domain knowledge (how much do we need to explain)
  - User's vocabulary (can we use technical terms in the text)
  - User's task (what does he need to know)

能完全获得这些信息不现实。

这一部分还有Possible Content Rules，Possible Ordering Rules，感觉不是像是重点，NLG第二个PDF的17，18页里有，就几句话。

3.构建NLG系统：

两大组成部分：

Knowledge and corpus analysis：

- Which patterns most important?
- What order to use?
- Which words to use?
- When to merge phrases?
- Etc.

knowledge 来源：1.模仿人类文献。2.咨询领域专家。3，用户实验

Evaluation：

1.系统是否能够帮助人类。2.人们是否喜欢文本。3.和人类文献对比

4.NLG对比NLP：

- 产生而不是理解语言
- 关注数据内容本身，AI技术和语言学的技术。
- 统计和深度学习的技术的使用越来越多。