应用：语言自动生成

一、自然语言生成概述

自然语言生成定义

自然语言生成历史

自然语言生成的架构和方法

自然语言生成评价方法

自然语言生成应用举例

NLG(NATURAL LANGUAGE GENERATION)

维基百科定义

- **Natural language generation** (**NLG**) is the natural language processing task of generating natural language from a machine representation system such as a knowledge base or a logical form

Reiter & Dale的定义

- NLG is characterized as ‘the sub- field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying non-linguistic repre- sentation of information’ (Reiter & Dale, 1997)

NLG划分

按照不同的输入划分

- 文本到文本的生成(text-to-text generation)
- 意义到文本的生成(meaning-to-text generation)
- 数据到文本的生成(data-to-text generation)
- 图像到文本的生成(image-to-text generation)

NLG发展历史

模板生成技术

模式生成技术

短语规则扩展技术

属性特征生成技术

模板生成技术

最早采用的自然语言生成技术

设计可能出现的语言情况，构造相应模板(包括常量、变量)

根据用户输入信息替代模板中变量，生成文本

优点：效率高，实现手段简单

缺点：处理仅在字符级上处理，生成文本质量不高，难以满足多变的需求

模式生成技术

模式生成技术(Schema based generation)是基于语义学中的修辞谓词来表达文本结构

- 文本表示成结构树形式(Root, Schema, Predicate, Argument, Modifier)
  - Root是根节点，表示一篇文章
  - Schema是子节点，表示一段话或几句话
  - Predicate是一个子树，表示一个句子(文章的基本单位)
  - Argument是叶子节点，表示句子中的基本语义成分
  - Modifier是叶子节点，代表具有对修饰成分Argument

具有较好的维护性，生成的文本质量高，但只能用于固定结构段落的生成

短语规则扩展技术

基于结构修辞理论(RST)，文章的各个组成部分都是由一些特定的关系按照一定的层次内聚在一起

包含两种模式：nucleus-satellite和multi-nucleus

- nucleus-satellite包括表达基本命题和表达附属命题，其组合表达目的、因果、转折、背景等关系
- multi-nucleus涉及一个或多个语段，用于说明顺序、并列等关系

具有更强的灵活性，在生成文本时也生成文本的总体框架结构，缺点在与基本数据结构、文本规则库较难建立

属性特征生成技术

在生成系统中，每个变化都是由一个属性特性表示出来

- 生成文本是主动or被动
- 生成的文本表示的动作是问题or命令

输出单元与特定的属性特征集相连，在生成过程中对每个信息增加对应的属性特征，确定输出结果

- 特征属性一般是语法特征
- 输出单元是词汇

优点在于概念简单，生成文本灵活，但难以维护各个属性间的内容关系，难以控制特征集选择

NLG 框架和方法

- Modular Approaches
- Planning-based Approaches
  - Planning through the Grammar
  - Stochastic Planning under Uncertainty using Reinforcement Learning
- Other Stochastic Approaches to NLG
  - NLG as a Sequential, Stochastic process
  - NLG as Classification and Optimisation
  - NLG as ‘Parsing’
  - Deep Learning Methods

MODULAR APPROACHES

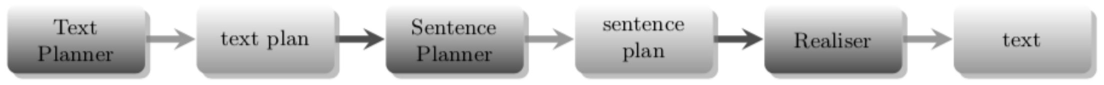


Figure 3: Classical three-stage NLG architecture, after Reiter and Dale (2000). Darker segments illustrate the three main modules; lighter segments show the outputs.

Pipeline architectures

1. Text Planner: content selection and text structuring (what to say)
2. Sentence Planner: sentence aggregation, lexicalisation and referring expression generation (how to say)
3. Linguistic Realiser: syntactic and morphological rules (say it)

PLANNING-BASED APPROACHES

- Planning through the Grammar
  - “Mary likes the white rabbit”为例
  - 词条”likes”的表示如右侧所示
  - Like作为动词， x likes y的句子需要填充两个名词
  - 场景中有一个人和两只兔子(一只是白色的)，推导过程先排除人(实体x是听众的一部分)将兔子放到y的位置，此时有两只兔子，“白色”就作为限定词
- Stochastic Planning under Uncertainty using Reinforcement Learning
  - 生成过程建模为马尔科夫决策过程

(12) **likes**(u, x, y) action:
PRECONDITIONS:

- The proposition that x likes y is part of the knowledge base (i.e. the statement is supported);
- x is animate;
- The current utterance u can be substituted into the derivation S under construction;

EFFECTS:

- u is now part of S
- New NP nodes for x in agent position and y in patient position have been set up (and need to be filled).

NLG AS A SEQUENTIAL, STOCHASTIC PROCESS

generative model based on a sequential, Markov process, combining strategic choices (of db records and fields) with tactical choices (of word sequences) into a single probabilistic model

- 排序树结构
- 根代表dialogue act type
- 叶子代表word或word sequence

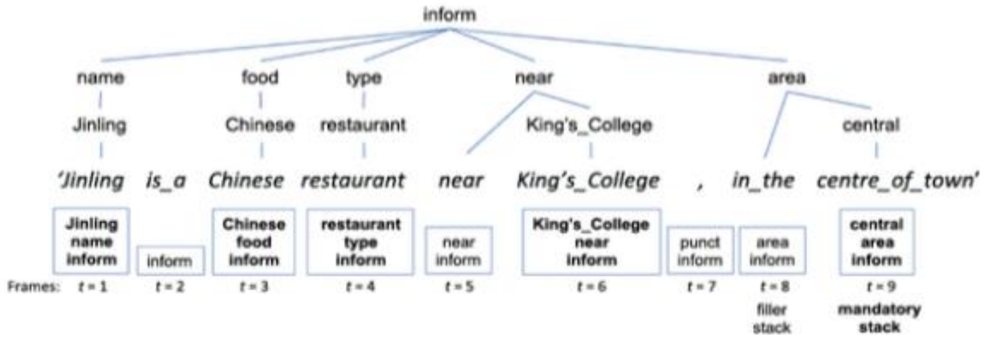


Figure 5: Tree structure for a dialogue act, after Mairesse and Young (2014). Leaves correspond to word sequences. Non-terminal nodes are semantic attributes, shown at the bottom as semantic stacks. Stacks in bold represent mandatory content.

NLG AS CLASSIFICATION AND OPTIMISATION

Classification

使用一系列分类器来执行参考表达式的生成

基于SVM的排序模型，从文本中提取出依赖表示

1. 将输入映射到浅层语法树以进行线性化
2. 插入参考表达式

(14) *Junge* *Familie*$_{v:0}$ *auf* *dem Heimweg*$_{poss:v}$ *ausgeraubt*$_{ag:p}$
Young family on the way home robbed
'A young family was robbed on their way home.'

Optimisation

1. Each NLG task is once again modelled as classification or label-assignment, but this time, labels are modelled as binary choices (either a label is assigned or not), associated with a cost function, defined in terms of the probability of a label in the training data;

2. Pairs of tasks which are strongly inter-dependent (for example, syntactic choices and REG realisations, in the example from Zarrieß & Kuhn, 2013) have a cost based on the joint probability of their labels;

3. An ILP model seeks the global labelling solution that minimises the overall cost, with the added constraint that if one of a pair of correlated labels $\langle l_i, l_j \rangle$ is selected, the other must be too.

NLG AS 'PARSING'

As probabilistic context-free grammer formalism(CFG) or 'inverse' of semantic parsing

例如：天气记录的数据集，规则如下

R-数据库中的一条记录，FS域的集合，F(x,y)代表x记录中的y域，W代表word sequence

- (15)描述风速应遵循温度和雨量记录
- (16)描述最小风速应遵循确定概率下的最大风速
- (17)双元语言模型将风速规则扩展为到word sequence

$$(15)\quad R(windSpeed) \rightarrow FS(temperature), R(rain)$$

$$(16)\quad FS(windSpeed, min) \rightarrow F(windSpeed, max)FS(windSpeed, max)$$

$$(17)\quad FS(windSpeed, min) \rightarrow W(windSpeed, min)$$

DEEP LEARNING METHOD

Encoder-Decoder Architecture

- RNN将输入编码为向量表示，作为解码器RNN的辅助输入

Conditioned Language Models

- 将生成器视作条件语言模型，输出是根据输入特征而定的分布中采样单词或字符来生成的，输入特征可以包括语义、上下文或文体属性

NLG 评价方法

NLG系统难于评价的原因：

- Variable input
  - NLG系统输入格式多样
  - 即使使用标准数据集，由于input variation 或implicit biases in the input data也不能直接进行对比
- Multiple possible outputs
  - 即使是单个系统的单个输入，可能的输出范围也是开放的，特别在输出是文本的NLG任务

NLG 评价方法

- 内在评价(Intrinsic Methods)——measures the performance of a system
  - 如：文本的质量、输出的可读性和正确性相关的问题
  - Method: human judgements, rely on corpora
- 外在评价(Extrinsic Evaluation)—— measure effectiveness in achieving a desired goal
  - 如：系统是否真正实现在offshore platform上的目标

SUBJECTIVE (HUMAN) JUDGEMENTS

评价输出的主观标准：

1. Fluency or readability the linguistic quality of the text
2. Accuracy, adequacy, relevance or correctness relative to the input, reflecting the system’s rendition of the content

OBJECTIVE MEASURES USING CORPORA

依赖于语料库的评价可以说是解决 "humanlikeness"的问题，即在可比条件下系统的输出与人类输出匹配的程度

1. N-gram overlap
2. String distance
3. Content overlap

	Metric	Description	Origins
N-gram overlap	BLEU	Precision score over variable-length n-grams, with a length penalty (Papineni et al., 2002) and, optionally, smoothing (Lin & Och, 2004).	MT
	NIST	A version of BLEU with higher weighting for less frequent n−grams and a different length penalty (Doddington, 2002).	MT
	ROUGE	Recall-oriented score, with options for comparing non-contiguous n−grams and longest common subsequences (Lin & Hovy, 2003).	AS
	METEOR	Harmonic mean of unigram precision and recall, with options for handling (near-synonymy) and stemming (Lavie & Agarwal, 2007).	MT
	GTM	General Text Matcher. F-Score based on precision and recall, with greater weight for contiguous matching spans (Turian et al., 2003)	MT
	CIDEr	Cosine-based n-gram similarity score, with n-gram weighting using TF-IDF (Vedantam et al., 2015).	IC
	WMD	Word-Mover Distance, a similarity score between texts, based on the (semantic) distance between words in the texts (Kusner et al., 2015). For NLP, distance is operationalised using normalised bag of words (NBOW) representations (Mikolov et al., 2013).	DS; IC

mt = Machine translation; as = automatic summarisation; ic = image captioning; ds = document similarity.

OBJECTIVE MEASURES USING CORPORA

String distance	Edit distance	Number of insertions, deletions, substitutions and, possibly, transposition required to transform the candidate into the reference string (Levenshtein, 1966).	N/A
	TER	Translation edit rate, a version of edit distance (Snover et al., 2006).	MT
	TERP	Version of TER handling phrasal substitution, stemming and synonymy (Snover et al., 2006).	MT
	TERPA	Version of TER optimised for correlations with adequacy judgements (Snover et al., 2006).	MT
Content overlap	Dice/Jaccard	Set-theoretic measures of overlap between two unordered sets (e.g. of predicates or other content units)	N/A
	MASI	Measure of agreement between set-valued items, a weighted version of Jaccard (Passonneau, 2006)	AS
	PYRAMID	Overlap measure relying on comparison of weighted Summarization Content Units (SCUs) (Nenkova & Passonneau, 2004; Yang et al., 2016)	AS
	SPICE	Measure of overlap between candidate and reference texts based on propositional content obtained by parsing the text into graphs representing objects and relations, by first parsing captions into scene graphs representing objects and relations (Anderson et al., 2016)	IC

EXTRINSIC EVALUATION

评价系统达到目标的有效性，而有效性取决于系统的应用领域和目的，如：

- persuasion and behaviour change, for example, through exposure to per- sonalised smoking cessation letters (Reiter et al., 2003)
- purchasingdecisionafterpresentationofargumentsforandagainstoptions on the housing market based on a user model (Carenini & Moore, 2006)
- engagement with ecological issues after reading blogs about migrating birds (Siddharthan et al., 2013)
- decision support in a medical setting following the generation of patient reports (Portet et al., 2009; Hunter et al., 2012)
- enhancing linguistic interaction among users with complex communication needs via the generation of personal narratives (Tintarev et al., 2016)
- enhancing learning efficacy in tutorial dialogue (Di Eugenio et al., 2005; Fossati et al., 2015; Boyer et al., 2011; Lipschultz et al., 2011; Chi et al., 2014)

NLG应用示例

论文写作

摘要生成

自动作诗

新闻写作

报告生成

百科写作

……

论文自动生成

SCIgen- An Automatic CS Paper Generator

只要输入作者名，就可以生成"SCI级别"的computer science论文

SCIgen多次成功逆袭IEEE的国际会议

- 在2005年，机器论文Rooter: A Methodology for the Typical Unification of Access Points and Redundancy被WMSCI会议所接收
- 2008年和2009年中国武汉举办的两个IEEE国际会议投稿，还获得高度评价

摘要生成

新闻摘要应用 Summly

- 提取新闻摘要功能，可以通过书签工具把文章发送到 Summly
- 2013年雅虎收购新闻摘要应用 Summly

微软“万小冰”提供金融领域公告摘要服务

- 实时抓取沪深两市公告作为基础数据
- 处理文件中表格、文字、数据等
- 对公告内容进行分类和结构化处理
- 建立自动摘要生成模型
- 在公告发布后**20秒**左右生成高质量摘要

科大国创:2018年半年度业绩预告

公告日期: 2018-07-14

发布时间: 2018-07-13 19:32

附件: 科大国创：2018年半年度业绩预告.pdf

公告摘要:

科大国创2018年半年度业绩预告,预计归属于上市公司股东的净利润，比上年同期增长25%—55%、盈利581.14万元—720.61万元。主要原因是 因公司业务存在明显的季节性波动，通常来说，公司营业收入及相应销售回款上半年较少，但期间的相关费用并没有减少，从而导致公司净利润的季节性波动明显。报告期内，公司整体经营平稳。 公司预计报告期内非经常性损益对归属于上市公司股东净利润的影响金额约为300万元左右。(摘要来自万小冰)

自动作诗

清华大学作诗机器人“薇薇”通过“图灵测试”

- 薇薇”创作的诗词中，有31%被认为是人创作
- 薇薇”通过社科院等唐诗专家评定，通过“图灵测试”

聊天机器人小冰创作的诗集《阳光失了玻璃窗》出版

新闻写作

Automated Insights公司开发的wordsmifh平台可以在每秒之内生成近2000篇新闻稿件

2015年，腾讯通过机器人新闻写作可以在政府发布CPI资料之后，只用了几分钟的时间就完成了相关新闻稿件的发布

今日头条和南方都市报先后与北大计算机所合作，分别推出奥运 AI 小记者 Xiaomingbot 和“小南”写稿机器人

报告生成

「心声医疗」帮医生自动生成诊断报告

- 通过CNN、RNN等AI技术分析心电图，生成诊断报告

CMU 邢波教授利用 AI 自动生成医学影像报告

- 运用图像说明技术(CNN-RNN框架)，可以为胸部X射线影像添加文本标签,自动生成文字描述

百科生成

2009年提出结构化方法生成维基百科文章

- 采用产生相关主题的模板，根据主题选择内容构成百科

谷歌大脑提出通过多文档摘要方法生成维基百科

- 输入维基百科的主题和参考文献的集合，目标是生成维基百科文章的文本
- 看作多文档摘要任务，并应用了Transformer 结构

小结

自然语言生成概述

- 定义及划分的四种类型
- 发展历史阶段
- 框架和方法
- 评价方法
- 应用场景及示例

二、数据到文本的生成技术

数据到文本生成的定义

数据到文本生成系统框架

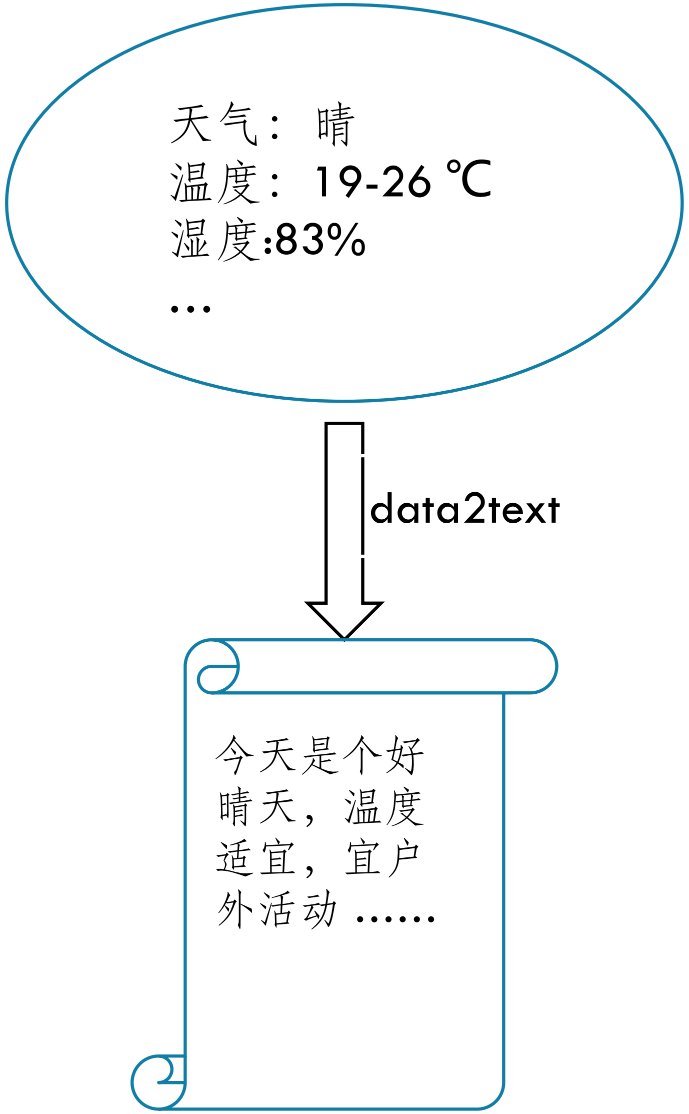
数据到文本生成的应用

数据到文本生成的研究前沿

小结

定义

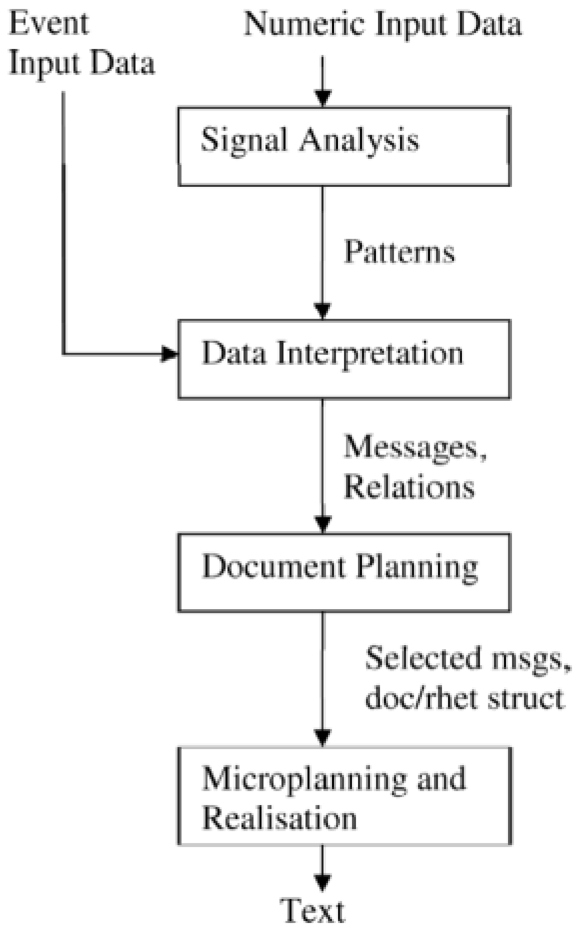
数据到文本的生成技术指根据给定的数值数据生成相关文本，例如基于数值数据生成天气预报文本、体育新闻、财经报道、医疗报告等



数据到文本生成系统框架

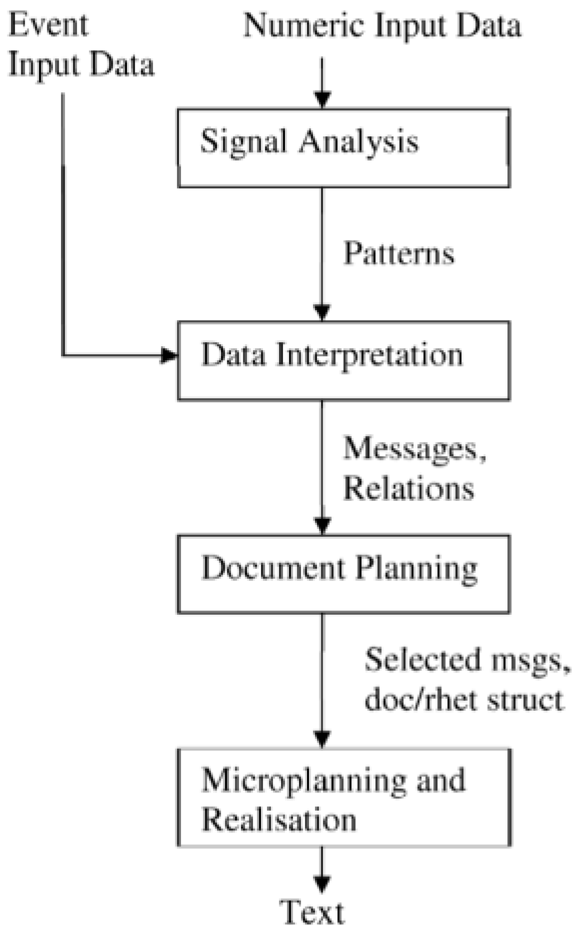
英国阿伯丁大学的 Ehud Reiter 在三阶段流水线模型：

- 信号分析模块(Siganl Analysis)
- 数据阐释模块(Data Interpretation)
- 文档规划模块(Document Planning)
- 微规划与实现模块(Microplanning and Realisation)



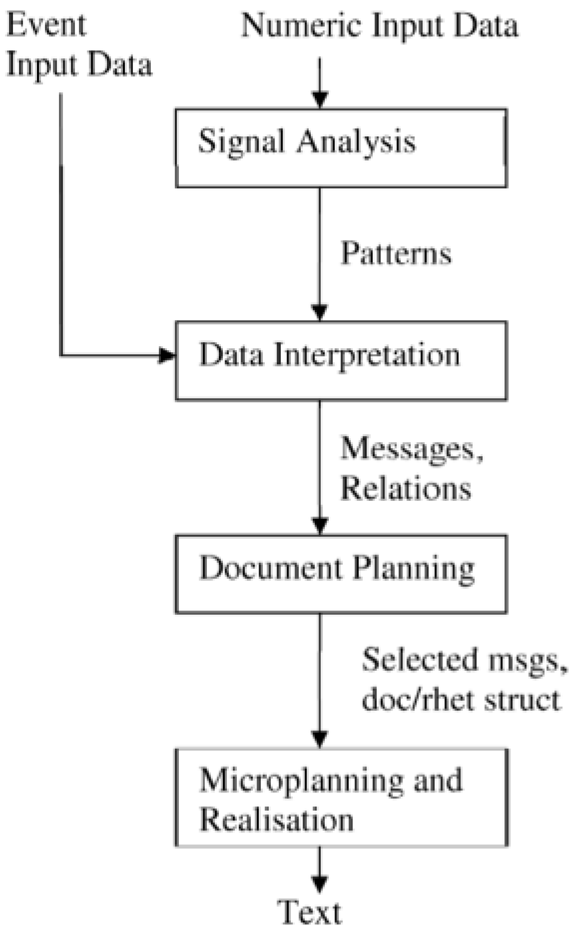
数据到文本生成系统框架

- 信号分析模块(Siganl Analysis)
  - 输入为数值数据，通过利用各种数据分析方法检测数据中的基本模式，输出离散数据模式
- 该模块与具体应用领域和数据类型相关，针对不同的应用领域与数据类型所输出的数据模式是不同的



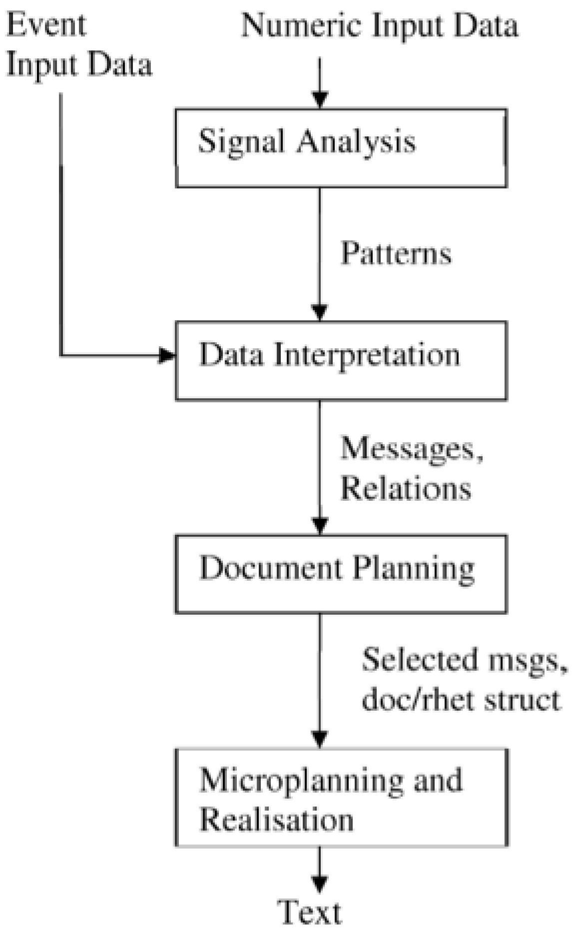
数据到文本生成系统框架

- 数据阐释模块(Data Interpretation)
  - 输入为基本模式与事件，通过对基本模式和输入事件进行分析，推断出更加复杂和抽象的消息，同时推断出它们之间的关系，最后输出高层消息以及消息之间的关系
  - 例针对股票数据，如果跌幅超过某个值则可以创建一条消息，还需要检测消息之间的关系，例如因果关系、时序关系等



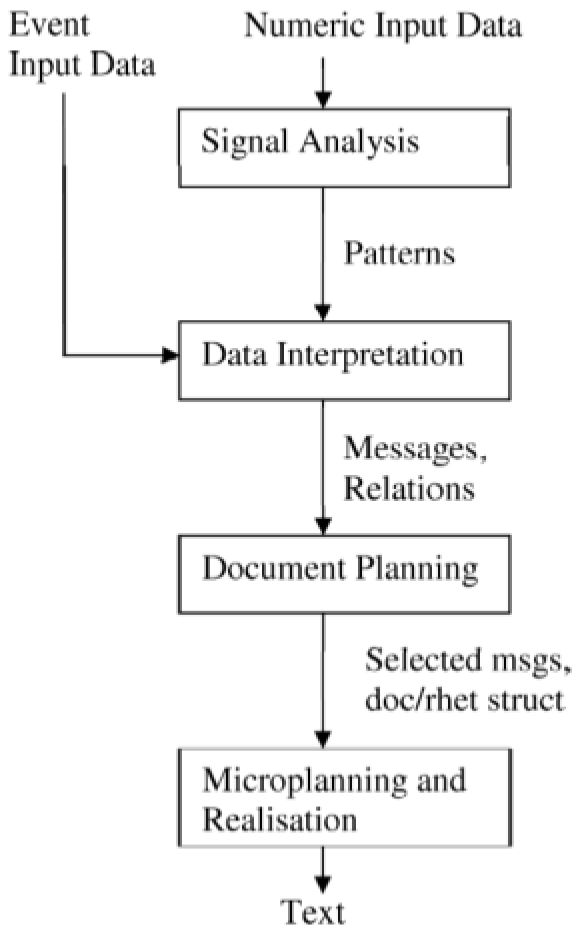
数据到文本生成系统框架

- 文档规划模块(Document Planning)
  - 输入为消息及关系，分析决定哪些消息和关系需要在文本中提及，同时要确定文本的结构，最后输出需要提及的消息以及文档结构
  - 信号分析与数据阐释模块会产生大量的消息、模式和事件，但文本通常长度受限，只能描述其中的一部分，因此文档规划模块必须确定文本中需要说明的消息



数据到文本生成系统框架

- 微规划与实现模块(Microplanning and Realisation)
  - 输入为选中的消息及结构，通过自然语言生成技术输出最终的文本
  - 主要涉及到对句子进行规划以及句子实现，要求最终实现的句子具有正确的语法、形态和拼写，同时采用准确的指代表达



数据到文本生成的应用

文本生成系统的应用领域:

- 天气预报领域的文本生成系统
- 针对空气质量的文本生成系统
- 针对财经数据的文本生成系统
- 面向医疗诊断数据的文本生成系统
- 基于体育数据生成文本摘要

数据到文本生成的应用

- 天气预报领域的文本生成技术应用最为成功
- FoG 系统 能够从用户操作过的数据中生成双语天气预报文本
- SumTime 系统能够生成海洋天气预报文本
- 英国阿伯丁大学的 Anja Belz 提出概率生成模型进行天气语言文本的生成
- Anja Belz 和 Eric Kow进一步基于天气预报数据分析对比了多种数据到文本的生成系统

2. FORECAST 6 - 24 GMT, Wed 12-Jun 2002
WIND(KTS)
10M: W 8-13 backing SW by mid afternoon and S 10-15 by midnight.
50M: W 10-15 backing SW by mid afternoon and S 13-18 by midnight.
WAVES(M)
SIG HT:0.5-1.0 mainly SW swell.
MAX HT: 1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon, then rising 1.0-1.5 by midnight.
PER(SEC)
WAVE PERIOD: Wind wave 2-4 mainly 6 second SW swell.
WINDWAVE PERIOD: 2-4.
SWELL PERIOD: 5-7.
WEATHER: Mainly cloudy with light rain showers becoming overcast around midnight.
VIS(NM): Greater than 10.
AIR TEMP(C): 8-10 rising 9-11 around midnight.
CLOUD(OKTAS/FT): 4-6 ST/SC 400-600 lifting 6-8 ST/SC 700-900 around midnight.

Figure 2. Forecast Text Produced by SUMTIMEMOUSAM for the AM of 12-Jun 2002. The Wind part of the forecast has been generated from the data, shown in Table 1

数据到文本生成的应用

针对空气质量的文本生成系统

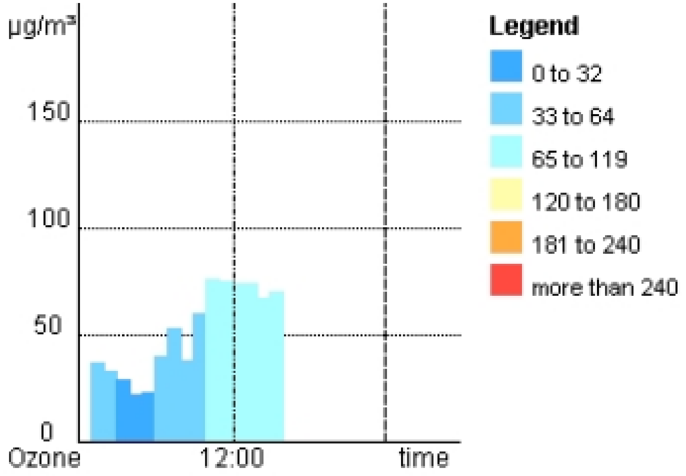


Figure 6: A sample English text

The ozone concentration (70 μg/m³) is relatively low. As a result, no harmful effects to human health are expected. Between 4 AM and 10 AM, the ozone concentration increased considerably from 22 to 76. The current ozone concentration (70 μg/m³) is close to the highest of 76 μg/m³ (at 10 AM). The lowest was 22 μg/m³ (at 4 AM).

数据到文本生成的应用

针对财经数据的文本生成系统

- 腾讯Dreamwriter，对数据进行学习，生成写作手法，进行新闻报道写作
- Narrative Science公司的生成系统可以自动生成新闻

腾讯Dreamwriter写稿机器人

Dreamwriter写稿流程

数据到文本生成的应用

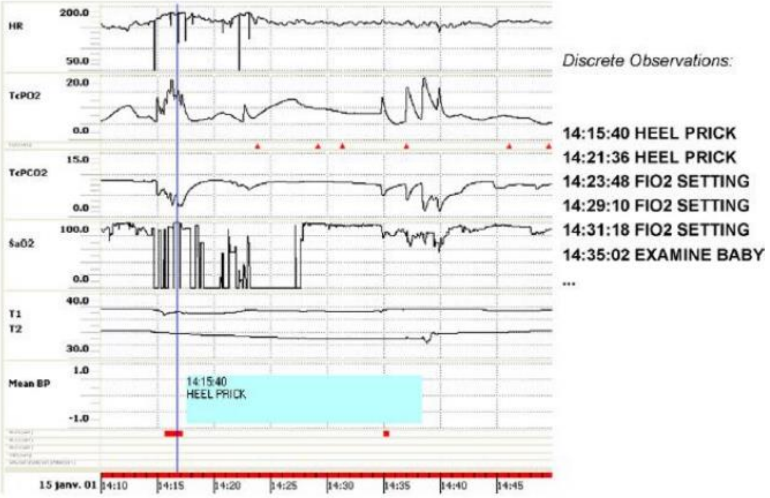


图 4.2: NICU 数据样例，从上到下分别表示 HR, TcPO2, TcPCO2, SaO2, T1 & T2, and Mear

面向医疗诊断数据的文本生成系统

You saw the baby between 14:10 and 14:50. Heart Rate (HR) = 159. Core Temperature (T1) = 37.7. Peripheral Temperature (T2) = 34.3. Transcutaneous Oxygen (TcPO2) = 5.8. Transcutaneous CO2 (TcPCO2) = 8.5. Oxygen Saturation (SaO2) = 89.

Over the next 30 minutes T1 gradually increased to 37.3.

By 14:27 there had been 2 successive desaturations down to 56. As a result, Fraction of Inspired Oxygen (FIO2) was set to 45%. Over the next 20 minutes T2 decreased to 32.9. A heel prick was taken. Previously the spo2 sensor had been re-sited.

At 14:31 FIO2 was lowered to 25%. Previously TcPO2 had decreased to 8.4. Over the next 20 minutes HR decreased to 153.

By 14:40 there had been 2 successive desaturations down to 68. Previously FIO2 had been raised to 32%. TcPO2 decreased to 5.0. T2 had suddenly increased to 33.9. Previously the spo2 sensor had been re-sited. The temperature sensor was re-sited.

图 4.3：BT-45 系统生成的对应文本 [Portet et al., 2009][87]

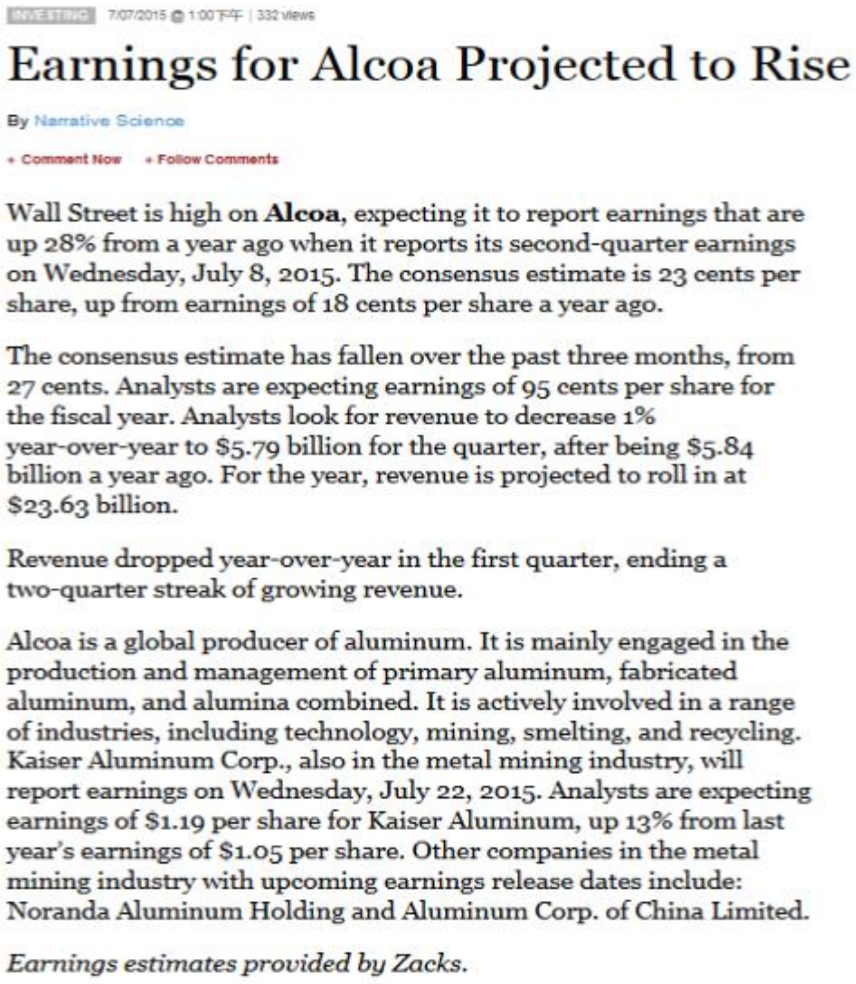
数据到文本生成的应用

工业界成立了多家从事文本生成的公司，能够为多个行业基于行业数据生成行业报告或新闻报道等

ARRIA

narrative science

ai AUTOMATED INSIGHTS


INVESTING 7/07/2015 @ 1:00下午 | 332 views

Earnings for Alcoa Projected to Rise

By Narrative Science

+ Comment Now + Follow Comments

Wall Street is high on **Alcoa**, expecting it to report earnings that are up 28% from a year ago when it reports its second-quarter earnings on Wednesday, July 8, 2015. The consensus estimate is 23 cents per share, up from earnings of 18 cents per share a year ago.

The consensus estimate has fallen over the past three months, from 27 cents. Analysts are expecting earnings of 95 cents per share for the fiscal year. Analysts look for revenue to decrease 1% year-over-year to $5.79 billion for the quarter, after being $5.84 billion a year ago. For the year, revenue is projected to roll in at $23.63 billion.

Revenue dropped year-over-year in the first quarter, ending a two-quarter streak of growing revenue.

Alcoa is a global producer of aluminum. It is mainly engaged in the production and management of primary aluminum, fabricated aluminum, and alumina combined. It is actively involved in a range of industries, including technology, mining, smelting, and recycling. Kaiser Aluminum Corp., also in the metal mining industry, will report earnings on Wednesday, July 22, 2015. Analysts are expecting earnings of $1.19 per share for Kaiser Aluminum, up 13% from last year's earnings of $1.05 per share. Other companies in the metal mining industry with upcoming earnings release dates include: Noranda Aluminum Holding and Aluminum Corp. of China Limited.

Earnings estimates provided by Zacks.

图 4.4：NarrativeScience 自动生成的样例新闻

小结

数据到文本生成的定义

数据到文本生成系统框架

- 三阶段流水线模型

数据到文本生成的应用

- 天气预报、财经、医疗等

三、文本到文本的生成

- 文本到文本的生成定义
- 对联自动生成
  - 对联特点
  - 对联自动生成方法
- 诗歌自动生成
  - 诗歌特点
  - 诗歌自动生成方法
- 总结与展望

定义

定义：对给定文本进行变换和处理从而获得新文本的技术

应用：

- 对联自动生成
- 诗歌自动生成
- 作文自动生成
- 对话生成
- 机器翻译

对联自动生成

对联特点：上联(FS, first sentence)和下联(SS, second sentence)

对联自动生成：给出上联，自动生成下联

海 sea	阔 wide	凭 allow	鱼 fish	跃 jump
天 sky	高 high	任 permit	鸟 bird	飞 fly

微软对联：http://duilian.msra.cn

对联特点

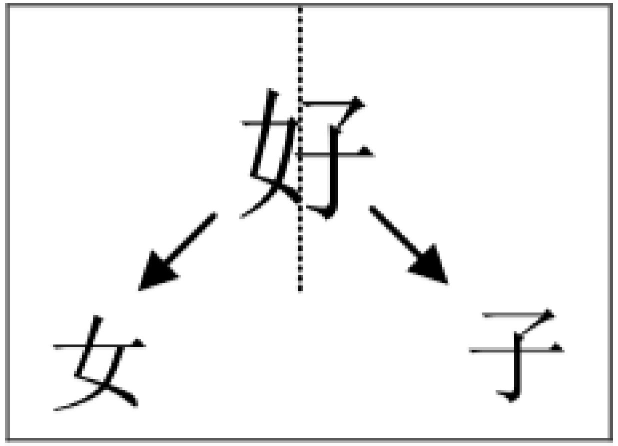
1. 上下联具有相同的长度，对应位置的分词一致
2. 上下联的音调协调。FS的最后一个字是仄音，SS的最后一个字是平音
3. 上下联对应的词具有相同的词性和特点
4. 上下联内容相关，但不重复

海	阔	凭	鱼	跃
sea	wide	allow	fish	jump
天	高	任	鸟	飞
sky	high	permit	bird	fly

对联特点

5. 上下联具有相同的写作风格，如 FS中有重复的字、词或发音，则SS中对应位置具有相同的重复；FS中具有字的分解，SS中对应位置也应该有对应字的分解

有	女	有	子	方	称	好
have	daughter	have	son	so	call	good
缺	鱼	缺	羊	敢	叫	鲜
lack	fish	lack	mutton	dare	call	delicious



对联自动生成

对联自动生成步骤：

1. Phrase-based SMT Model 生成出N-Best 候选
2. 基于语言学的筛选
3. 基于其他特征的重排序

对联自动生成模型

Phrase-based SMT Model

- 上联表示 F = {f1, f2, …, fn}, 下联表示 S ={s1, s2, …, sn}，其中fi, si是对联中对应第i个汉字
- 生成对联，目标最大化 p(S/F)
- 采用phrase-based log-linear model

$$
\begin{aligned}
S^* &= \arg\max_{S} p(S \mid F) \\
&= \arg\max_{S} \sum_{i=1}^{M} \lambda_i \log h_i(S, F)
\end{aligned}
$$

其中 hi(S,F)为特征函数，M为特征函数的个数，λi是估计量

对联自动生成模型

应用于Phrase-based SMT Model的特征：

S, F 切分为短语后分别表示为$\overline{s1},\ldots,\overline{sl}$ 和$\overline{f1},\ldots,\overline{fl}$

1. Phrase translation model
2. Inverted phrase translation model
3. Lexical weight
4. Inverted lexical weight
5. 下联的语言模型得分(character-based trigram)

$h_1(S,F)=\prod_{i=1}^{I} p(\overline{f_i}\mid\overline{s_i})$	Phrase translation model
$h_2(S,F)=\prod_{i=1}^{I} p(\overline{s_i}\mid\overline{f_i})$	Inverted phrase translation model
$h_3(S,F)=\prod_{i=1}^{I} p_w(\overline{f_i}\mid\overline{s_i})$	Lexical weight
$h_4(S,F)=\prod_{i=1}^{I} p_w(\overline{s_i}\mid\overline{f_i})$	Inverted lexical weight
$h_5(S,F)=p(S)$	Language model

基于语言学的筛选

1. Repetition filter
   检查上联和下联对应位置是否有重复的字(词)
2. Pronunciation repetition filter
   检查上联和下联对应位置字的发音
3. Character decomposition filter
   检查上联和下联字的分解是否对应
4. Phonetic harmony filter
   根据对联下联发音特点，过滤最后一词发音不符的下联候选

基于其他特征重排序

FS:　海阔凭鱼跃
SS1: 天高任鸟飞
SS2: 天高任狗叫
SS1和SS2具有相同的得分(SMT、语言模型), 但"天高"和"狗叫"放一起没有意义

1. Mutual information (MI) score

$$MI(S)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} I(s_i,s_j)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log\frac{p(s_i,s_j)}{p(s_i)p(s_j)}$$

基于其他特征重排序

海阔凭鱼跃，天高任鸟飞
FS:“海”和“阔”，“海”和“鱼”，“鱼”和“跃”具有较强联系
SS:“天”和“高”，“天”和“鸟”，“鸟”和“飞”具有较强联系

2. MI-based structural similarity

$$F = \{f_1, f_2, \ldots, f_n\}$$

$$V_f = \{v_{12}, v_{13}, \ldots, v_{1n}, v_{23}, \ldots, v_{n-1n}\}$$ (Vij是fi, fj的互信息得分)

$$MISS(F, S) = \cos(V_f, V_s) = \frac{V_f \bullet V_s}{|V_f| \times |V_s|}$$

Reranking model: Ranking SVM

生成对联的评价方法

指标： BLEU, human evaluation

1. Bleu与人工评价具有0.92相关性
2. 所选特征对于生成对联的有效性
3. 人工评价自动生成对联中可接受的比例

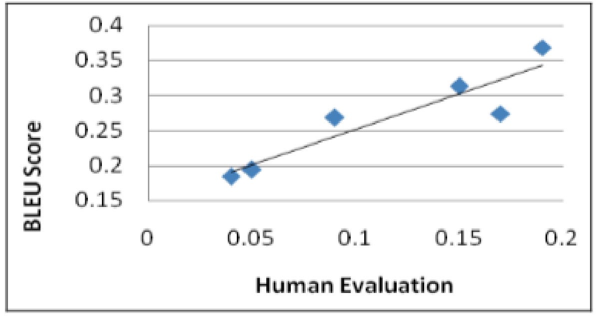


Figure 4: BLEU Predicts Human Judgments.

	Features	BLEU
Phrase-based SMT Model	Phrase TM(PTM) + LM	0.276
	+ Inverted PTM	0.282
	+ Lexical Weight (LW)	0.315
	+ Inverted LW	0.348
Ranking SVM	+ Mutual information (MI)	0.356
	+ MI-based structural similarity	0.361

Table 3. Feature Evaluation.

	Top-1	Top-10
Top-n inclusion rate	0.21	0.73

Table 4. Overall Performance Evaluation.

诗句自动生成

输入代表诗主题的词,逐句生成直到末句，最后诗题

微软古诗：http://duilian.msra.cn/jueju/

诗句特点

1. 严格的声调模式(平仄)

   五言绝句和七言绝句具有4中常见的平仄结构，右侧为其中一种

2. 押韵

   押韵的字具有具有相同元音结尾，诗句第二句和第四句的结尾押韵

3. 结构限制

   绝句具有“起, 承, 转, 合”的结构

* + - - +
- - + + -
* - - + +
* + + - -

+代表仄，-代表平，*代表任意一种

登鹳雀楼
On the Stork Tower
王之涣

白日依山尽，(-+--+)
white sunlight along hill fade
黄河入海流。(--++-)
Yellow River into sea flow
欲穷千里目，(+--++)
wish exhaust thousand mile eyesight
更上一层楼。(+++--)
More up one story tower

诗句自动生成

1. 基于模板生成第一句
   a) 用户从已经聚类的短语对应的词中选择关键词，根据关键词得到候选短语
   b) 所有候选短语放入可以满足平仄模式的所有可能位置，构成短语格子
   c) 采用语言模型计算格子中所有路径的得分
   d) Forward-Viterbi-Backward-A* 算法选择N-best候选句子
2. SMT model生成四句诗
   a) 上句生成下句作为翻译任务，采用SMT model(对联自动生成的SMT)
   b) 三个不同的SMT系统，分别生成第二句、第三句和第四句
   c) Coherence model将之前的句子中的关键词用于SMT 系统，排序生成句子候选

诗句自动生成

Coherence Model

$$S^* = \arg\max_S p(S \mid F) = \arg\max_S \sum_{i=1}^{M} \lambda_i \log h_i(S, F)$$

1. 上述用于SMT的公式只考虑到相邻绝句的信息，而诗的四个句子都要连贯
2. Coherence Model 计算下一句与之前生成所有的句子的互信息，当做SMT系统的第六个特征

$$h_6(S,F) = \sum_{i,j} MI(s_i, s_j) = \sum_{i,j} \log \frac{p(s_i, s_j)}{p(s_i)\,p(s_j)}$$

si,sj代表SetA和SetB中的字，SetA由已生成的诗句的字组成，SetB由下一句的字组成

$h_1(S,F) = \prod_{i=1}^{I} p(\bar{f}_i \mid \bar{s}_i)$	Phrase translation model
$h_2(S,F) = \prod_{i=1}^{I} p(\bar{s}_i \mid \bar{f}_i)$	Inverted phrase translation model
$h_3(S,F) = \prod_{i=1}^{I} p_w(\bar{f}_i \mid \bar{s}_i)$	Lexical weight
$h_4(S,F) = \prod_{i=1}^{I} p_w(\bar{s}_i \mid \bar{f}_i)$	Inverted lexical weight
$h_5(S,F) = p(S)$	Language model

Table 1. Features Used in the SMT Model.

总结与展望

总结：

- 定义
- 对联自动生成
  - SMT model
  - 候选对联重排序
- 诗句自动生成
  - 基于模板生成第一句
  - SMT model 和 coherence model

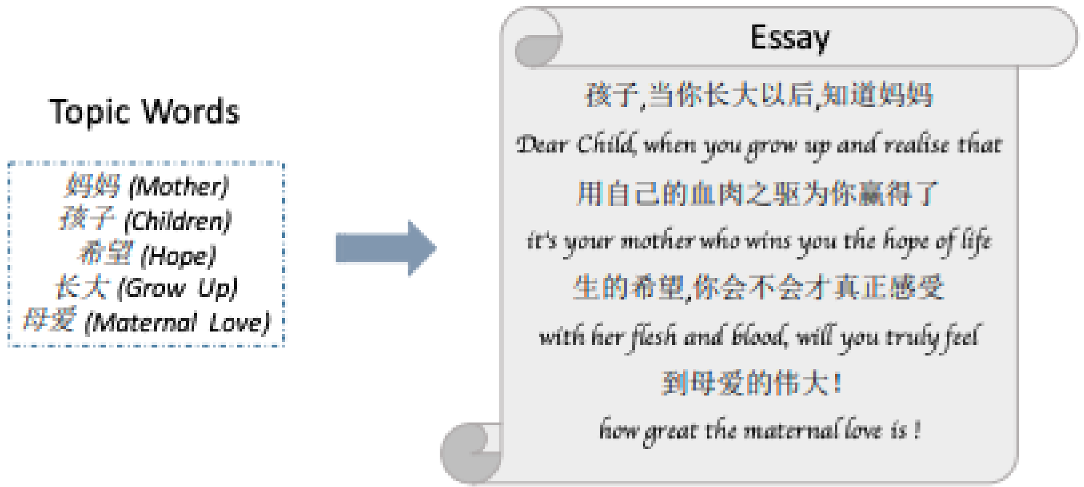


Figure 1: Chinese example for essay generation with five topics.

展望：

- 自动生成作文
- 自动生成对话

四、图像到文本的生成技术

定义

三阶段的流水线模式

相关研究现状

- 依靠三阶段流水线模式
- 依靠计算机视觉提取图像中物体及定位技术
- 图像语义标注与自然语言句子生成联合建模

小结

定义及类别

图像到文本的生成技术是指根据给定的图像生成描述该图像内容的自然语言文本，例如新闻图像附带的标题、医学图像附属的说明、儿童教育中常见的看图说话、以及用户在微博等互联网应用中上传图片时提供的说明文字

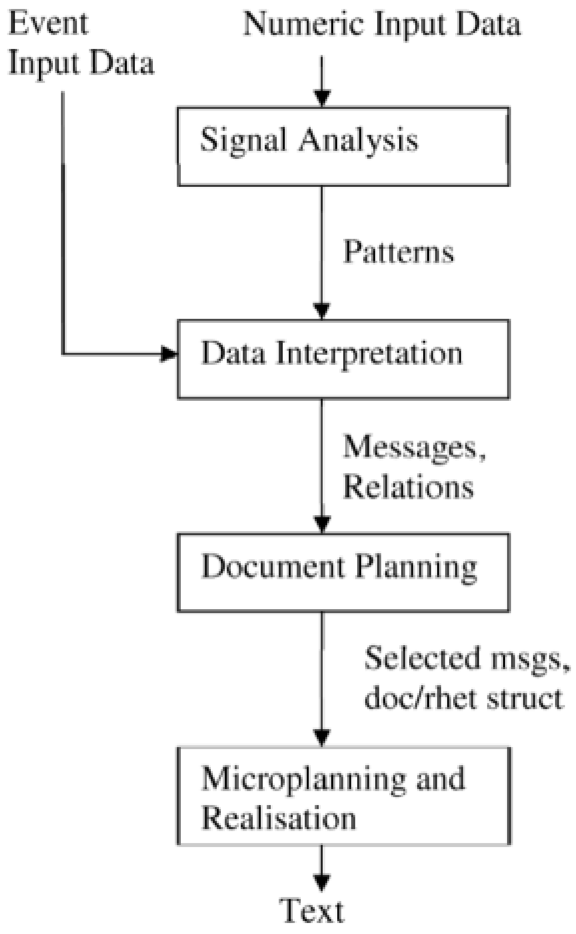
所生成自然语言文本的详细程度及长度的不同，分为图像标题自动生成和图像说明自动生成

图像到文本的生成技术

遵循三阶段流水线模型

根据图像内容理解的特点，做出三部分调整

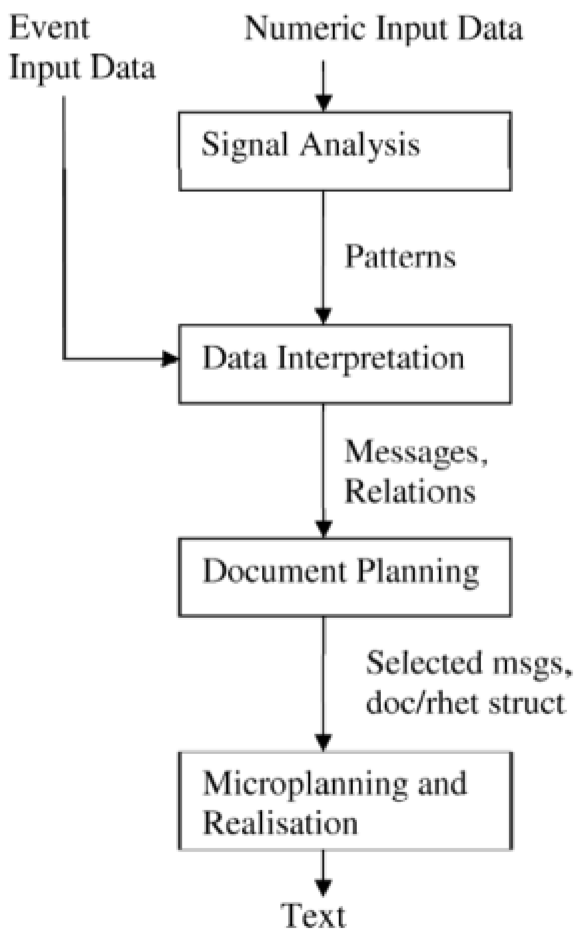
- 内容抽取方面
- 句子内容选择方面
- 句子实现方面



三阶段的流水线模式

内容抽取方面，

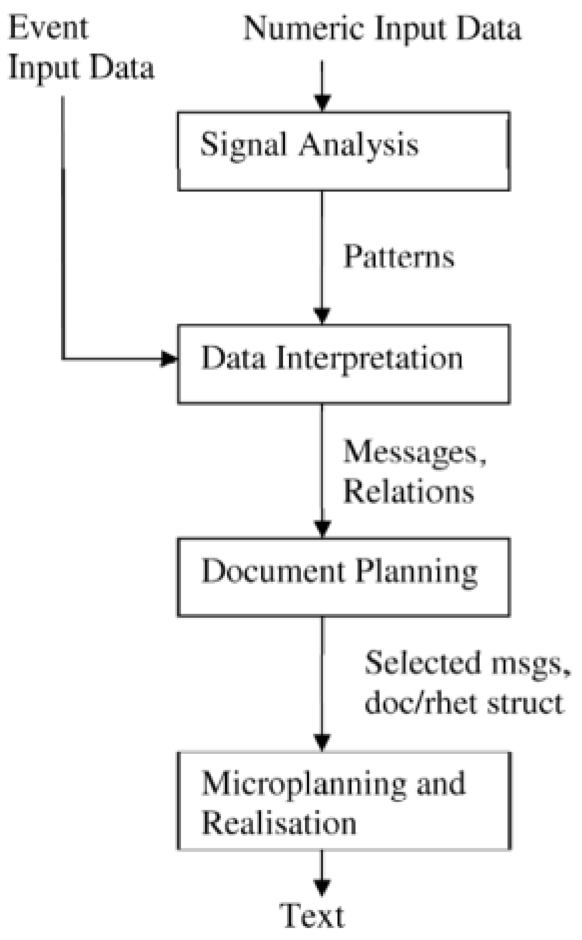
- 需要从图像中抽取物体、方位、动作、场景等概念，其中物体可以具体定位到图像中的某一具体区域，而其他概念则需要进行语义标引。这部分主要依靠模式识别和计算机视觉技术



三阶段的流水线模式

句子内容选择方面

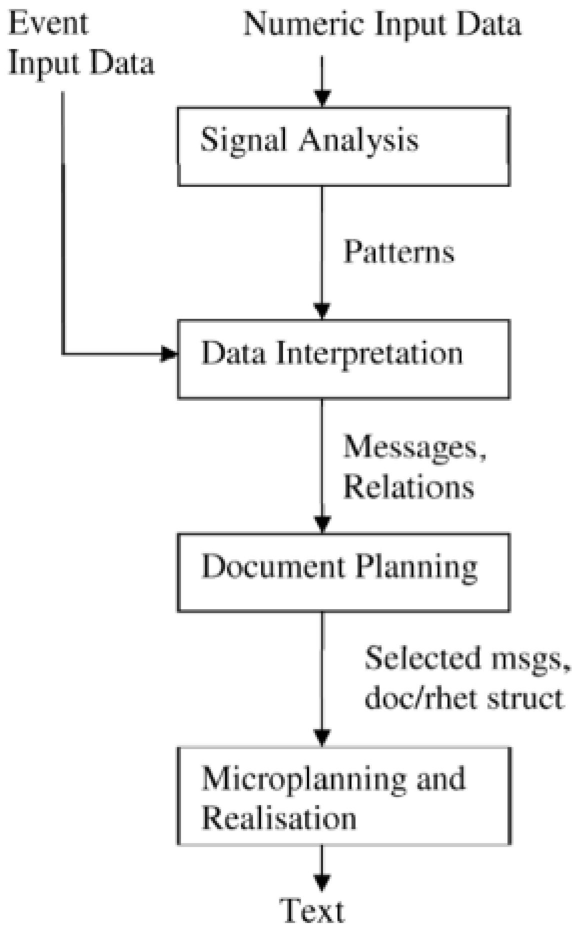
- 需要依据应用场景，选择最重要（如图像画面中最突出的，或与应用场景最相关的），且意义表述连贯的概念。这部分需要综合运用计算机视觉与自然语言处理技术



三阶段的流水线模式

句子实现部分

- 根据实际应用特点选取适当的表述方式将所选择的概念梳理为合乎语法习惯的自然语言句子。这部分主要依靠自然语言处理技术



相关研究现状

依靠三阶段流水线模式

- 基于图像描述模板
- 基于概率图模型

依靠计算机视觉提取图像中物体及定位技术

- 基于概率图和语言模型
- 基于核函数的典型关联分析

图像语义标注与自然语言句子生成联合建模

- 基于多模态m-RNN
- 基于视觉中注意力机制

基于图像描述模板

图像被细致的分割并标注为物体及其组成部分，以及图像所表现的场景，并在此基础上选择与场景相关的描述模板，将物体识别的结果填充入模板得到图像的描述文字

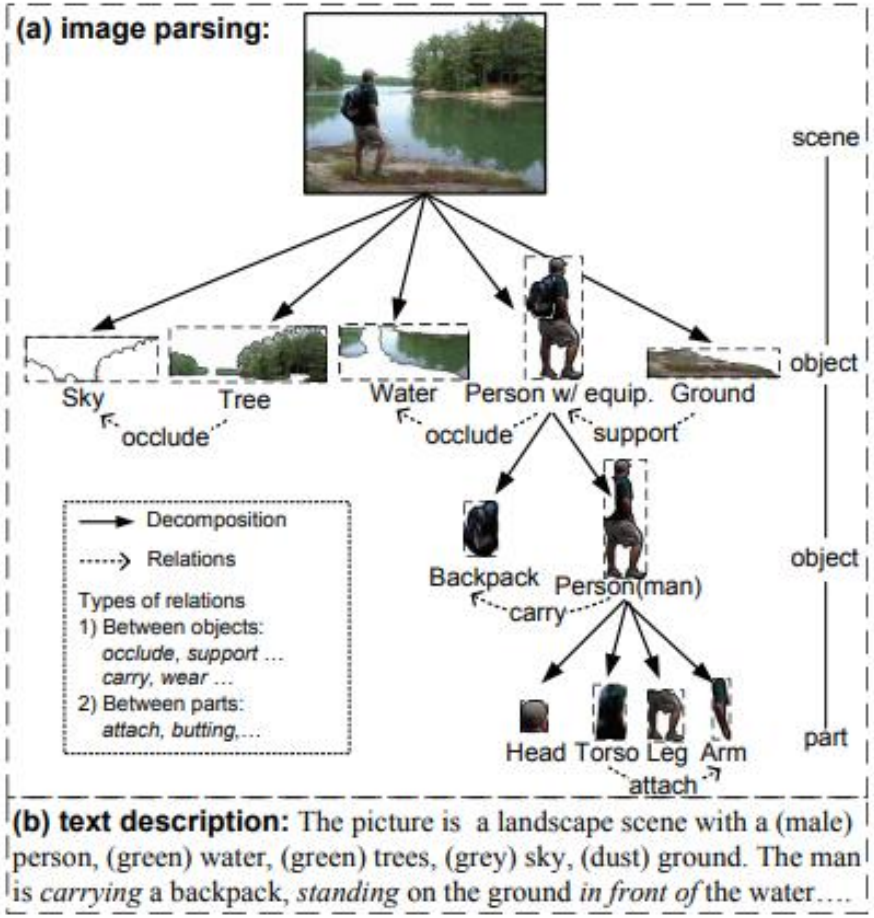


Fig. 1. Two major tasks of the I2T framework: (a) image parsing and (b) text description. See text for more details.

基于概率图模型

采用概率图模型对文本信息和图像信息同时建模，并从新闻图片所在的文字报道中挑选合适的关键词作为体现图像内容的关键词，并进而利用语言模型将选取的内容关键词、及必要的功能词汇链接为基本合乎语法规则的图像标题

Thousands of Tongans have attended the funeral of King Taufa'ahau Tupou IV, who died last week at the age of 88. Representatives from 30 foreign countries watched as the king's coffin was carried by 1,000 men to the official royal burial ground.		**King Tupou, who was 88, died a week ago.**
Contaminated Cadbury's chocolate was the most likely cause of an outbreak of salmonella poisoning, the Health Protection Agency has said. About 36 out of a total of 56 cases of the illness reported between March and July could be linked to the product.		**Cadbury will increase its contamination testing levels.**
A Nasa satellite has documented startling changes in Arctic sea ice cover between 2004 and 2005. The extent of "perennial" ice declined by 14%, losing an area the size of Pakistan or Turkey. The last few decades have seen ice cover shrink by about 0.7% per year.		**Satellite instruments can distinguish "old" Arctic ice from "new".**
A third of children in the UK use blogs and social network websites but two thirds of parents do not even know what they are, a survey suggests. The children's charity NCH said there was "an alarming gap" in technological knowledge between generations.		**Children were found to be far more internet-wise than parents.**

Table 1: Each entry in the BBC News database contains a document an image, and its caption.

依靠计算机视觉提取图像中物体及定位技术

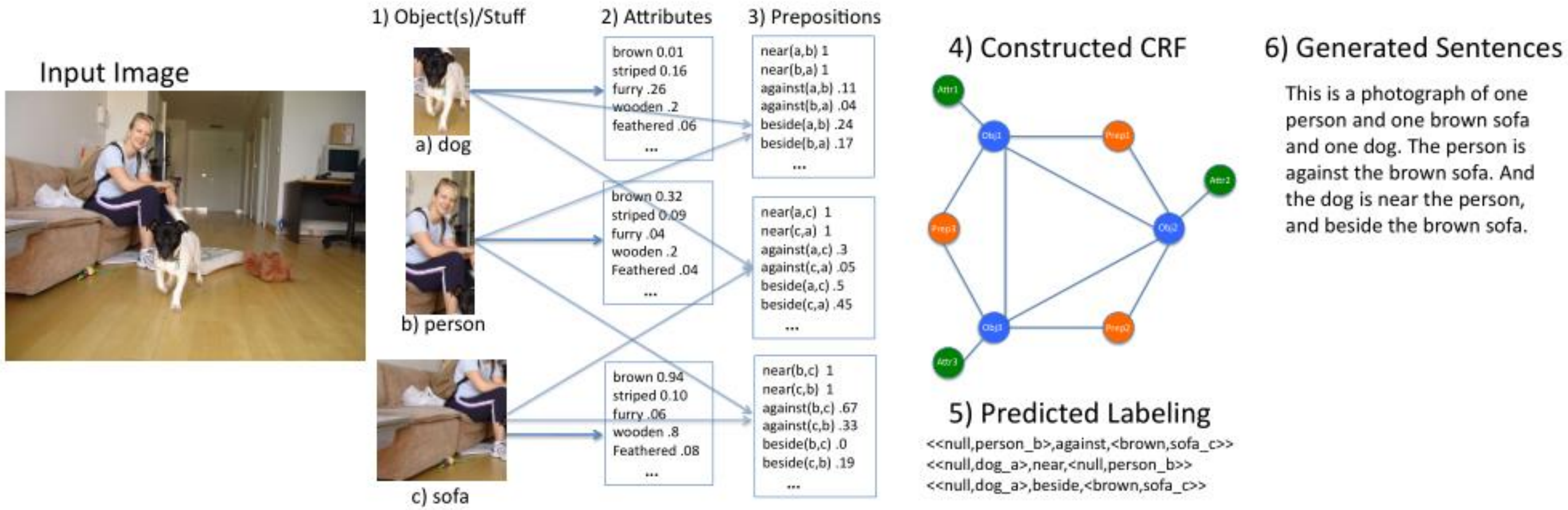


Figure 2. System flow for an example image: 1) object and stuff detectors find candidate objects, 2) each candidate region is processed by a set of attribute classifiers, 3) each pair of candidate regions is processed by prepositional relationship functions, 4) A CRF is constructed that incorporates the unary image potentials computed by 1-3, and higher order text based potentials computed from large document corpora, 5) A labeling of the graph is predicted, 6) Sentences are generated based on the labeling.

依靠计算机视觉提取图像中物体及定位

依靠计算机视觉领域现有的物体识别技术从图像中抽取物体（包括人物、动物、花草、车、桌子等常见的物体类型），并对其定位以获得物体之间的上下位关系，进而依赖概率图模型和语言模型选取适当的描述顺序将这些物体概念、介词短语块串联成完整的句子

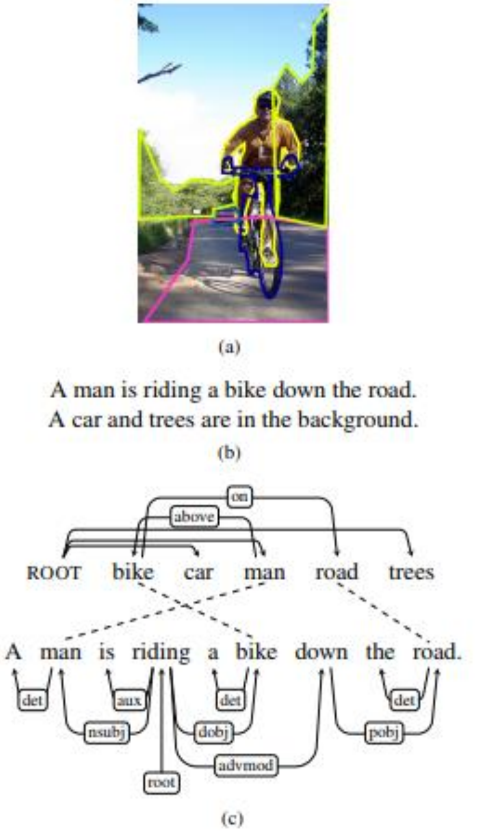


Figure 1: (a) Image with regions marked up: BIKE, CAR, MAN, ROAD, TREES; (b) human-generated image description; (c) visual dependency representation expressing the relationships between MAN, BIKE, and ROAD aligned to the syntactic dependency parse of the first sentence in the human-generated description (b).

基于核函数的典型关联分析

使用Kernel Canonical Correlation Analysis(KCCA)将图片和文本映射到共享的潜在语义空间

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watch
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

Figure 1: An example of an image from the Flickr 8K dataset. Each of the captions literally describe what is being depicted in the photograph while also mentioning different entities and exhibiting linguistic variation

图像语义标注与自然语言句子生成联合建模

1. 在图像端采用多层深度卷积神经网络（Deep Convolution Neural Network，DCNN）对图像中的物体概念进行建模
2. 在文本端采用循环神经网络（Recurrent Neural Network，RNN）或递归神经网络（Recursive Neural Network）对自然语言句子的生成过程进行建模

多模态M-RNN模型

Mao 等人通过 DCNN 得到的图像信息与文本信息融合到同一个循环神经网络（m-RNN）中，将图像信息融入到了自然语言句子生成的序列过程

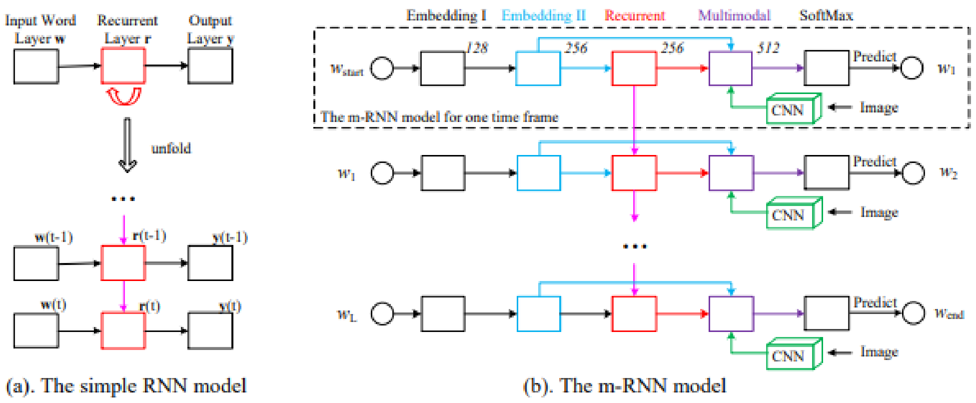


Figure 2: Illustration of the simple Recurrent Neural Network (RNN) and our multimodal Recurrent Neural Network (m-RNN) architecture. (a). The simple RNN. (b). Our m-RNN model. The inputs of our model are an image and its corresponding sentence descriptions. $w_1, w_2, ..., w_L$ represents the words in a sentence. We add a start sign w_{start} and an end sign w_{end} to all the training sentences. The model estimates the probability distribution of the next word given previous words and the image. It consists of five layers (i.e. two word embedding layers, a recurrent layer, a multimodal layer and a softmax layer) and a deep CNN in each time frame. The number above each layer indicates the dimension of the layer. The weights are shared among all the time frames. (Best viewed in color)

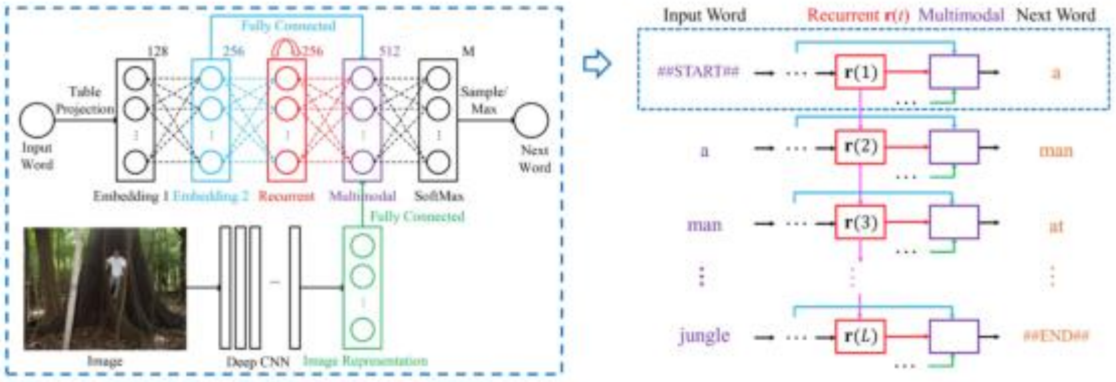
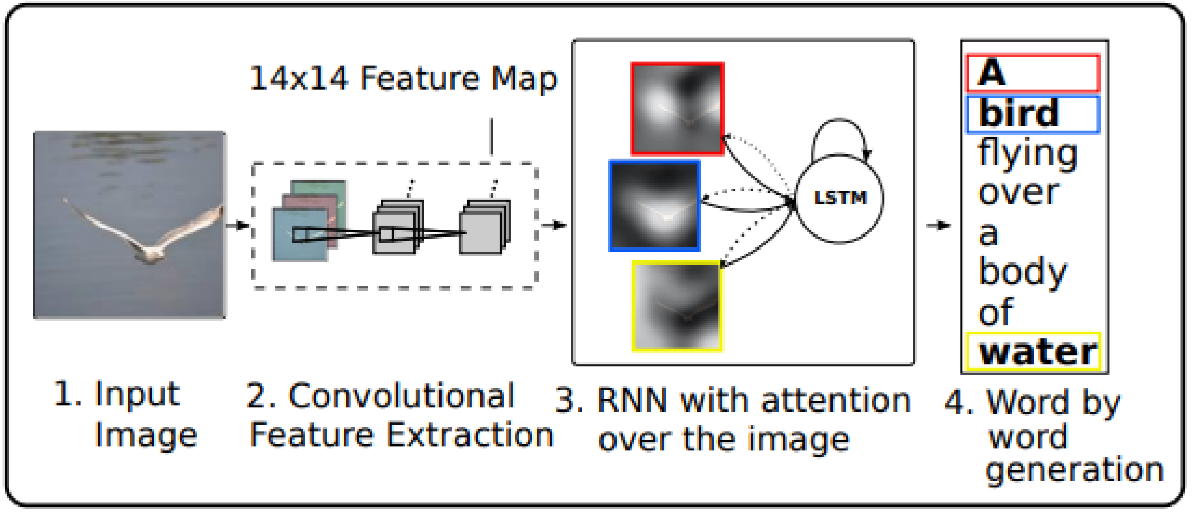


图 5.2：多模态 m-RNN 模型[100]

视觉“注意”引导的图像标题生成

Xu等人提出利用计算机视觉领域中的“注意”(Attention)机制来促进词语和图像块之间的对齐，从而在句子生成过程中，模拟人视觉的“注意”转移过程能够与词语序列的生成过程相互促进，使生成的句子更符合人的表述习惯

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in Sections 3.1 & 5.4



基于卷积神经网络 CNN 和多示例学习

微软的研究人员利用卷积神经网络 CNN 和多示例学习（Multiple Instance Learning，MIL）对图像建模，并利用判别式语言模型生成候选句子，并采用统计机器翻译研究中经典的最小误差率训练(Minimum Error Rate Training，MERT)来发掘文本和图像层面的特征对候选句子进行排序

	D-ME+DMSM	a plate with a sandwich and a cup of coffee
	MRNN	a close up of a plate of food
	D-ME+DMSM+MRNN	a plate of food and a cup of coffee
	k-NN	a cup of coffee on a plate with a spoon
	D-ME+DMSM	a black bear walking across a lush green forest
	MRNN	a couple of bears walking across a dirt road
	D-ME+DMSM+MRNN	a black bear walking through a wooded area
	k-NN	a black bear that is walking in the woods
	D-ME+DMSM	a gray and white cat sitting on top of it
	MRNN	a cat sitting in front of a mirror
	D-ME+DMSM+MRNN	a close up of a cat looking at the camera
	k-NN	a cat sitting on top of a wooden table

Table 2: Example generated captions.

小结

定义

三阶段流水模型

- 内容抽取方面
- 句子内容选择方面
- 句子实现方面

相关研究现状

- 依靠三阶段流水线模式
- 依靠计算机视觉提取图像中物体及定位技术
- 图像语义标注与自然语言句子生成联合建模