基于深度学习的自然语言处理：边界在哪里?

刘群，华为诺亚方舟实验室

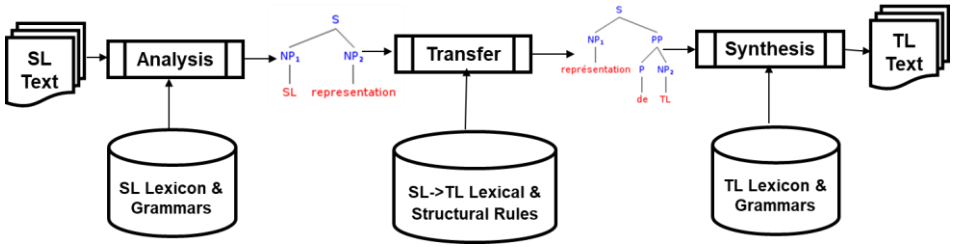
www.huawei.com

第四届语言与智能高峰论坛, 北京语言大学，2019-8-24

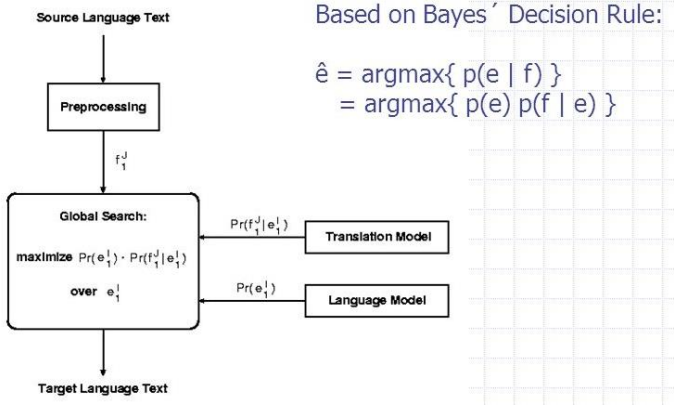
HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

自然语言处理的范式迁移

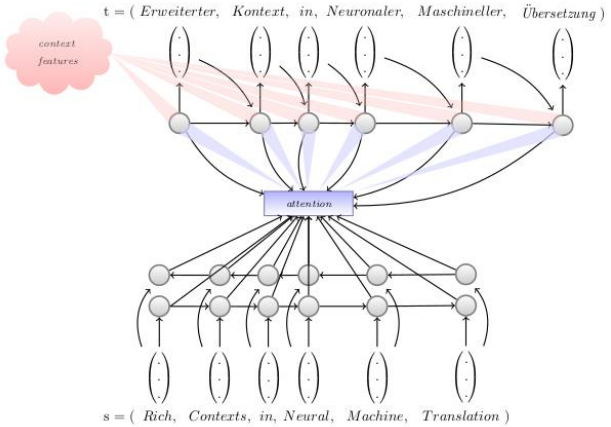


以机器翻译为例



Based on Bayes´ Decision Rule:

$$\hat{e} = \text{argmax}\{ p(e \mid f) \}$$
$$= \text{argmax}\{ p(e)\, p(f \mid e) \}$$



目录

深度学习解决了自然语言处理的哪些问题?

HUAWEI

词语形态问题

句法结构问题

多语言问题

联合训练问题

领域迁移问题

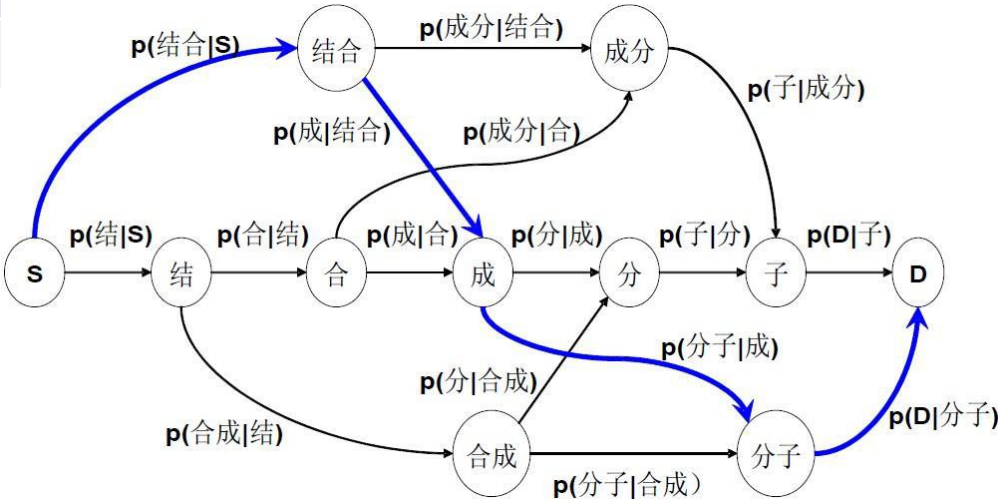
在线学习问题

语言形态问题

- 在RBMT和SMT框架下，形态分析是机器翻译首先需要解决的问题：
  - 对于中文来说：
    - 基于汉字的翻译效果很差，因此分词是必须的；
    - 中文词语不是一个well-defined的语言单位，分词缺乏统一的规范，分词粒度难以把握。
  - 对于形态丰富语言来说：
    - 形态分析本身难度很大，需要语言学家深度介入；
    - 形态本身是一层结构，很难融入到SMT已有的框架中。

语言形态问题

研究/生命/的/起源	研究生/命/的/起源
他/从/马/上/下来	他/从/马上/下来
乒乓球/拍卖/完了	乒乓/球拍/卖/完了
和/特朗普/通话	和/特朗/普通话



汉语词语切分歧义

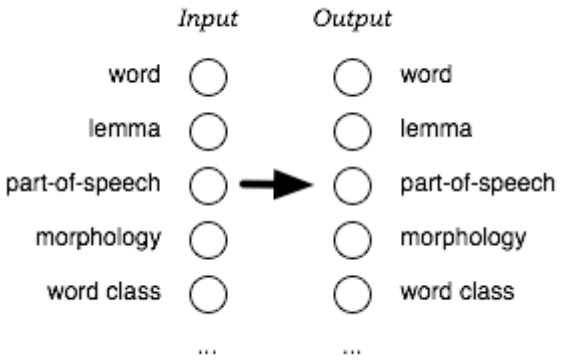
语言形态问题

Word	Translation
Turkish:	
terbiye	good manners
terbiye+siz	rude
terbiye+siz+lik	rudeness
terbiye+siz+lik+leri	their rudeness
terbiye+siz+lik+leri+nden	from their rudeness
terbiye+siz+lik+leri+nden+mis	it was because of their rudeness
Farsi:	
drāmd	income
pr+**drāmd**	wealthy
pr+**drāmd**+tar	more wealthy
pr+**drāmd**+tar+in	the most wealthy
pr+**drāmd**+tar+in+hā	the most wealthy people
pr+**drāmd**+tar+in+hā+yshān	the most wealthy group of them
pr+**drāmd**+tar+in+hā+yshān+nd	they are the most wealthy group of them

复杂形态语言的机器翻译

语言形态问题

SMT的解决方案（之一）



Factored statistical machine translation

Koehn & Hoang, 2007, https://www.aclweb.org/anthology/D07-1091

语言形态问题

- NMT框架下，语言形态不再构成严重问题：
  - 中文词语切分不再是必须的，汉字作为建模的基本单位可以取得很好的效果
  - 对于形态复杂的语言，基于subword的模型和基于character的模型为翻译建模提供了统一而优雅的解决方案

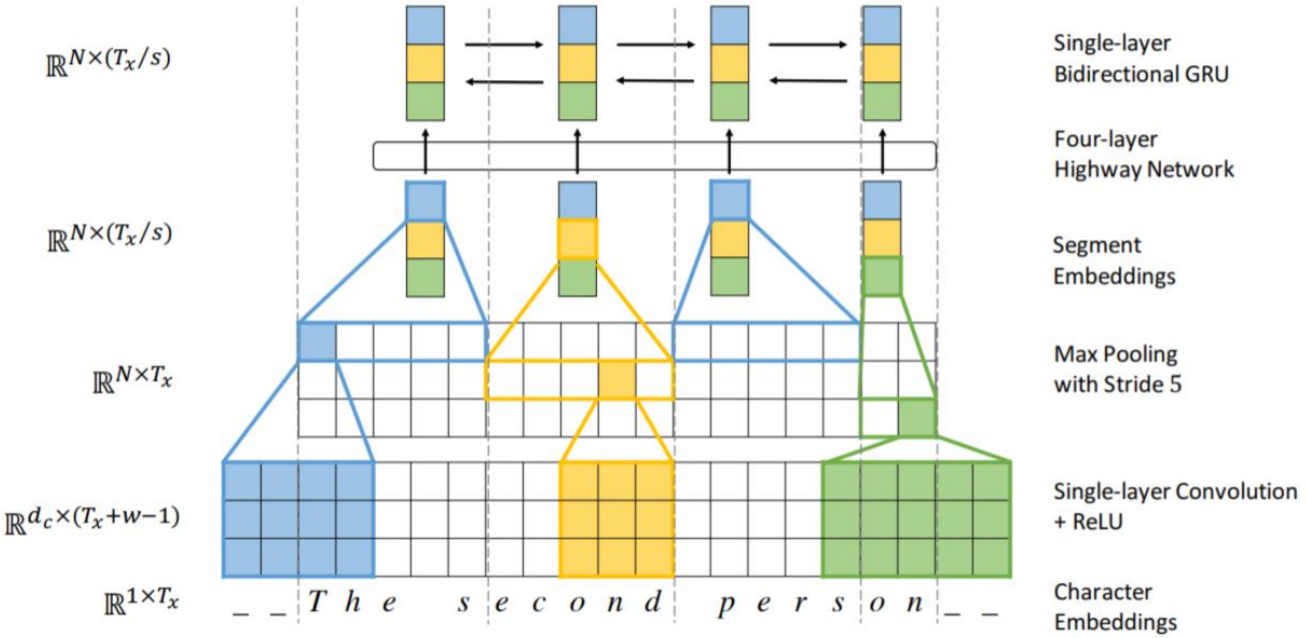
语言形态问题

NMT的解决方案

Sentence:	龙年新春，繁花似锦的深圳处处洋溢着欢乐祥和的气氛。
Word:	龙年 新春， 繁花似锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。
Character:	龙 年 新 春， 繁 花 似 锦 的 深 圳 处 处 洋 溢 着 欢 乐 祥 和 的 气 氛。
Hybrid :	<B>龙 <E>年 新春， <B>繁 <M>花 <M>似 <E>锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛。
BPE:	龙年 新春 ， 繁花@@ 似@@ 锦 的 深圳 处处 洋溢 着 欢乐 祥和 的 气氛 。
Wordpiece:	_龙 年 _新春 _， _繁花 似 锦 _的 _深圳 _处处 _洋 溢 _着 _欢乐 _祥 和 _的 _气氛 _。

图片来自：Wang et al. 2017 arxiv:1711.04457

语言形态问题



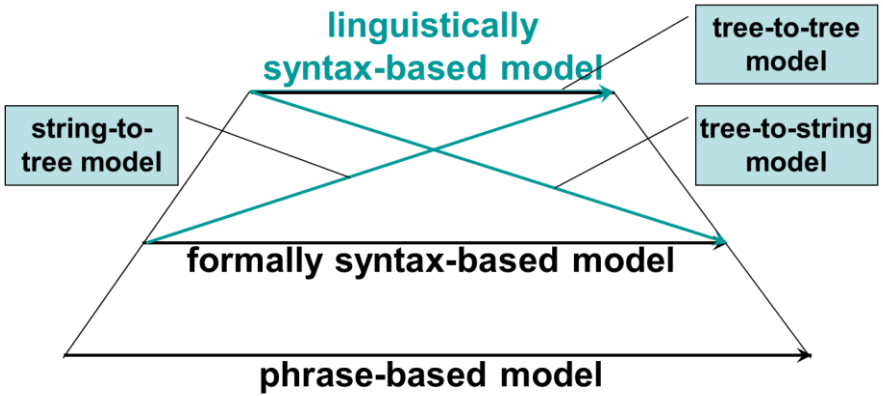
图片来自： Lee et al., TACL 2017, https://aclweb.org/anthology/Q17-1026)

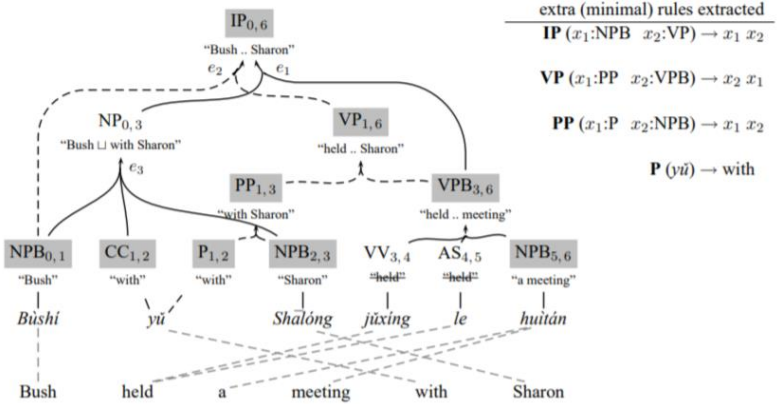
句法结构问题

- 在RBMT和SMT框架下，句法分析对机器翻译的质量起着重要的作用
  - RBMT中，句法分析是核心模块之一，没有句法分析机器翻译几乎寸步难行；
  - SMT中，基于短语的方法获得很大成功，对于句法结构相似的语言效果很好，但对于句法结构相差较大的语言，基于句法的方法仍然比基于短语的方法效果明显提高
    - 句法结构带来模型的复杂性增加
    - 句法分析的错误影响翻译性能
- 在NMT框架下，神经网络可以很好地捕捉句子的结构，无需进行句法分析，系统可以自动获得处理复杂结构句子翻译的能力。

句法结构问题

SMT框架下的解决方案





Mi & Huang, 2018, https://www.aclweb.org/anthology/D08-1022

句法结构问题

NMT框架下语言句法结构差异大部分情况下不再构成问题

- 第二家加拿大公司因被发现害虫而被从向中国运输油菜籽的名单中除名。
- The second Canadian company was removed from the list of transporting rapeseed to China due to the discovery of pests.

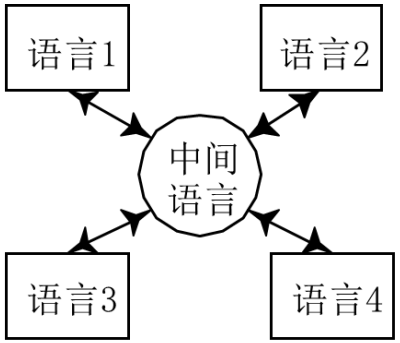
- 张三因被发现考试作弊而被从向欧洲派遣的留学生名单中除名。
- John Doe was removed from the list of foreign students sent to Europe after he was found to have cheated on a test.

令人惊讶的是，NMT模型的训练只是使用双语的纯文本信息，没有使用任何句法信息。

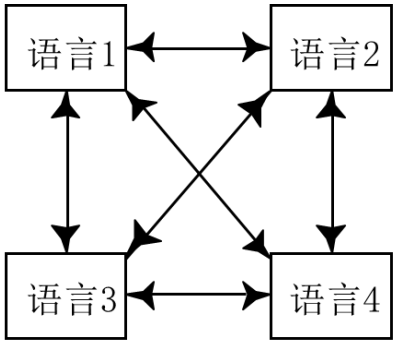
多语言问题

- 在RBMT时代，开发多语言机器翻译系统代价极高，其中较为理想的中间语言（Interlingua）方案，由于系统过于复杂，成为“不可承受之重”
- 在SMT时代，由于广泛采用Pivot机制，多语言机器翻译成为可能。但Pivot机制也导致错误的传播和翻译性能的下降，各种语言之间的相似性无法被利用来改进机器翻译的质量
- 在NMT时代，单一的多语言机器翻译系统被提出来并被验证有效，中间语言的理想初步得以实现。另外，Multilingual BERT的推出，为104种语言提供了统一的预训练语言模型，大大增加了对多语言nlp的支持，对资源稀疏语言的帮助尤其明显。

多语言问题



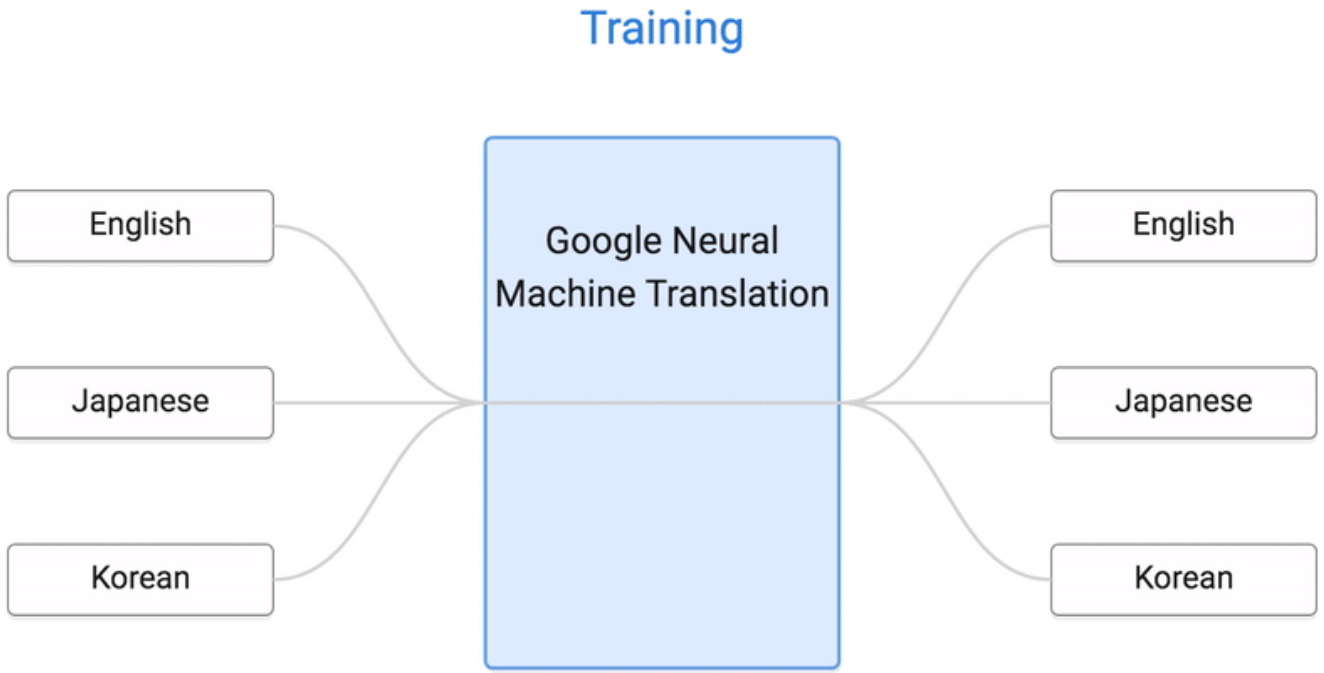
中间语言方法



转换方法

Makoto Nagao (Kyoto University) said: “.. when the pivot language [i.e. interlingua] is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place. This level of subtlety is a practical impossibility.” (Machine Translation, Oxford, 1989)

多语言问题



Zero-Shot Translation with Google’s Multilingual Neural Machine Translation System
https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

多语言问题

Google Multilingual BERT: Covering 104 lanugages

联合训练问题

- 在SMT框架下，
  - 由于各模块独立训练，导致错误传播问题严重，联合训练成为提高性能的有效手段；
  - 但联合训练又会导致模型复杂度大大增加，开发和维护都变得困难。同时由于搜索范围急剧扩大，系统开销也严重增加。另外，由于模块太多，只能有限的模块进行联合训练，不可能都纳入联合训练
- 在NMT框架下，
  - 端到端训练成为标准模式，所有模块构成一个有机的整体，针对同一个目标函数同时训练，有避免了错误传播，提高了系统性能

目录

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

可信任性问题

可控制性问题

超长文本问题

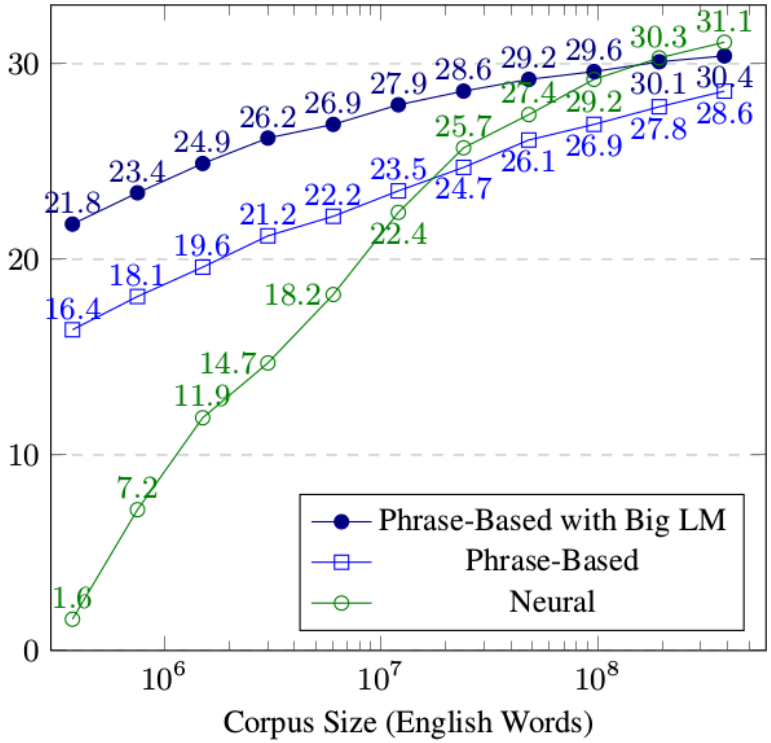
缺乏常识问题

资源稀缺问题

- 资源稀缺问题远比大部分人想象的要严重得多
  - 绝大多数自然语言都是资源稀缺语言
  - 绝大多数的专业领域都缺乏足够的数据资源
  - 工业界遇到的大部分问题都只有极少或者完全没有标注数据
- 以WMT2019的Biomedical MT Task为例：

Language pairs	Medline training		Medline test		Terminology test
	Documents	Sentences	Documents	Sentences	Terms
de/en	3,669	40,398	50	589	-
en/de			50	719	-
es/en	8,626	100,257	50	526	-
en/es			50	599	6,624
fr/en	6,540	75,049	50	486	-
en/fr			50	593	-
pt/en	4,185	49,918	50	491	-
en/pt			50	589	-
zh/en	-	-	50	283	-
en/zh			50	351	-

资源稀缺问题



Koehn & Knowles, 2017, https://arxiv.org/pdf/1706.03872.pdf

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

可信任性问题

可控制性问题

超长文本问题

缺乏常识问题

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

可信任性问题

可控制性问题

超长文本问题

缺乏常识问题

可信任性问题

- 对于一些关键性的应用，比如疾病诊断，如果一个AI系统不能对其作出的诊断给出合理的解释，这样的诊断是难以获得医生和病人的信任的，也是无法投入使用的。
  - 这种情况下，可信任信问题本质上可以归结于可解释性问题。
- 对于非关键性应用，系统不应该犯严重错误，以机器翻译为例：
  - 重要的人名、地名、机构名不应该翻译错误；
  - 不能翻译成相反的意思，但这种情况很难避免，因为意思相反的表达在统计上的表现是非常相似的；
  - 不应该犯过于幼稚的错误，一些简单的、常见的表达不应该翻译错误。

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

可信任性问题

可控制性问题

超长文本问题

缺乏常识问题

可控制性问题

- 对于重要的人名、地名、机构名、术语，我们希望严格按照给定的方式进行翻译，不能随便乱翻。

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

可信任性问题

可控制性问题

超长文本问题

缺乏常识问题

超长文本问题

- 现在的神经网络在处理长文本方面，已经取得了很大进步：
  - 早期的NMT系统在翻译长句子的时候，翻译质量会大幅下降，但现在已经好得多了
  - 预训练语言模型如BERT和GPT，通常的训练长度都在几百词到上千词，GPT-2模型生成的千词以内的问题已经非常流畅，连贯性也非常好
- 但长文本处理问题仍然面临很大的挑战：
  - 目前在基于篇章的机器翻译研究中，对改进翻译质量起作用的上下文只有前1-3个句子，更长上下文反倒会降低当前句的翻译质量
  - 基于Transformer的预训练语言模型消耗计算资源巨大，计算所需时空消耗会随着句子长度呈平方或者三次方增长，现有模型无法支持更长的文本。

还有哪些自然语言处理问题深度学习没有解决？

资源稀缺问题

可解释性问题

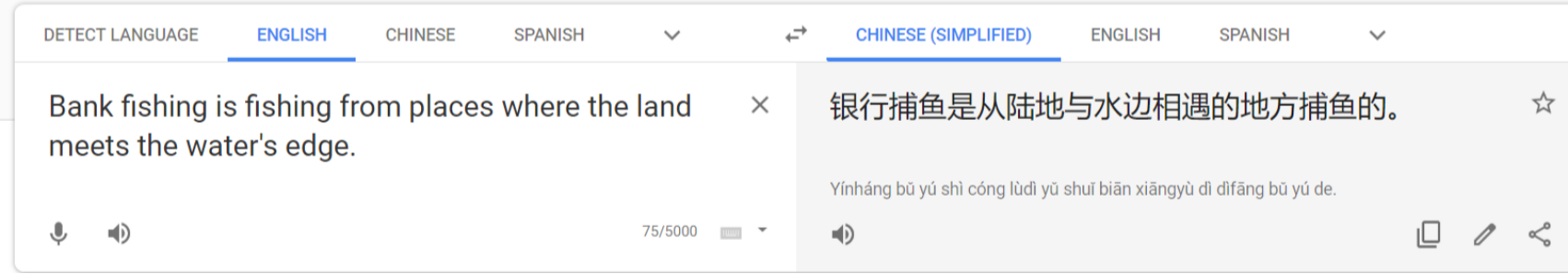
可信任性问题

可控制性问题

超长文本问题

缺乏常识问题

缺乏常识问题



- Bank（银行、岸）这样一个经典的歧义词在NMT时代仍然无法避免翻译错误，即使有fishing、water这样的相关提示词存在

（感谢董振东老师提供的例子）

缺乏常识问题

GPT-2虽然具有强大的文本生成能力，可以生成非常流畅和连贯的文本，但仍然会犯一些常识性错误：

- **HUMAN INPUT**

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

- **MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)**

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

目录

基于深度学习的自然语言处理：边界在哪里？

数据边界

语义边界

符号边界

因果边界

基于深度学习的自然语言处理：边界在哪里？

数据边界

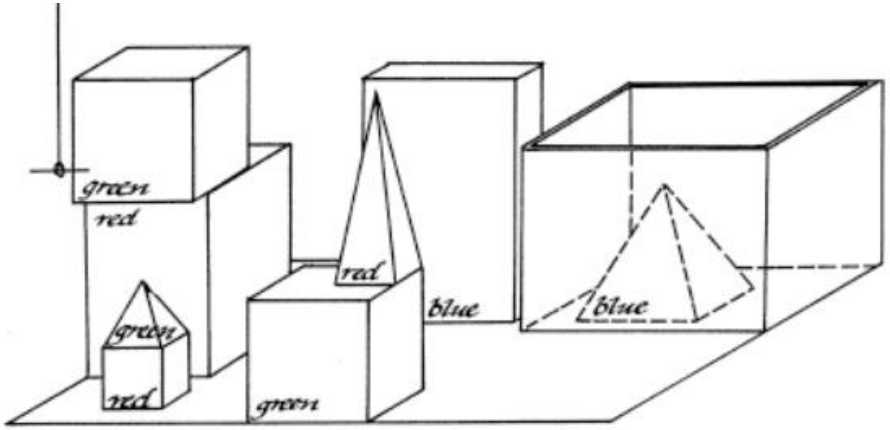
语义边界

符号边界

因果边界

语义边界

- 人工智能之所以在围棋、电子竞技等项目上，包括早期的Winograd系统上，都能大获成功，一个主要原因是这些问题都是well-defined的问题。



- 在这样的系统中，我们对客观世界有着精确的建模，系统的所有操作，对在这个世界模型上。

语义边界

- 音箱、语音助手之类的产品能够取得成功，很大程度上也是因为这些系统对对应着明确定义的任务，但一旦用户的问话超出这些预定义的任务，系统也很容易出错。
- 机器翻译的成功是一个比较特殊的例子，因为译文的意义是受到原文意义严格约束的，只要有足够的数据，可以取得较好的效果，但大部分的nlp问题不具备这样的特点
- 大部分的自然语言模型，还只是流于对词语符号之间的关系建模，没有对所描述的问题语义进行建模
- 人在理解自然语言的时候，脑子里是有一个客观世界的模型的，只有当我们能够把一个自然语言的表达转换成大脑里的客观世界模型的某种对应形式的时候，才是真正理解了语言。而目前大部分的nlp系统是没有这个客观世界模型的。

语义边界

- 对客观世界建模是非常复杂的，以“颜色”这个属性为例，可以用三个8位数进行建模，可以组合出数千万种颜色，但刻画颜色的词语只有数十个，词语和颜色模型的对应关系很难准确地进行描述
- 知识图谱（包括知识本体）是人类专家试图对客观世界建立通用性模型的一种长期努力
- 知识图谱已经包含了大量知识，能够排除很多的常识性错误，但现在知识图谱在nlp任务上还没有办法得到大规模成功的应用。
- 我认为理想的nlp系统，是需要有一个描述客观世界的语义模型的，可以把这个语义模型理解成一种隐状态。知识图谱是这种模型的一种可能的形式。
- 带隐状态的模型是非常难训练的。

基于深度学习的自然语言处理：边界在哪里？

数据边界

语义边界

符号边界

因果边界

符号边界

- 一些心理学家把人类的心理活动分为意识和潜意识。
- 一种不太准确的理解，把可以用语言描述的心理活动称作意识，而无法用语言描述的心理活动称为潜意识。
- 人类可以使用语言进行逻辑推理，这是一种强大的能力
- 神经网络可以较好地模拟潜意识活动，也可以从语言输入、并输出语言表达，但神经网络的计算是无法用语言来描述的，而且使用语言的逻辑推理也无法表示成神经网络。这是神经网络方法的重要缺陷
- 一个简单的例子：使用有限状态自动机，可以精确地定义一些特定的表示形式，如数词、年份、网址等等，但再好的神经网络也很难准确地学习到有限状态自动机的表达能力。这是很多实用的nlp系统仍然离不开规则方法的原因。

基于深度学习的自然语言处理：边界在哪里？

数据边界

语义边界

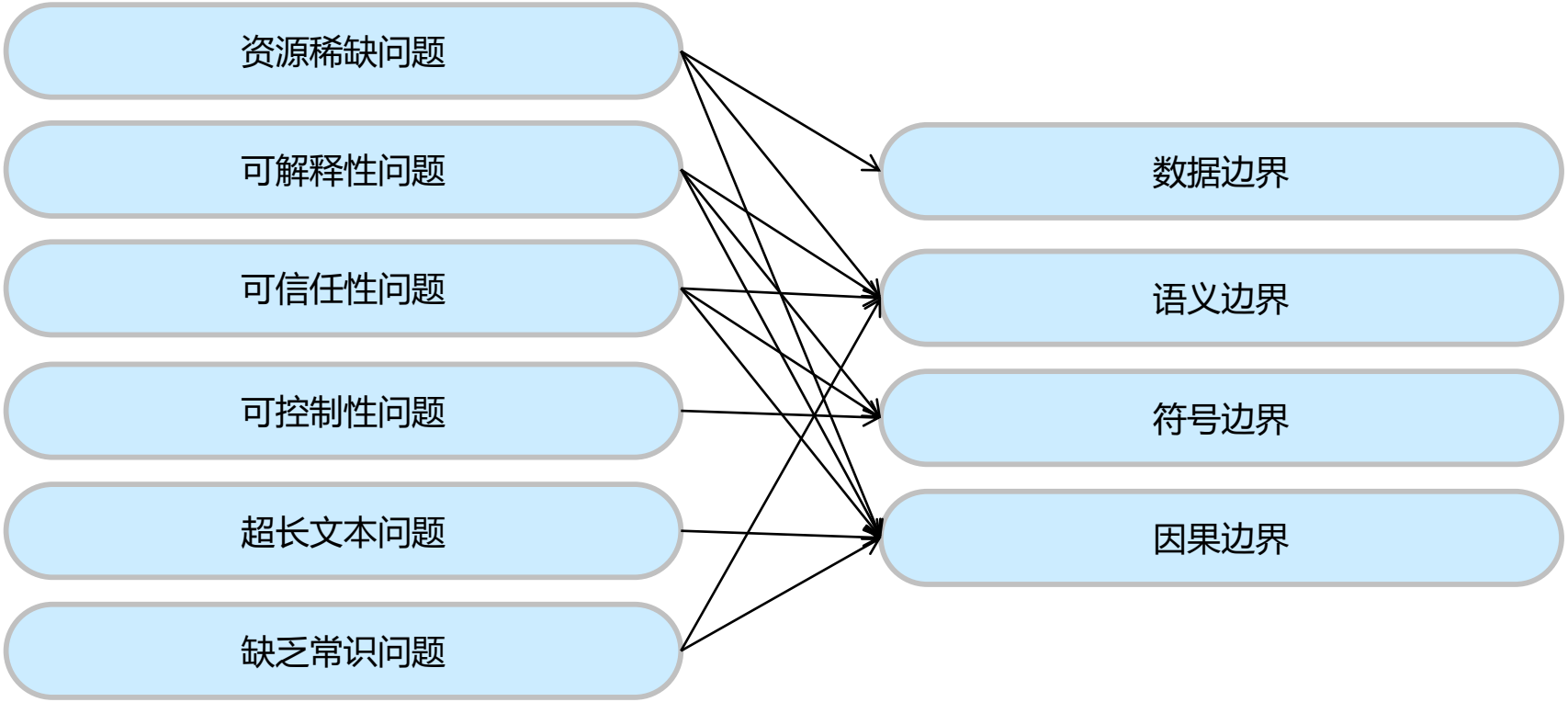
符号边界

因果边界

因果边界

- 人类对客观世界中发生的事情中的因果关系，有明确的理解。所以很容易去芜存真，抓住问题的本质。但神经网络中，做出判断的依据是根据数据学习到的，并没有理解真正的因果关系，因而很容易做出错误的判断
- 仅仅根据统计数据得出的推断，很难反应真正的因果关系。真正的因果关系，只有通过精心设计的实验才能得出（例如药物的有效性实验）

NLP所面临问题和深度学习方法边界的关系



Thank You!

www.huawei.com

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI