自然语言处理中的可理解分析

张奇

复旦大学

自然语言处理中的可理解性

Interpretability

理解　　解释

理解与**解释**是释义学的一对概念。

释义学主张的针对文本而进行的中心工作。对这对概念的关系有不同的阐述。古典释义学家把理解置于首要地位，如德国狄尔泰把理解看作人以心灵力量的整体去认识白己及自己所创造的精神世界的能力。释义学理论家着重研究解释理论，意大利E．贝蒂把解释看作一个过程，其目的和结果是理解，**解释的任务在于使某物得到理解**。赫什认为理解在于原原本本地构造文本的意义，而解释则是解释者对文本的意义所作的阐释，渗杂有解释者附加的成分。

自然语言处理中的可理解性

Interpretability

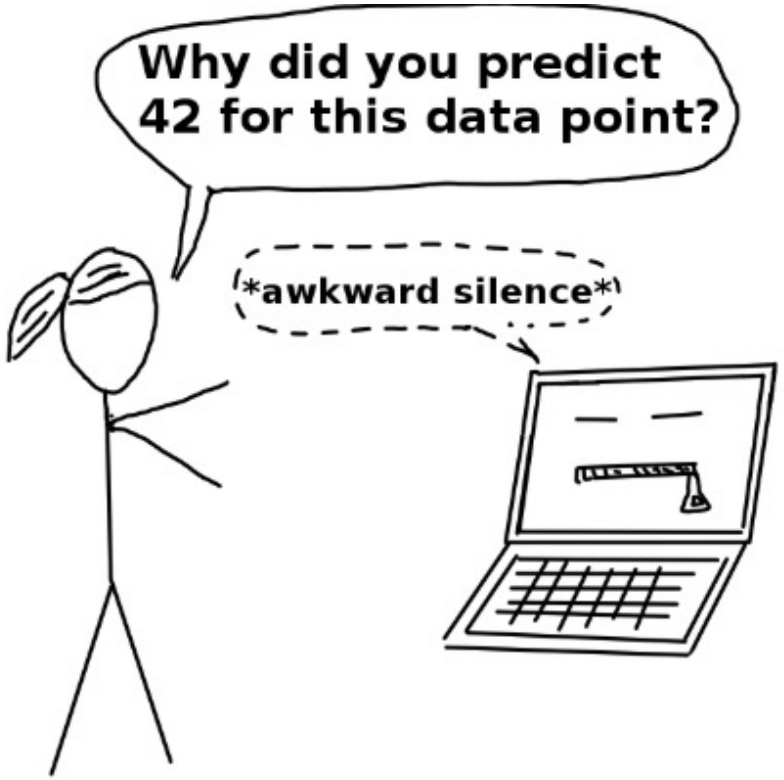
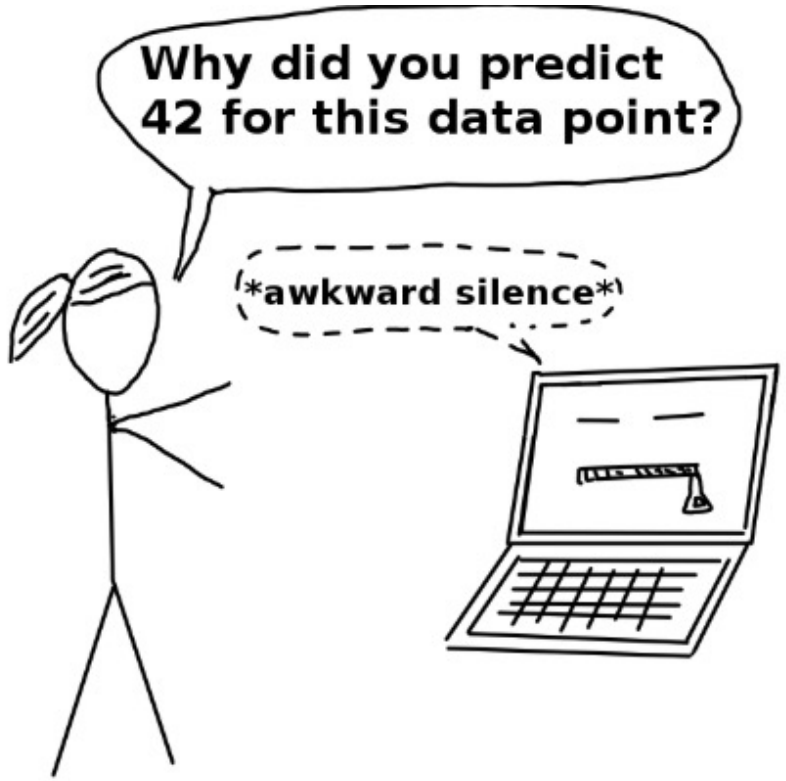


figure from: https://docs.paperspace.com/machine-learning/wiki/interpretability

自然语言处理中的可理解性

- 理解模型部件的功能属性
- 解释模型预测的行为
- 模型诊断

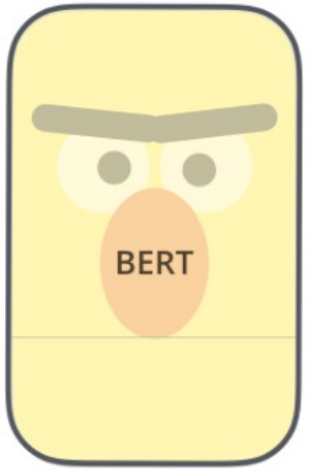


自然语言处理中的可理解性

- 理解模型部件的功能属性
- 解释模型预测的行为
- 模型诊断

例如：

- 预训练模型里存储了哪些知识？
- 基于Transformer的摘要系统和基于LSTM的摘要系统功能上有什么差异？



词法？句法？语义？

自然语言处理中的可理解性

- 理解模型部件的功能属性
- 解释模型预测的行为
- 模型诊断

例如：

- 情感分类任务中，模型预测其为正面情感的依据是什么？

我 喜欢 吃 苹果 。

evidence

自然语言处理中的可理解性

- 理解模型部件的功能属性
- 解释模型预测的行为
- 模型诊断

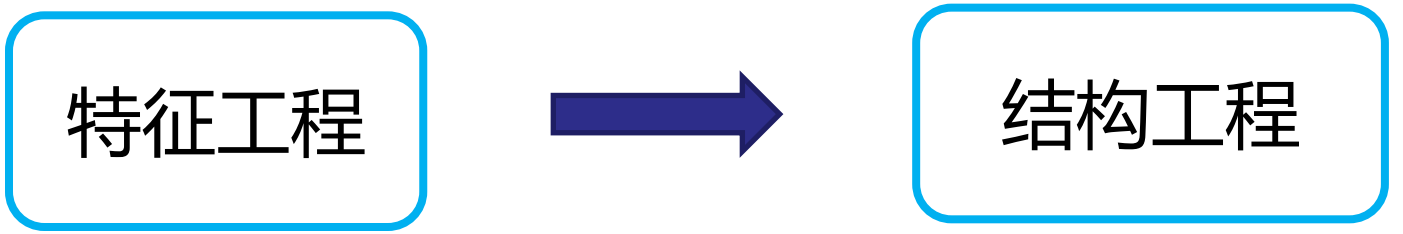
例如：

- 一个SOTA（State-of-the-art）系统足够鲁棒到可以处理不同类型的样本么？
- 不同模型的优缺点是什么？SOTA的瓶颈是什么？

为什么要做可理解分析?

- 理解模型部件的功能属性
  - 指导我们进行模型选择
- 解释模型预测的行为
- 模型诊断

By Neural Networks



如何选择适配特殊任务特殊数据集的结构？

为什么要做可理解分析?

- 理解模型部件的功能属性
- 解释模型预测的行为
  - 增加模型的可信赖性
  - 判断模型是否过拟合
- 模型诊断

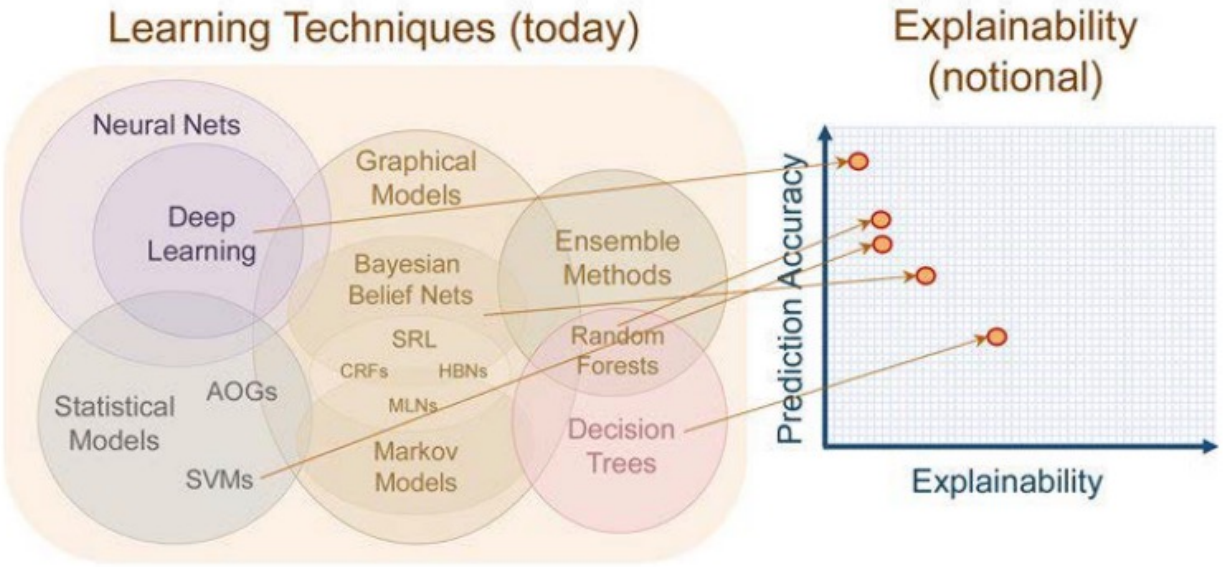
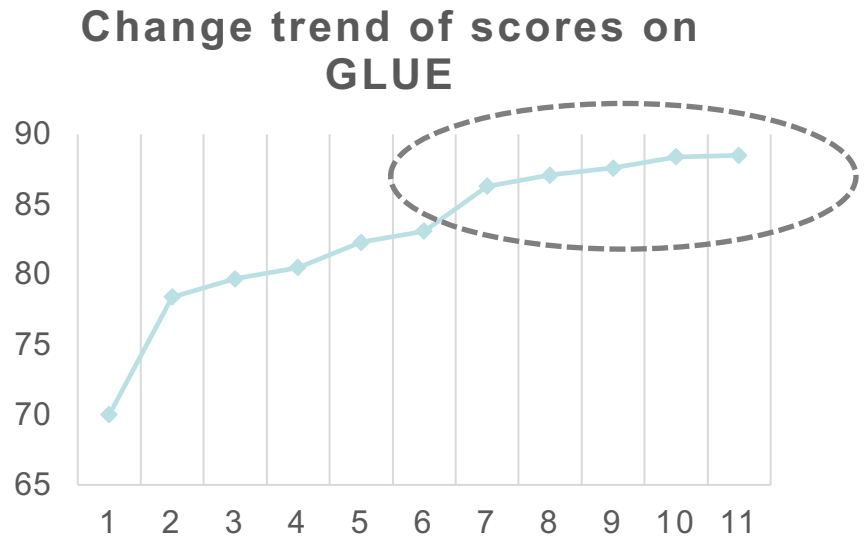
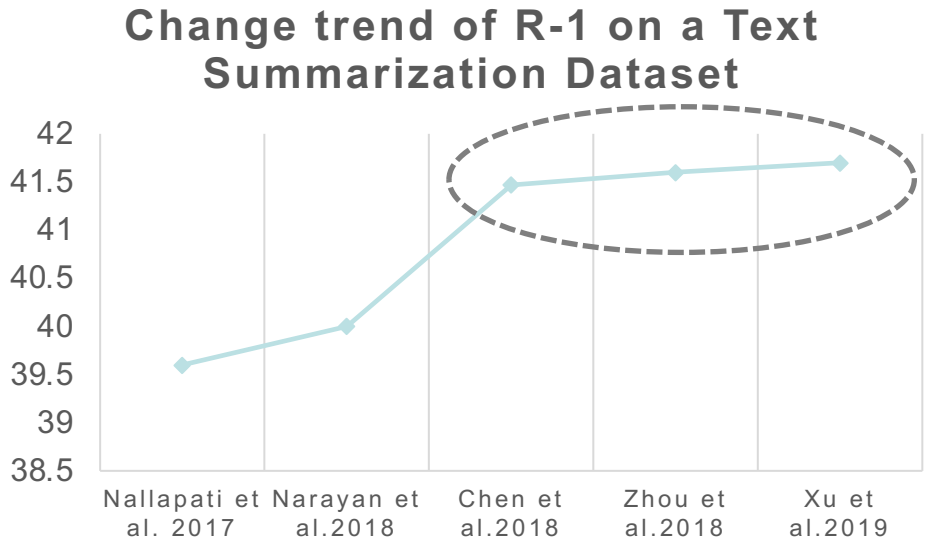


Figure from: https://arxiv.org/pdf/1911.12116.pdf

为什么要做可理解分析?

- 理解模型部件的功能属性
- 解释模型预测的行为
- 模型诊断
  - 指出模型不足，便于继续优化

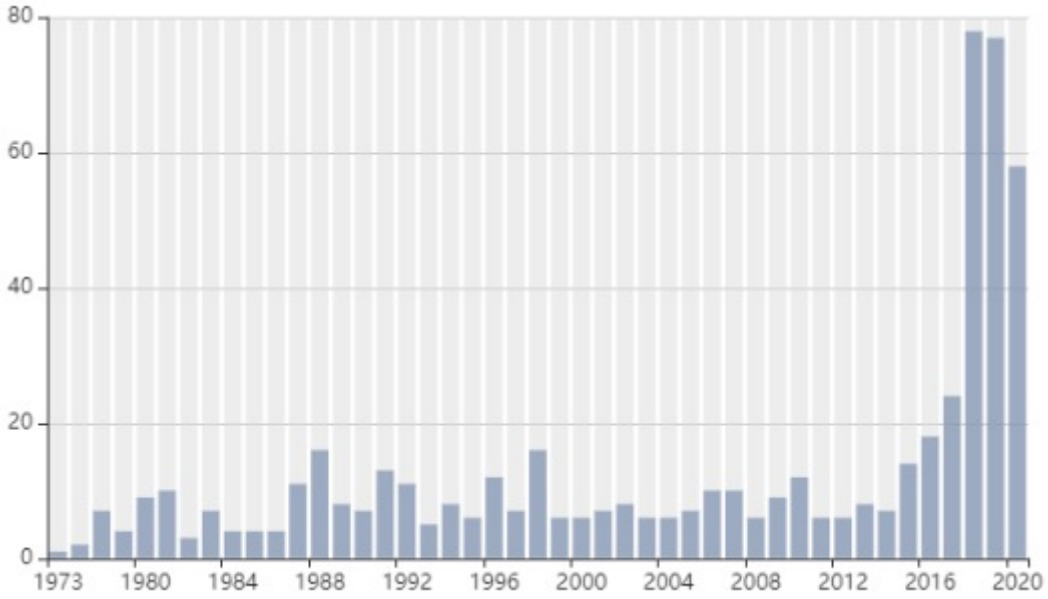




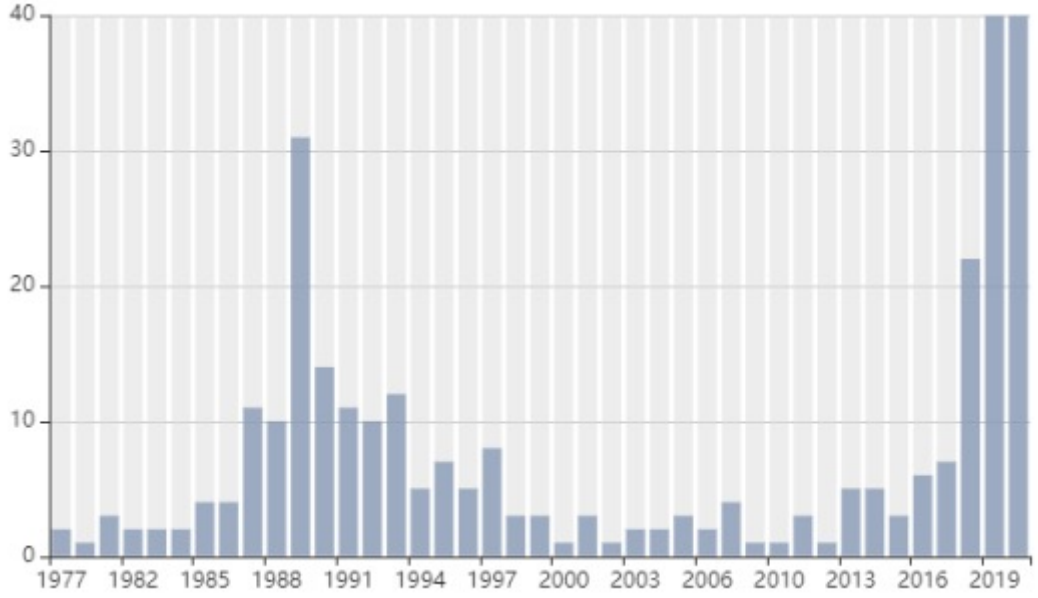
可理解分析的发展现状

简单的统计（论文来自于：*ACL, EMNLP, NAACL, NIPS, ICML, ICLR, AAAI* ）

“Interpretability”



“Explanation/Explainable”



研究现状：理解模型部件的功能属性

功能部件

Embedding Layer

RNN

Transformer

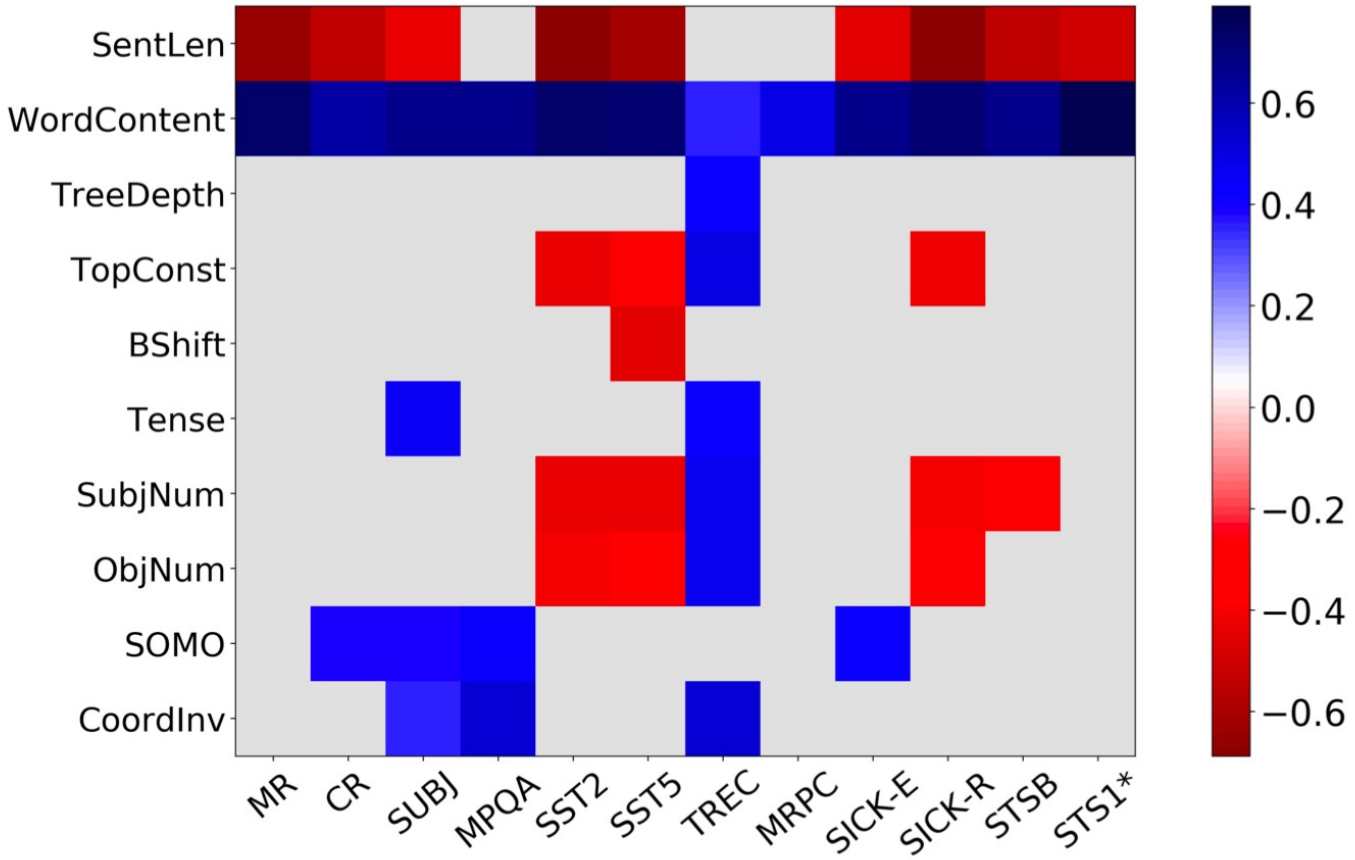
Attention

任务

LM

MT

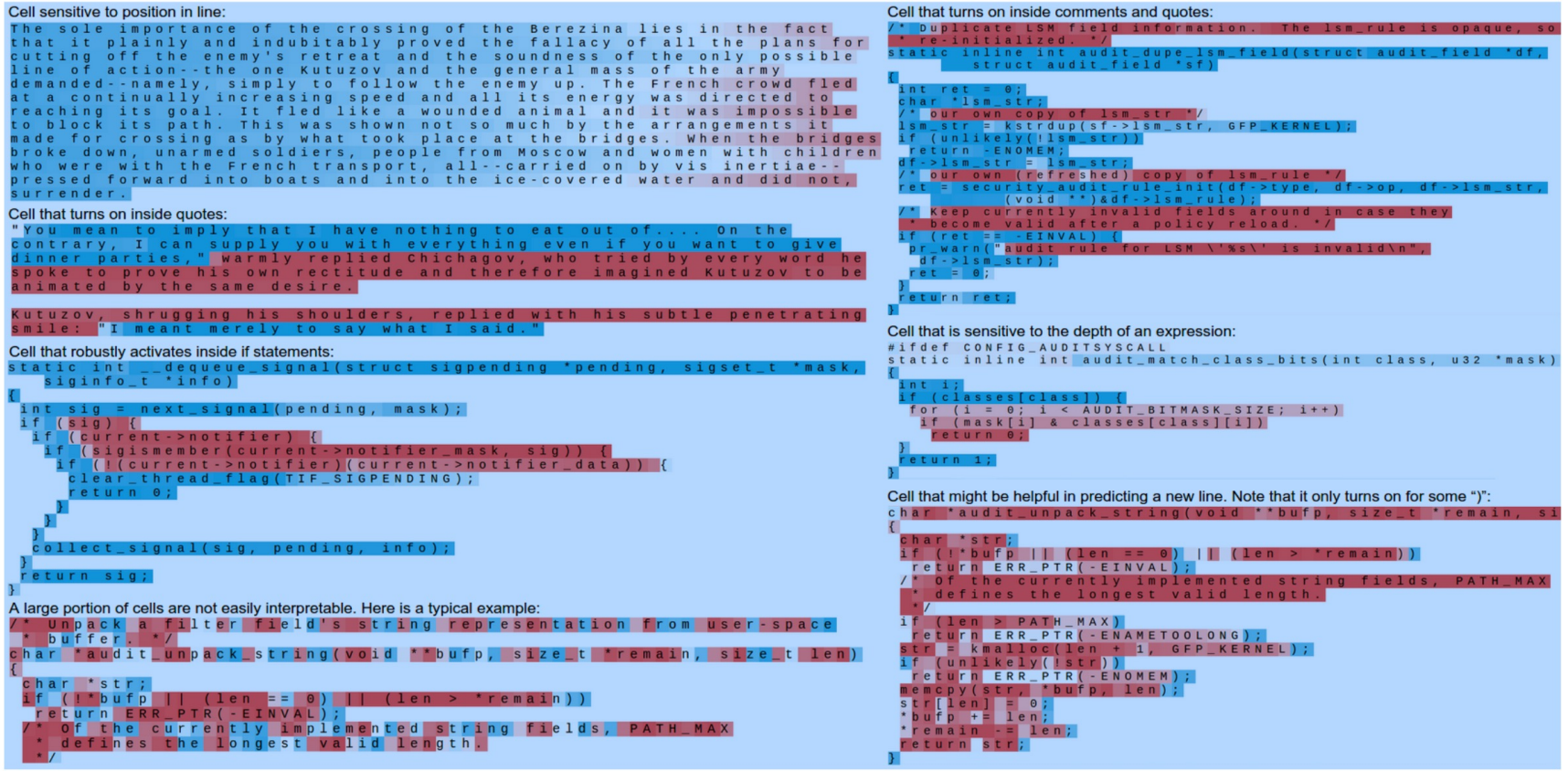
研究现状：理解模型部件的功能属性



What you can cram into a single $&#* vector: Probing sentence embeddings for linguistic properties (Alexis et al. 2018)

- 引入指针任务(Probing Task)的概念，用来“窥探”词向量所编码的知识
- 定义了10个指针任务，这些任务可以用来揭示词向量编码的不同类型的知识（词法，句法，语义）
- 后续很多工作都沿用此方法对学习到的表示 (representation) 进行解读

研究现状：理解模型部件的功能属性



Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
    siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/*  Unpack a filter field's string representation from user-space
 *  buffer.  */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   *  defines the longest valid length.
   */
```

Cell that turns on inside comments and quotes:

```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
           struct audit_field *sf)
{
  int ret = 0;
  char *lsm_str;
  /* our own copy of lsm_str */
  lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
  if (unlikely(!lsm_str))
    return -ENOMEM;
  df->lsm_str = lsm_str;
  /* our own (refreshed) copy of lsm_rule */
  ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
                 (void **)&df->lsm_rule);
  /* Keep currently invalid fields around in case they
   * become valid after a policy reload. */
  if (ret == -EINVAL) {
    pr_warn("audit rule for LSM \'%s\' is invalid\n",
      df->lsm_str);
    ret = 0;
  }
  return ret;
}
```

Cell that is sensitive to the depth of an expression:

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

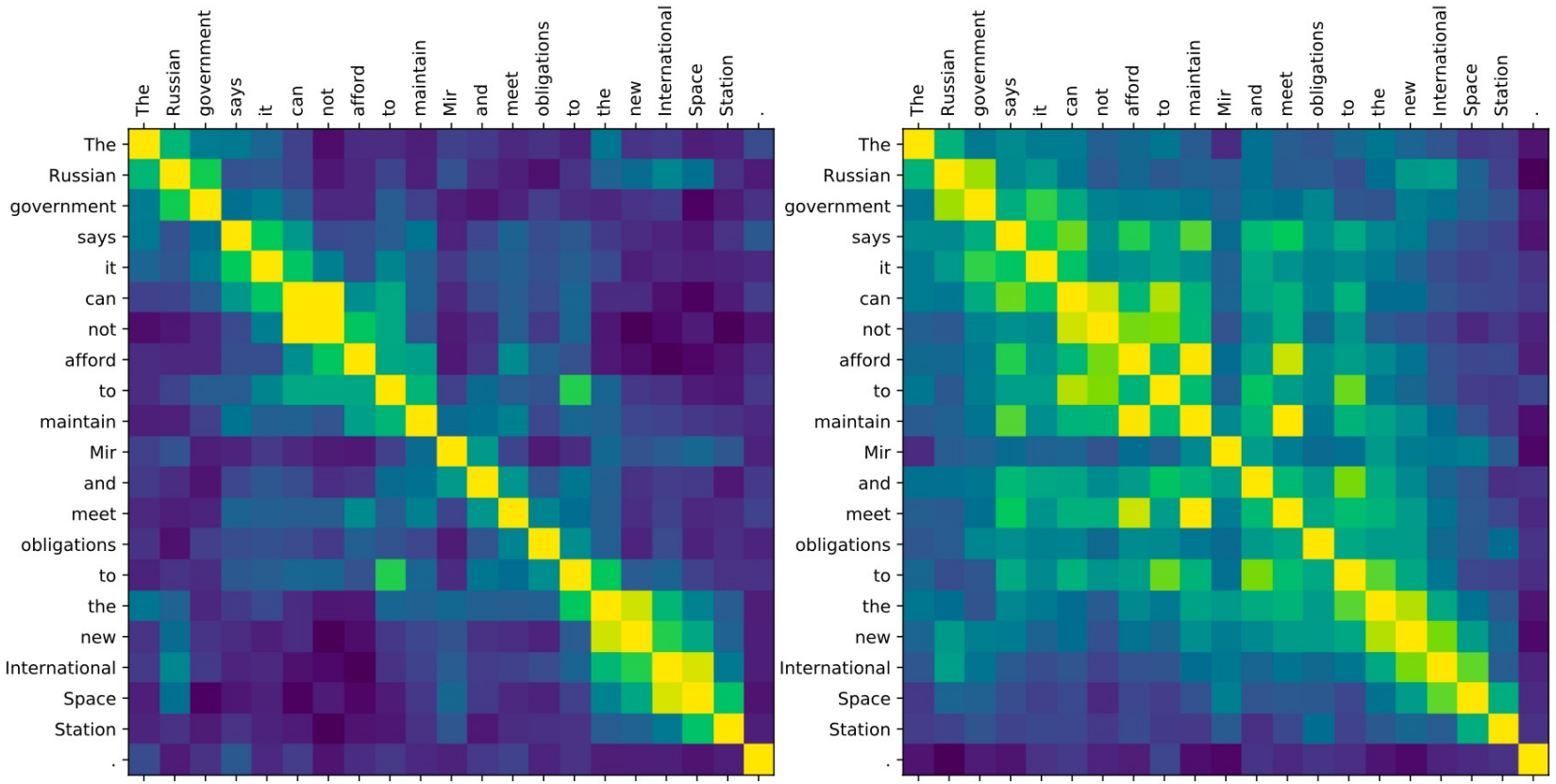
Cell that might be helpful in predicting a new line. Note that it only turns on for some “)”:

```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
  if (len > PATH_MAX)
    return ERR_PTR(-ENAMETOOLONG);
  str = kmalloc(len + 1, GFP_KERNEL);
  if (unlikely(!str))
    return ERR_PTR(-ENOMEM);
  memcpy(str, *bufp, len);
  str[len] = 0;
  *bufp += len;
  *remain -= len;
  return str;
}
```

Visualizing and Understanding Recurrent Networks (Karpathy et al 2016)

- 通过字符级别语言模型作为实验平台，探究了普通RNN，LSTM之间的细微差异，尤其是在捕捉长距离依赖时，神经元的行为差异

研究现状：理解模型部件的功能属性



g Contextual Word
ngs: Architecture and
ıtation (Peters et al.2018)

对比了LSTM, CNN,
mer作为基本部件训练语
行为的差异

研究现状：理解模型部件的功能属性

功能部件

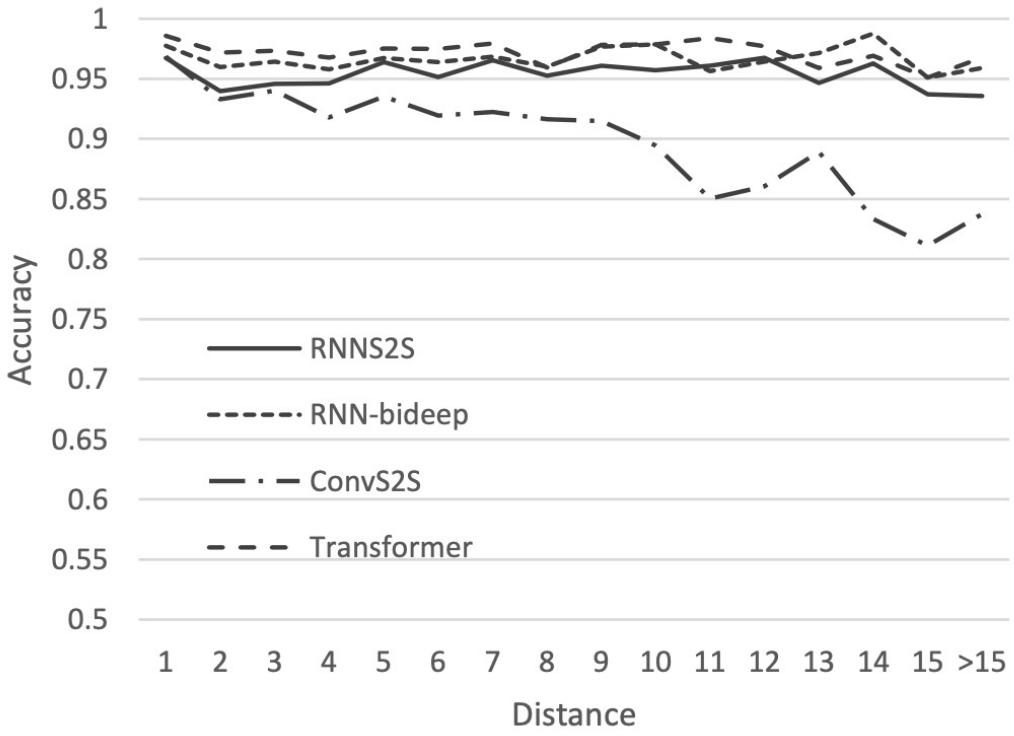
Embe
La

RN

Trans

Atte

任务



Why Self-Attention?
A Targeted Evaluation of Neural Machine Translation Architectures (Tang et al. 2018)

- 该文详细对比了CNN, RNN, Self-attention各自的擅长之处，并告诉后续研究者不同场景应该选择的部件

研究现状：理解模型部件的功能属性

Reference	Component	Property	Method
(Elloumi et al., 2018)	CNN activations	Style, accent, broadcast program	Classification
(Belinkov and Glass, 2017)	CNN/RNN activations	Phonetic units	Classification
(Dalvi et al., 2019a)	NMT and LM neurons	POS, morphology, lexical semantics	Classification
(Shi et al., 2016a)	NMT encoder neurons	Sentence length	Regression
(Belinkov et al., 2017b)	NMT states	POS, lexical semantics	Classification
(Belinkov et al., 2017a; Dalvi et al., 2017)	NMT states	POS, morphology	Classification
(Bisazza and Tump, 2018)	NMT states	Morphology	Classification
(Tran et al., 2018)	RNN / self-attention states	Subject-verb agreement	Likelihood comparison, direct classification
(Wang et al., 2017b)	RNN gates	Phoneme boundaries	Change in activation signal
(McCoy et al., 2018)	RNN sentence embedding	Hierarchical structure	Classification
(Blevins et al., 2018)	RNN states	POS, ancestor label prediction, dependency relation prediction	Classification
(Shi et al., 2016b)	RNN states	POS, top syntactic sequence, smallest constituent, tense, voice	Classification
(Gulordava et al., 2018)	RNN states	Number agreement	Likelihood comparison

Analysis Methods in Neural Language Processing: A Survey， https://arxiv.org/pdf/1812.08951.pdf

研究现状：解释模型预测的行为

- 经典方法
  - Gradient-based
    - IG score (Mukund et al.2017)
  - Attention
    - Attention is not explanation
    - Attention is not not explanation
  - Evidence Generation
    - e-snli (Camburu et al.2018)

研究现状：解释模型预测的行为

- 经典方法
  - Gradient-based
    - IG score (Mukund et al.2017)
  - Attention
    - Attention is not explanation
    - Attention is not not explanation
  - Evidence Generation
    - e-snli (Camburu et al.2018)

IG：将梯度回传到输入空间，通过观察梯度的绝对值反映输入特征的贡献

Did the Model Understand the Question? (Mudrakarta et al.2018)

- 在视觉问答中，即使模型回答对了问题，可是模型却把依据的重点弄错了，比如关注了不该关注的词
- 利用IG Score技术，去观看模型回答问题时，真正的依据是什么

Question: how symmetrical are the white bricks on either side of the building
Prediction: very
Ground truth: very

研究现状：解释模型预测的行为

- 经典方法
  - Gradient-based
    - IG score (Mukund et al.2017)
  - Attention
    - (1) Attention is not explanation
    - (2) Attention is not not explanation
  - Evidence Generation
    - e-snli (Camburu et al.2018)

通过观察Attention的权重值来判断输入空间不同特征的贡献

- 文章(1)说attention技术不可以作为解释模型决策的归因，因为attention的行为和决策的行为不完全一致

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore	*after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore*
original α	adversarial $\tilde{\alpha}$
$f(x\|\alpha,\theta)=0.01$	$f(x\|\tilde{\alpha},\theta)=0.01$

- 文章(2)回应文章(1)，认为attention是可以作为解释模型角色的归因