Fake news detection and fact verification

汇报人：刘豪

- 虚假新闻定义
- 虚假新闻检测和事实验证
- GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification
- 虚假新闻检测方法
- 讨论

虚假新闻

- 虚假新闻：有故意误导，欺骗读者的可证实为假的新闻文章。它是和公共新闻事件相关的信息。
- 相关概念：谣言，阴谋论。

虚假新闻检测

- 通过新闻项目的信息（文本，图片，传播，源）获取输入数据，对其进行真假分类。

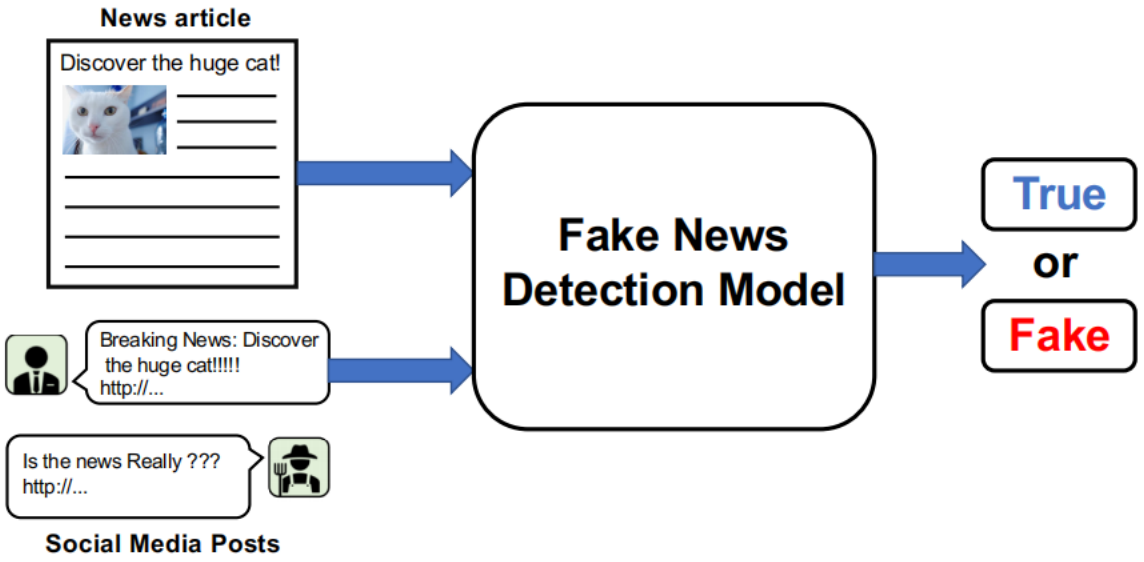


Fig. 1 **Fake news detection task** classifies whether each news item is fake, by getting from, as input

事实验证

- 通过明确的证据，判决一项声明是否是正确的。
- 事实验证一般验证一句话或几句话的可靠性。

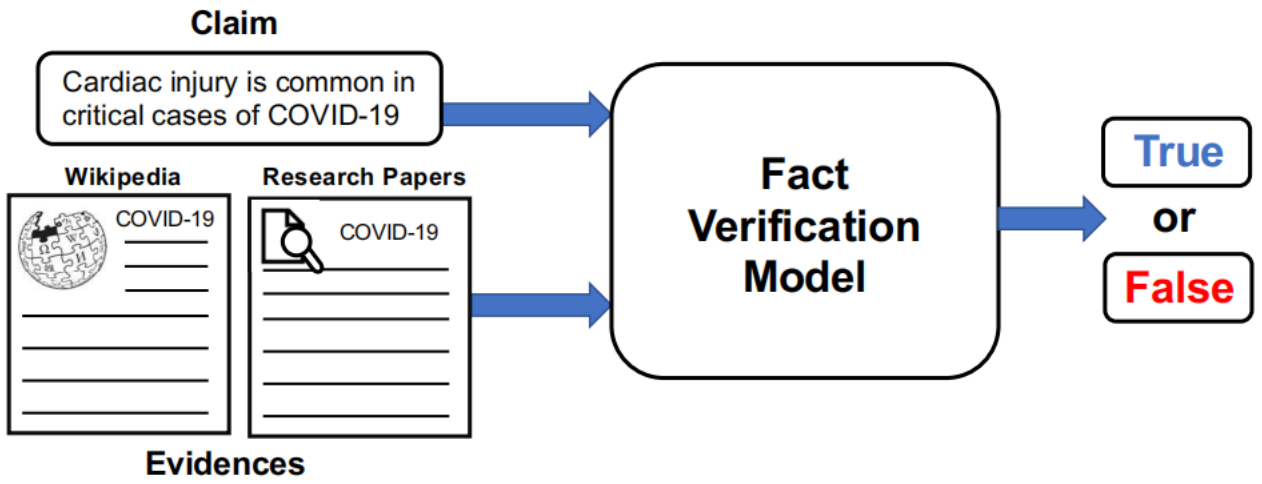


Fig. 2. **Fact verification task** is to make the decision as to whether a claim is correct, based on the explicitly-available evidence, such as Wikipedia articles and research papers.

虚假新闻检测方法

基于知识的假新闻检测

- **人工**：专家检测和众包。
  - 中国互联网联合辟谣平台；科学辟谣；腾讯较真等。
- **自动检测**：通常使事实核查的过程。
  - 对新闻进行摘要简化为一个事实核查任务，通过检索专家检测网站提取证据，验证新闻。
  - 要么直接构建一个知识库。

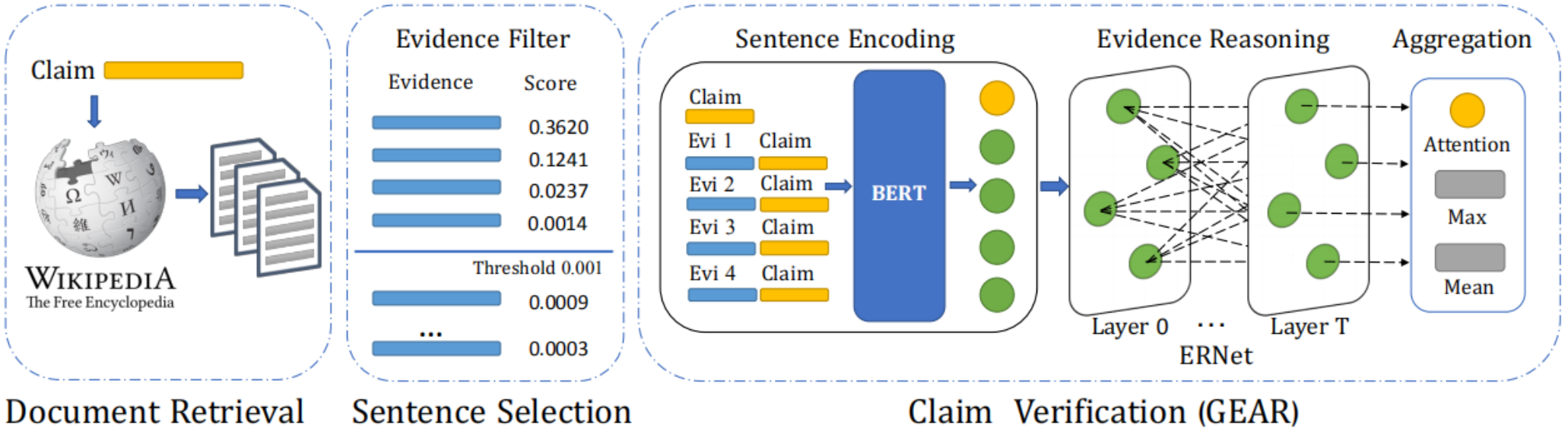
GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification

- 使用BERT、证据推理网络（ERNet）和证据聚合器来编码、传播和聚合来自多个证据的信息。
- 提出一个模型基于图的GEAR模型，可以对证据进行聚合推理，是得信息能够在完全连接的图上进行传输，然后利用不同的聚合器来收集多证据信息。

"SUPPORTED" Example	
Claim	The Rodney King riots took place in the most populous county in the USA.
Evidence	(1) The 1992 Los Angeles riots, ***also known as the Rodney King riots*** were a series of riots, lootings, arsons, and civil disturbances that ***occurred in Los Angeles County***, California in April and May 1992. (2) ***Los Angeles County***, officially the County of Los Angeles, ***is the most populous county in the USA***.

"REFUTED" Example	
Claim	Giada at Home was only available on DVD.
Evidence	(1) ***Giada at Home*** is a television show and first ***aired*** on October 18, 2008, ***on the Food Network***. (2) ***Food Network*** is an American ***basic cable and satellite television channel***.

事实验证中根据证据通常给验证的声明有三种标签"支持"、"反驳"、"信息不足"。现在问题是许多声明需要多个证据联合验证推理。



文档检索、证据选择、特征编码以及证据推理聚合四部分。

Document Retrieval and Sentence Selection

- 给定一个claim，利用来自AllenNLP工作中constituency parser从声明中提取潜在实体。然后，它使用这些实体作为搜索查询，并通过在线Mediawiki API找到相关的维基百科文档。存储每个查询中排名最高的7个结果以形成候选文章集。
- 最后，该方法删除不在离线维基百科转储中的文章，并根据文章标题和声明之间的重叠部分对文章进行过滤。句子选择组件从检索到的文档中的所有句子中选择与声明最相关的证据。

Claim Verification with GEAR

Sentence encoder

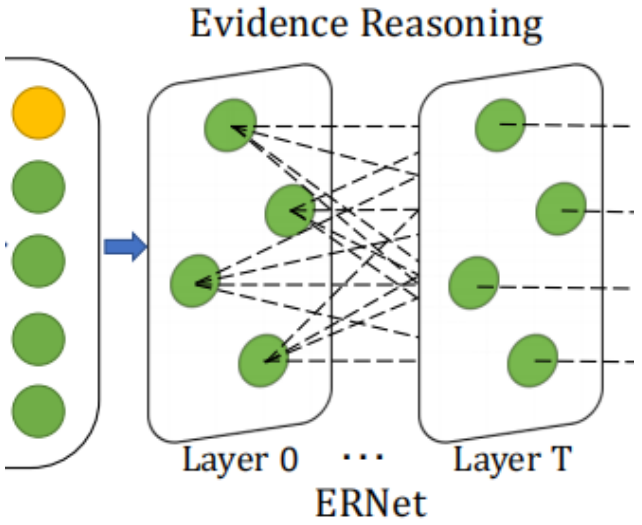
$$\mathbf{e}_i = \mathbf{BERT}\left(e_i, c\right),$$
$$\mathbf{c} = \mathbf{BERT}\left(c\right).$$

Evidence Reasoning Network

$$p_{ij} = \mathbf{W}_1^{t-1}(\mathbf{ReLU}(\mathbf{W}_0^{t-1}(\mathbf{h}_i^{t-1} \| \mathbf{h}_j^{t-1})))$$

$$\alpha_{ij} = \mathrm{softmax}_j(p_{ij}) = \frac{\exp(p_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(p_{ik})}.$$

$$\mathbf{h}_i^t = \sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{h}_j^{t-1}$$



Evidence aggregator

- Attention

$$p_j = \mathbf{W}_1'(\mathrm{ReLU}(\mathbf{W}_0'(\mathbf{c}\|\mathbf{h}_j^T))),$$
$$\alpha_j = \mathrm{softmax}(p_j) = \frac{\exp(p_j)}{\sum_{k=1}^{N}\exp(p_k)},$$
$$\mathbf{o} = \sum_{k=1}^{N}\alpha_k\mathbf{h}_k^T,$$

- Max
- Mean
- 最后一层预测

$$l = \mathrm{softmax}(\mathrm{ReLU}(\mathbf{W}\mathbf{o} + \mathbf{b})),$$

ERNet Layers	Aggregator		
	Attention	Max	Mean
0	66.17	65.36	65.03
1	67.13	66.63	66.76
2	**67.44**	**67.24**	**67.56**
3	66.53	66.72	66.89

Table 5: Label accuracy on the difficult dev set with different ERNet layers and evidence aggregators (%).

ERNet Layers	Aggregator		
	Attention	Max	Mean
0	77.12	76.95	76.30
1	77.74	77.66	77.62
2	**77.82**	**77.66**	**77.73**
3	77.70	77.55	77.60

Table 6: Label accuracy on the evidence-enhanced dev set with different ERNet layers and evidence aggregators (%).

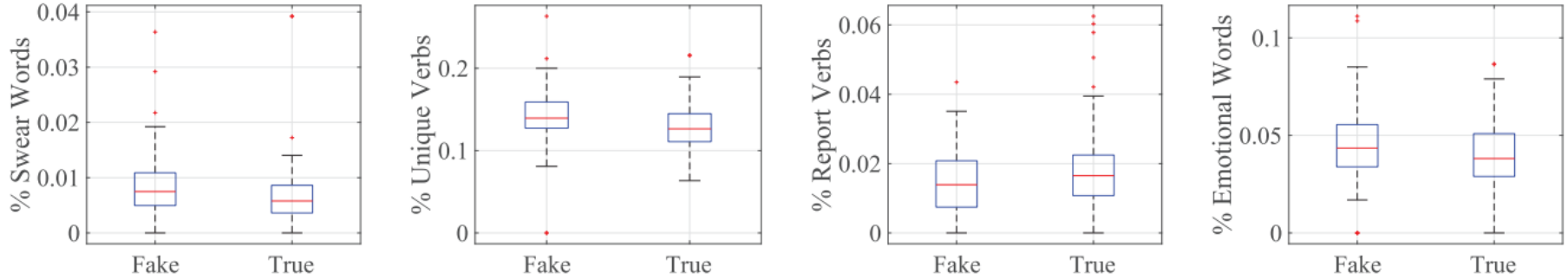
一个对BERT进行微调，一个没有微调。

文章总结

- 多证据联合验证claim的真伪
- 利用维基百科外部知识
- 专家检测的知识
- 使用到虚假新闻检测中

基于风格的假新闻检测

- 假新闻的文本更加不正式，有主观性，更加情绪化。那么就可以对行文客观性进行打分。



- 假新闻使用的图片具有吸引力，但和主题联系可能并不紧密。
- 一些特征融合技术检测其中图片部分和文本之间的差距。

基于社会背景的假新闻检测

- 一般来说，我们想代表的社交媒体环境有三个主要方面：新闻背景、用户、生成的帖子和网络。
- **新闻背景**：一般假新闻伴随一起公共新闻事件产生。
- **用户**：1）利用用户从相关帖子内容中提出的观点或潜在立场来推断新闻文章的准确性。2）判断恶意用户，主要是社交机器人。证据表明社交机器人会更早的传播虚假新闻。

生成的帖子：点赞、评论、其他用户对新闻的反应。（有证据表示虚假新闻的转发人和社交机器人概率差不多）

用户网络：通过对相关社交媒体帖子的相互关系来预测新闻的可信度。假设是一个新闻事件的可信度与相关社交媒体帖子的可信度高度相关。

新闻传播级联：虚假新闻相比于真实新闻传播更快，更广泛、更受欢迎，具有更高的病毒结构评分。

讨论

- 早期虚假新闻的检测：只能依靠有限的新闻内容以及社会背景知识。
- 和其他方向：社交机器人检测和观点动力学的联系。
- 研究的关注点：大面积预警早期虚假新闻还是类似与辟谣检测广为流传的虚假新闻。