给模型“提高知识水平”：通过知识库增强文本匹配任务

原创 韦国奥 SIAT-NLP AI论道 2020-03-06

什么是文本匹配 （Text Matching） 任务

文本匹配任务一直是自然语言处理领域的热门话题。简单的来说，文本匹配任务就是给定一个文本对 $input = (q_1, q_2)$，要求模型输出一个匹配分值 $score = f(q_1, q_2)$。文本匹配是信息检索（IR），自然语言推理（NLI），以及检索式问答（Answer Selection QA）的等任务的核心，因为抽象地来说，这些任务都要求一个文本对作为输入，输出可以是2维向量（表示二元标签的概率），亦或是一个相似度分值。

以往的文本匹配模型单纯基于词嵌入层的语义特征。模型将输入文本的 one-hot 向量通过词嵌入层转为词向量表示，然后通过 representation-based, interaction-based 或是 hybrid [1] 的隐层结构来捕捉文本对之间的交互信息。词嵌入层一般通过 Word2vec，fastText，GloVe 等预训练词向量模型初始化。

然而预训练词向量有自身的局限性。下图是 [2] 中给出的一个检索式问答的样本案例。单纯从词义相似度来考虑，以往的模型更有可能选择 Negative Answer，因为输入问题的关键词 **Pokemon** 在 Negative Answer 中出现了多次，导致其与问题的相似度更高，而它不知道宝可梦也与任天堂、田尻智强相关。词向量模型的假设是在上下文中相近的词拥有更相近的语义，大多使用通用文本语料库进行训练，因此它无法很好地处理命名体、低频词、专有名词以及特定领域的词间关系。这些问题在包含大量领域知识的 QA 任务中尤为凸显。

Question	When was ***Pokemon*** first started ?
Positive Answer	Is a media franchise published and owned by Japanese video game company ***Nintendo*** and created by ***Satoshi Tajiri*** in ***1996*** .
Negative Answer	The official logo of ***Pokemon*** for its international release ; "***Pokemon***" is short for the original Japanese title of "***Pocket Monsters***" .

检索式问答案例

为了解决上述问题，学者们都在探索新的方法，尝试将一些结构化知识和规则引入到自然语言模型当中。回到上面的例子，如果能给模型一个额外的输入，(Pokemon, produced_by, Nintendo)，模型就更有可能选择 Positive Answer。要获取这些额外的先验知识，近年发展势头迅猛的各种知识库项目是极佳的选择。接下来本文将介绍一些结合知识库与文本匹配的前沿论文。

知识增强的文本匹配模型

Neural Natural Language Inference Models Enhanced with External Knowledge (ACL 2018) [3]

这篇文章意图解决NLI任务，即判断两个句子前提 p和假设 h之间的逻辑关系。该方法使用了 WordNet 作为外部知识的来源。

> WordNet 是最著名的词典知识库。WordNet 主要定义了名词、动词、形容词和副词之间的语义关系。例如名词之间的上下位关系（如：“猫科动物”是“猫”的上位词），动词之间的蕴含关系（如：“打鼾”蕴含着“睡眠”）等。[4]

模型结构如下图所示。

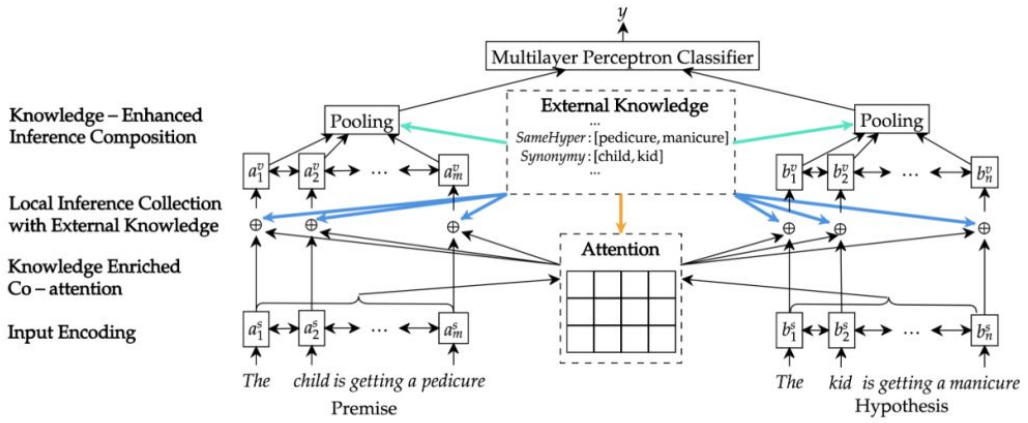


Figure 1: A high-level view of neural-network-based NLI models enriched with external knowledge in co-attention, local inference collection, and inference composition.

模型首先通过BiLSTM分别获得句子的上下文相关表示，

$$\boldsymbol{a}_i^s = \text{Encoder}(\mathbf{E}(\boldsymbol{a}), i)$$
$$\boldsymbol{b}_j^s = \text{Encoder}(\mathbf{E}(\boldsymbol{b}), j)$$

本文的核心在于，模型构造了一个相似度矩阵 (co-attention matrix)，并将先验知识嵌入到相似度矩阵中。矩阵的每个元素通过以下公式获得，

$$e_{ij} = (\boldsymbol{a}_i^s)^{\mathrm{T}} \boldsymbol{b}_j^s + F(\boldsymbol{r}_{ij})$$
$$F(\boldsymbol{r}_{ij}) = \lambda \mathbf{1}(\boldsymbol{r}_{ij})$$
$$\mathbf{1}(\boldsymbol{r}_{ij}) = \begin{cases} 1 & \text{if } \boldsymbol{r}_{ij} \text{ is not a zero vector} \\ 0 & \text{if } \boldsymbol{r}_{ij} \text{ is a zero vector} \end{cases}$$

其中$\boldsymbol{r}_{ij}$是一个5维的 “knowledge vector”。若$\boldsymbol{a}_i$与$\boldsymbol{b}_j$的词在 WordNet 中存在如下关系，该向量对应的元素则为1：近义词，反义词，上位词，下位词，同上位词（Co-hyponyms）。作者通过函数 F将 WordNet 中的知识转为了词间的相关性特征。直观上来说这篇文章的核心思想很容易理解，即有一定语义关系的词语应该有更高的相似性或逻辑相关性。

模型在 SNLI 数据集上取得了 SOTA 的效果，并且从实验分析上来看，在数据集越小时，越多的外部知识能带来更高的提升效果。

Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval (ACL 2018) [5]

本文采用了 CN-DBpedia 构建的知识图谱（Knowledge Graph，KG）作为外部知识的来源。CN-DBpedia是一个由百度百科、互动百科、中文维基百科构建的大型中文知识库，其格式为（subject, relation, object）形式的三元组，其中包含 10,341,196 个实体 以及 88,454,264 种关系。

文中提出的模型 EDRM 将 IR 任务中的文本对通过词+实体的特征来表示。对于一段文本中出现的实体，作者构造三种特征来进行表示：Entity Embedding，Description Embedding，Type Embedding来获得Enriched-entity Embedding。其中Entity Embedding直接通过嵌入层表示，

$$\vec{v}_e^{\text{emb}} = \text{Emb}_e(e)$$

Description Embedding 通过将实体的文本描述输入 CNN 和 Max-polling 层获得，

$$\vec{g}_e^j = \text{ReLu}\left(W_{\text{CNN}} \cdot \vec{V}_w^{j:j+h} + \vec{b}_{\text{CNN}}\right)$$

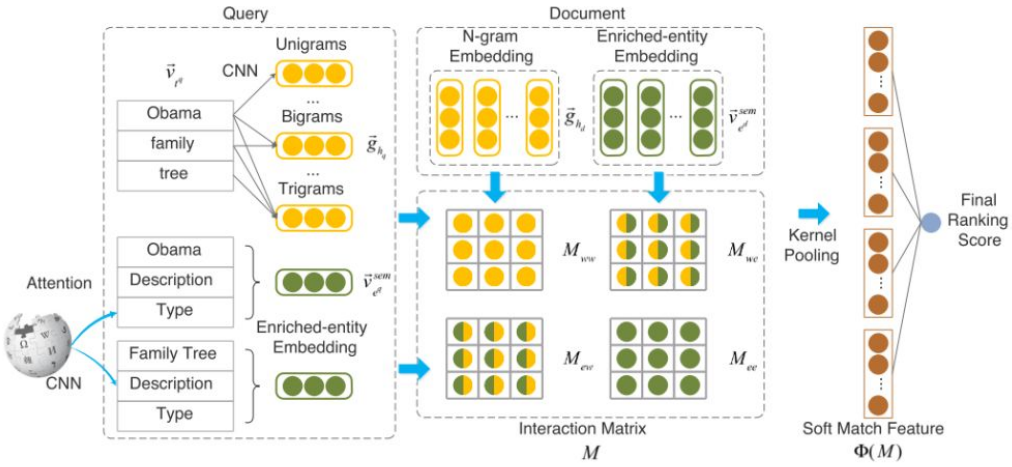
Type Embedding 通过嵌入层加上一个 Attention 操作获得，

$$\vec{v}_{f_j}^{\text{emb}} = \text{Emb}_{\text{tp}}(e)$$
$$a_j = \frac{\exp(P_j)}{\sum_l^n \exp(P_l)}$$
$$P_j = (\sum_i W_{\text{bow}}\, \vec{v}_{t_i}) \cdot \vec{v}_{f_j}$$

其中 f_j为实体的类型集合， $\vec{v}_t$为文本的词向量。最终三种特征向量通过如下方式结合，获得了Enriched-entity Embedding，

$$\vec{v}_e^{\text{sem}} = \vec{v}_e^{\text{emb}} + W_e\left(\vec{v}_e^{\text{des}} \oplus \vec{v}_e^{\text{type}}\right)^T + \vec{b}_e$$

模型之后采用了和 K-NRM [6] Conv-KNRM [7] 相同的处理方式，创新的地方是其分别对文本对的文本特征和实体特征求出了4个相似度矩阵，最后将所有结果拼接获得分值。最终模型在 Sougo Log 数据集上相比 K-NRM 和 Conv-KNRM 在 NDCG 以及 MRR 上获得了最高 28.42% 的提升。



EDRM

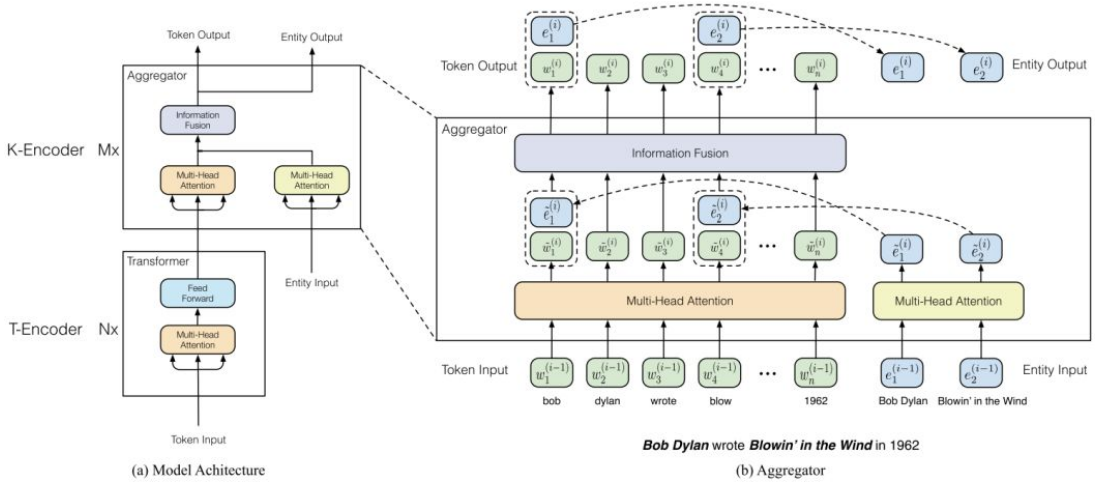
EDRM 与以往的文本匹配模型在特征的处理上思路类似，但是给文本增加了基于知识图谱的实体特征，使其在计算时能获得更丰富的语义信息。

ERNIE: Enhanced Language Representation with Informative Entities (ACL 2019) [8]

2018 年，Transformer 以及 BERT 横空出世，开始霸占各大 NLP 任务榜单。其优秀的表现也让其成为了众多前沿研究的 baseline。BERT本身是一种基于大语料库的无监督预训练模型，为了让其学习到更多结构化的知识，许多前沿工作开始研究如何把外部知识注入到 BERT 当中。

由清华大学发布的 ERNIE-thu 使用维基百科语料库进行预训练，并使用语料中的 Anchor Link 来获取实体，通过 Wikidata 训练出的 TransE [9] 向量作为实体的特征。TransE 是知识图谱嵌入（Knowledge Graph Embedding）方法的一种，和词嵌入类似，这些方法的核心是把知识图谱中的所有实体和关系映射到连续的向量空间当中，由此让机器能够更好地理解知识图谱的结构化信息。

在 BERT-base 的基础上，ERNIE-thu 保留了 6 层的 Encoder，但是将原本的 Decoder 改成了文中提出的 K-Encoder 来学习知识图谱中的实体信息，如下图所示



ERNIE

K-Encoder 输入的先经过如下变换，

$$\left\{\tilde{\boldsymbol{w}}_1^{(i)}, \ldots, \tilde{\boldsymbol{w}}_n^{(i)}\right\} = \text{MH-ATT}\left(\left\{\boldsymbol{w}_1^{(i-1)}, \ldots, \boldsymbol{w}_n^{(i-1)}\right\}\right)$$
$$\left\{\tilde{\boldsymbol{e}}_1^{(i)}, \ldots, \tilde{\boldsymbol{e}}_m^{(i)}\right\} = \text{MH-ATT}\left(\left\{\boldsymbol{e}_1^{(i-1)}, \ldots, \boldsymbol{e}_m^{(i-1)}\right\}\right)$$

文本的 Token Embedding 以及 Entity Embedding 经过 Multi-head Attention 之后，输入如下的 Infomation Fusion Layer，

$$\boldsymbol{h}_j = \sigma\left(\tilde{\boldsymbol{W}}_t^{(i)}\tilde{\boldsymbol{w}}_j^{(i)} + \tilde{\boldsymbol{W}}_e^{(i)}\tilde{\boldsymbol{e}}_k^{(i)} + \tilde{\boldsymbol{b}}^{(i)}\right)$$
$$\boldsymbol{w}_j^{(i)} = \sigma\left(\boldsymbol{W}_t^{(i)}\boldsymbol{h}_j + \boldsymbol{b}_t^{(i)}\right)$$
$$\boldsymbol{e}_k^{(i)} = \sigma\left(\boldsymbol{W}_e^{(i)}\boldsymbol{h}_j + \boldsymbol{b}_e^{(i)}\right)$$

其中 $\boldsymbol{h}_j$为融合了词特征和实体特征的隐层向量。这个隐层向量在分别通过两个 FFN 输出新的 Token Embedding 和 Entity Embedding。如果一个 Token 没有对应的实体，上述的公式便省去和 $\tilde{\boldsymbol{e}}_k^{(i)}$有关的计算项。需要注意的是计算中一个实体仅与对应词的第一个 Token 进行知识融合。实验表明 ERNIE-thu 在 Entity Typing 和 Relation Classification 任务上获得了较好的提升。

未来展望

在自然语言处理中引入知识库，其中一个目的是为了通过人类已有的知识对深度学习进行指导。清华大学教授刘知远曾经说过，

> 大多数深度学习模型都和词嵌入向量一样缺少可解释性，这也是深度学习被广泛诟病的地方。然而对于自然语言处理而言，我们不仅希望模型能够理解或生成文本，更希望知道模型这样做的原因。[10]

目前的许多深度学习模型虽然通过外部知识获得了很好的效果提升，但是仍未提高我们对模型决策的理解，并且没有较好地利用知识库中的结构化信息。

如何更好的处理知识库的结构化数据，是未来科学界面临的重要问题之一。从另一个角度来看，大多数知识库的结构可以转化为知识图谱，而今年兴起的图神经网络（Graph Neural Networks, GNN）在处理图结构信息的任务上发挥出了强大的威力。事实上，已经有很多研究方向专注于通过知识图谱与 GNN 解决 NLP 任务。希望未来在这些方向上也能出现像 Word2vec 和 BERT 一样对 NLP 领域产生革命性影响的成果。

引用文献

[1] Guo, Jiafeng, et al. "A deep look into neural ranking models for information retrieval." Information Processing & Management (2019): 102067.

[2] Shen, Ying, et al. "Knowledge-aware attentive neural network for ranking question answer pairs." The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018.

[3] Chen, Qian, et al. "Neural natural language inference models enhanced with external knowledge." arXiv preprint arXiv:1711.04289 (2017).

[4] 科普 | 典型的知识库/链接数据/知识图谱项目 - Laubass的文章 - 知乎 https://zhuanlan.zhihu.com/p/41118663

[5] Liu, Zhenghao, et al. "Entity-Duet Neural Ranking: Understanding the Role of Knowledge Graph Semantics in Neural Information Retrieval." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[6] Xiong, Chenyan, et al. "End-to-end neural ad-hoc ranking with kernel pooling." Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval. 2017.

[7] Dai, Zhuyun, et al. "Convolutional neural networks for soft-matching n-grams in ad-hoc search." Proceedings of the eleventh ACM international conference on web search and data mining. 2018.

[8] Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
