第六届“泰迪杯”
数据挖掘挑战赛

优
秀
作
品

作品名称：基于知识与语义的深度文本匹配模型-知语

荣获奖项：特等奖

作品单位：山东大学

作品成员：张晓 杜存宵 刘东岳

指导老师：聂礼强

基于知识与语义的深度文本匹配模型-知语

Abstract

在大数据时代，智能阅读系统的需求日益增长，人们需要一款智能阅读软件来方便自己的生活，如电子书阅读，说明书阅读，都可以借由该技术简化。另一方面，随着数据的不断增多，以及深度学习在自然语言处理领域的发展，智能阅读模型也有了发展的基础，目前来看，智能阅读模型正处在方兴未艾的发展中。

对于本次赛题给出的智能阅读模型系统，我们重新定义为文本匹配问题，也即给定两段文本，模型要正确判断文本是否匹配。

对于赛题复杂的要求，多变的环境句式，我们采用了以深度学习为基础的，基于知识与语义双重视角匹配的模型。我们首先构建了知识库，然后进行了知识检索，在这里使用了 jieba 分词以及用 Lucene 将知识库构建索引，并且使用 BM25 的匹配方式进行计算；然后我们使用 Embedding 映射后，通过 GRU 进行提取时序特征信息以及使用注意力机制对不同的特征进行加权；最后我们使用了 CNN 进行特征提取以及用 softmax 进行归一化处理。我们的模型在构造的线下测试集上得到了有利的验证，也说明了模型的鲁棒性，实用性。

关键词: 知识库、知识检索、知识过滤、GRU、注意力机制、CNN

In the era of big data, the demand of intelligence-read-system is growing up. We need this system to help our lives w.r.t reading e-book, instruction reading, etc. On the other hand, with the big data to training deep learning models, and the trend of deep learning in Natural Language Processing, this system gains much improvement.

For the intelligent reading model system given in this contest, we redefine the problem of text matching. Given two sections of text, the model should correctly judge whether the text is matched.

For the complex requirements of the competition and the changeable environmental sentence pattern, we have adopted a model based on deep learning and based on the dual visual angle of knowledge and semantics. We first built the knowledge base, then carried out the knowledge retrieval, used the Jieba participle, and built the index with the Lucene knowledge base, and used the BM25 matching method to calculate; then we used the Embedding mapping, and we extracted the temporal feature information by GRU, and weighted the difference by using the attention mechanism. Finally, we use CNN to extract feature and normalize it with softmax.Our model has been verified in the constructed offline test set, and it also shows the robustness and practicability of the model.

Key words: knowledge base, knowledge retrieval, knowledge filtering, GRU, attention mechanism, CNN.

1. 引言

在大数据时代，智能阅读系统的需求日益增长，人们需要一款智能阅读软件来方便自己的生活，如电子书阅读，说明书阅读，都可以借由该技术简化。另一方面，随着数据的不断增多，以及深度学习在自然语言处理领域的发展，智能阅读模型也有了发展的基础，目前来看，智能阅读模型正处在方兴未艾的发展中。

对于本次赛题给出的智能阅读模型系统，我们重新定义为文本匹配问题，也即给定两段文本，模型要正确判断文本是否匹配。

目前主流的文本匹配模型有很多，大致可以分为两个主流。一方面是基于传统统计学特征的方法，比如

$tf-idf$ [14]，比如 $BM25$ [16]这种语义匹配的计算，这些传统统计学的方法依赖于人工构建的特征，往往特征是非常难以构建而且无法迁移的。另一大类的方法是基于深度学习的方法，比如基于卷积神经网络的匹配模型，基于循环神经网络 (长短时记忆神经网络) 的匹配模型，基于特征交互的匹配模型等等。这些模型可以很方便的在语义，也即文本的内容是否匹配上做出判断。由于目前计算资源以及可用数据的增多，再加之专家系统构建的困难程度，研究者们已经开始逐步转向了基于深度学习的匹配方法。

在看到主流方法蓬勃发展的同时，我们也要指出这些方法都不约而同的存在一些问题。无论是基于统计学特征的方法，还是深度学习的表示学习方法，都无法有效融合外部知识，也无法做出与知识有关的判断。如文本 1: “外星球来的孩子作者是谁?” 与文本 2: “外星球来的小女孩, 作者: 金涛著” 就很容易被判别为正确匹配，然而只要了解过这本《外星球来的孩子》的书籍的人，都应该知道它的作者不是金涛而是杨红樱，由于这些主流方法没有考虑额外的知识信息，另外文本 2 确实在内容上是可以认为在回复文本 1，只不过是事实错误，因此如何解决这种语义匹配而却与事实相矛盾的问题成为了当前智能阅读系统的关键。

在语义匹配的基础上融合知识并不是一件平凡的事情，它具有以下几个挑战:

- 如何根据文本检索到正确的知识?
- 如何将检索到的知识融入到深度学习模型中来?
- 如何既能兼顾知识也兼顾语义的匹配?

为了解决以上的这些挑战及难点，我们设计了我们的模型-基于**知**识与**语**义匹配的深度学习文本匹配模型，简称为**知语模型**。知语模型大致有三个阶段，在阶段 I，我们会根据二元组中的问题检索出相应的 10 个最相关的知识，在阶段 II，我们将利用注意力机制 [6]根据问题去选择知识三元组中问题所关注的属性，在最后一个阶段，我们将问题，回答，以及检索得到的知识作为一个三元组输入进入深度匹配模型，根据长短时记忆神经网络给出最后的答案。

综上所述，我们提出了基于知识与语义匹配的深度文本匹配模型，可以有效建模知识，语义，等多个角度的匹配程度，在我们自己构造的线下数据集上取得了远远高于 baseline 的成绩，这也说明了知语模型的有效性。

总结来看，我们主要有以下三个贡献:

- 我们首次提出了将知识融入到深度学习的模型，增强了模型的表示能力。
- 我们的模型可以在多个角度 (知识，语义) 方面同时判别文本是否匹配。
- 我们的模型可以在泰迪杯数据上取得良好的成绩，证明了模型的实用性。

2. 相关工作

我们的工作主要与文本匹配相关联，在这里我们将介绍文本匹配的主流方法。大体来说，现今文本匹配的思想主要分为两种，一种是基于统计全局特征的传统方法，另一种是基于深度学习等的表示学习方法。

首先我们介绍统计全局特征的方法，基于统计全局特征的方法一般使用一些规则提取特征，之后使用例如支持向量机 [9]，朴素贝叶斯 [15]，逻辑回归 [18]等方法进行判别。最早的统计全局特征方法是由 Karen Spark Jones 提出的 tf-idf，它不需要复杂的分类器进行判别，而且有较深的数学基础作为支撑。之后 Hinton 提出了词向量的概念，将词的表示方法进行了丰富。除此之外，也有一些学者提出了 n-grams [2]来丰富词与词的关系作为新的特征，也有另外一些学者研究主题模型希望得到更好的表示，Blei 提出了如 LDA 主题模型 [21]、S.T.Dumais 等提出了 LSI/PLSI 概率潜在语义索引等方法。然而这一大类方法严重依赖特征提取，需要大量的专家设计良好的特征，可迁移性非常差，难以广泛使用。

另一种主流方法是表示学习，它利用神经网络与大规模的数据来对文本进行表征学习，不需要人工提取特征。在这一研究领域，Hinton 是最早提出神经语言模型的人 [12]，在他之后，Bengio 跟进了他的研究，他对如何进行词嵌入提出了良好而有力的模型。2014 年，谷歌的研究人员提出了新的算法 word2Vec [11]，后来有人提出了 glove 方法 [13]获得更加精准的表征。于此同时，对于句子的表征学习也在蓬勃发展，同年，耶鲁大学的 kim 提出了 TextCNN [8]算法，使用卷积神经网络 (CNN) 来对文本进行建模，希望使用卷积来提取类似于 n-grams 的文本特征，首次在公有数据集上超过了传统方法的表征。之后也有人使用了循环神经网

络 [22]来进行实现，由于循环神经网络可以建模时序，所以在表征句子这种序列上，取得了不错的成绩。近年来随着自注意力机制的提出，卡内基梅隆大学的人员关注如何使用自注意力机制表征文档，提出了 HAN（层次注意力网络）[20]算法来建模更加复杂的文档，他们使用词，句子，段落之间三种的注意力来表示一篇文档，由于自注意力机制可以良好的捕捉更加长远的时序关系，所以他们在 yelp 等公开的数据集上取得了非常好的表现。于此同时，中国中科院的研究人员也提出使用建模图片的方式来建模文本，他们将自己的模型命名为 RCNN [7]。近年来随着斯坦福大学构建了公有的自然语言推理数据集-SNLI，文本匹配的算法更加丰富。在对 SNLI 的研究中，陈亮等人提出了 Re-read LSTM [17]来对一个句子进行多层次的建模; 宫一尘等人提出了 DIIN 模型来对文本之间的信息进行大规模的交互，取得了在该数据集榜首第一名的成绩; 王志国也提出了 BIMPM 在包括 SNLI 数据集在内的多个数据集同时取得了不俗的表现。第二种方法目前需要的手工特征非常少，所以已经成为了研究界乃至工业界的主流，然而第二种方法的问题在于难以进行超参数的选取，以及大量的计算资源及标注数据的获得。不过由于 GPU 并行运算的提高以及日新月异的诸如 MXNet [3]，tensorflow [1]等深度学习框架的提出，再加之如数据堂[1]等数据众包公司的出现，这一困难已经不再沉重。

然而我们同时也要看到无论是第一种方法还是第二种方法都只能藉由文本内部的特征进行表示，如果说判断两段文本是否匹配需要用到外部的知识，那么以上提出的模型都会失效。比如判断“枣庄市在中国哪个省”与“河北省”这两段文本来说，在缺乏外部知识的给出下是很难得到正确的答案的，因为二者确实在语义上，内容上是连贯的，只不过是事实错误。因此我们的模型-知语，考虑了外部知识对于神经网络的加强，这也正式我们提出的算法与之上所介绍方法的区别。

2.1. 任务重定义

官方赛题的要求是：聚焦于智能交互在电子书阅读的应用，日常生活中人们要阅读大量的 txt 文本，其内容可能是小说、教程、文集、词典等。很多情况下我们只是需要从文本中查找某一些片段来解决我们的问题。比如，通过查找法律文献中的一些段落来解决我们

[1]http://www.datatang.com/.

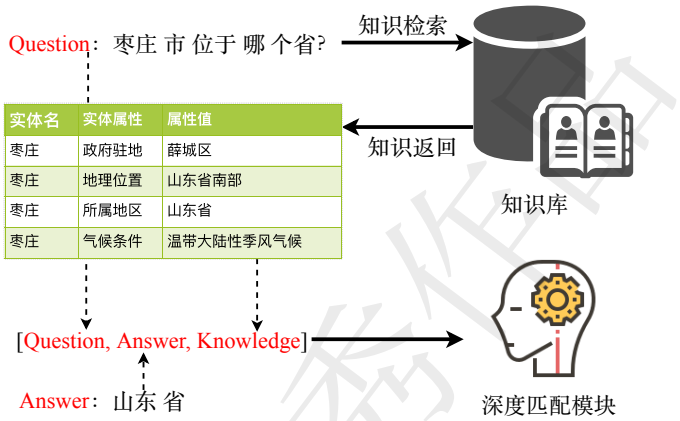

图 1. 对于给定的二元组 (question, answer)，首先使用 question 在知识库中检索出与该二元组相关的知识，将新构成的 (quesiton, answer, knowledge) 三元组作为深度匹配模块的输入。

的法律疑惑，这时并不需要精读整个法律文献；对于小说，有时候我们也只是想知道其中一些特殊细节，并不想花时间去通读整个小说。所以我们需要在给出问题时，将与问题相匹配的文档中的内容返回，或者返回相应答案的所在行。

我们将这个问题看作是一个文本匹配的问题，给出一个 [问题, 回复] 二元组，模型要正确的判别问题与回复是否匹配。下面给出形式化定义，我们需要学习一个匹配函数 $F(\cdot)$，能够正确的判断给定问题 Q 以及回复 A 是否匹配。

因为我们认为该任务需要考虑额外的知识表征，所以我们将上面的任务进行重定义，即给出 Q 和 A 的情况下，允许使用第三方的知识库进行额外的知识检索来帮助模型学习，所以匹配函数 $F(\cdot)$ 的输入是一个三元组 [问题，知识，回复]，我们使用 K 来表示知识，最后模型给出的分数 y 即为:

$$y = F(y|[Q, K, A]) \tag{1}$$

，其中 y 的取值区间为 [0,1]。

2.2. 流程概览

为了学习这个函数，我们构建了我们的模型-知语，它共有三个阶段，首先会根据 Q 在知识库中检索出相关的 m 条知识，由于知识库非常巨大，所以在此检索阶段我们并未使用深度学习。在第二阶段，我们会让问题与知识进行交互，利用注意力机制，m 条知识分别

赋予不同的权重，以便模型更加有效的利用有利的知识而舍弃那些无用的知识。在第三个阶段，我们使用循环神经网络以及卷积神经网络，来对输入的加权知识，问题，回复三元组进行有效的建模，融合，判别。

2.3. 阶段一：知识检索

我们将第一个阶段称为知识检索，此阶段的流程如图1所示。我们首先使用 jieba[2]对输入的二元组进行分词。另一方面借助 lucene[10]来将庞大的知识库构建为索引，以方便实时检索；在这里我们采用的知识库是由复旦大学的肖教授构建的知识工厂[3]，知识库共有 60,000,000 条记录，涵盖天文百科等多个方面。构建完全知识库之后，我们使用基于 BM25 的匹配方法来进行计算。BM25 是一种基于统计学的匹配方式，它使用 idf 方法和 tf 方法的某种乘积来定义单个词项的权重，然后把和查询匹配的词项的权重相加作为分数，一般被认为可能比 tf-idf 更好，特别是在短文档集上。

2.4. 阶段二：知识过滤

由于在大规模的知识库中往往会检索出一些与问题所提问属性无关的东西，例如给定问题 Q = [‘枣庄’，‘市’，‘位于’，‘哪个’，‘省’，‘?’]，仅仅使用阶段一中的检索结果可能会将枣庄的气候条件，枣庄的政府驻地都进行返回，而我们仅仅需要枣庄的所属地区即可，所以我们需要进行知识过滤，将检索到的知识进行一定程度上的过滤，以帮助模型正确的选择所需要的知识。如图2所示，在此阶段对于输入的问题 Q= $[Q_1, Q_2, \cdots, Q_n]$，以及所检索到的知识序列 K = $[K_1, K_2, \cdots, K_n]$，我们首先使用嵌入层对二者进行由独热向量到实值的向量空间中。词嵌入是一种可以将稀疏高维独热的向量映射到低维实值空间的一种方法，对于给定的序列 Q 以及 K，经过词嵌入过后依然可以得到一个特征序列，如下公式所示：

$$E_Q = Embedding(Q),$$
$$E_K = Embedding(K)$$

之后我们使用门单元神经网络 (GRU) [4]来继续对 E_Q 和 E_K 提取时序特征信息。GRU 是循环神经网络

[2]https://pypi.python.org/pypi/jieba/.

[3]http://kw.fudan.edu.cn/about/.

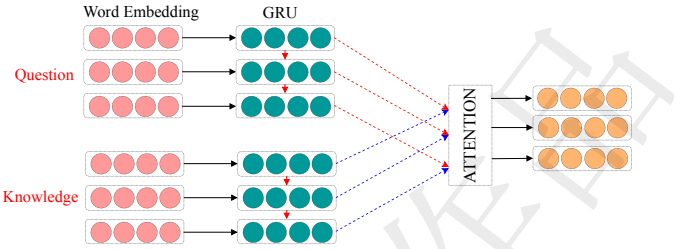

图 2. 对于给定的二元组 (question, knowledge)，使用 attention 机制去注意到特定的 knowledge 中的属性，进行知识过滤。

的一种经典的实现方式，它可以有效避免传统循环神经网络中出现的梯度消失以及梯度爆炸的问题，而且一般来说性能方面优于长短时记忆神经网络，同时它还兼具了循环神经网络可以有效建模上下文的特点，故我们采用 GRU 作为我们的实现。对于输入的序列 E_Q 中的第 t 个特征 E_Q^t 来说，GRU 会做如下计算：

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]),$$
$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]),$$
$$\widetilde{h_t} = \sigma(W_r \cdot [h_{t-1}, x_t]),$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t}$$

其中 h_t 是 GRU 在该时间步上返回的输出，W_r，W_Z 都是模型的参数，h_{t-1} 是上一个时间步的输出。

经过 GRU 之后，我们由 E_Q 与 E_K 得到了有时序性的序列 H_Q 以及 H_K。

对于 H_Q 以及 H_K 来说，我们希望由 H_Q 可以使用注意力机制来对 H_K 序列中的元素进行不同程度的加权，以凸显出重要的知识而忽略那些没有重要作用的知识。注意力机制是一种由人类仿生学启发得到的方法，它的原理在于使用一个查询向量 query 去对一个特征集合中不同的特征进行加权。这样可以对于不同的任务来凸显不同的特征，注意力机制在自然语言处理 [5]，计算机视觉 [19]中广泛被使用，得到了有力的认可。一般的来说，给定序列 H_Q，我们采用序列 H_Q 的最后一个时间步的结果作为 query，记作 h_Q，我们希望利用 h_Q 能够将 H_K 的不同权重进行加权，加权后的特征表达如下计算：

$$I_t = \sum_{i=1}^{n} \alpha_i * h_K^i,$$

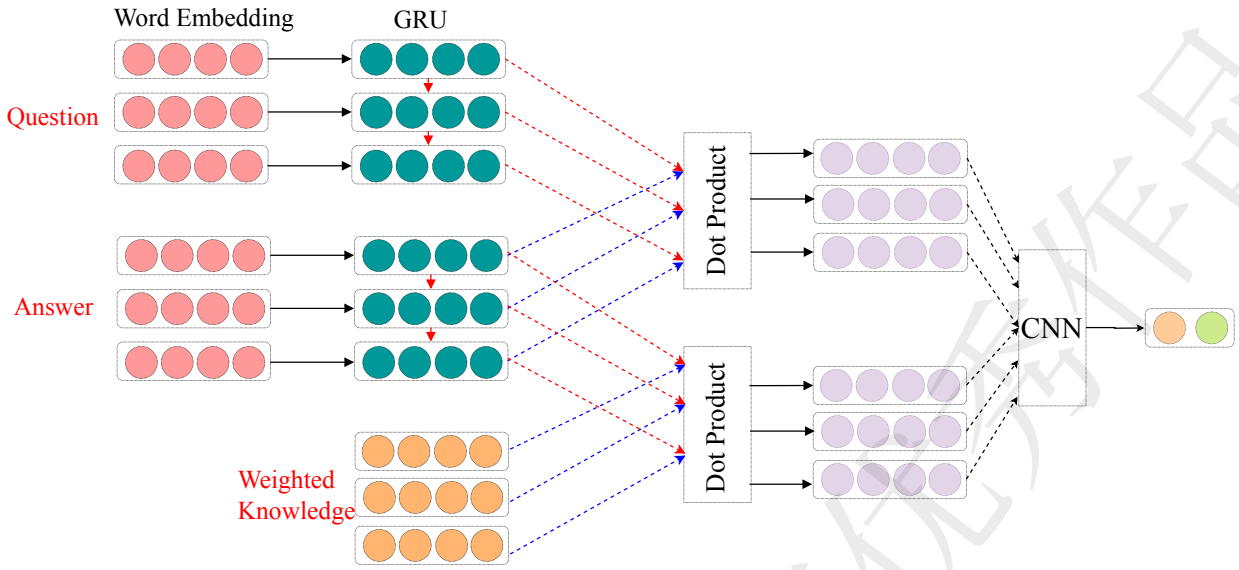

图 3. 深度语义匹配模型，首先对于 question 和 answer 使用 Embedding Layers 和 GRU 进行编码，之后使用点乘计算 Question-Answer，Question-Weighted_Knowledge 之间的关系，最后使用 CNN 聚集交互特征并且预测。

其中 α 的计算方式如下：

$$\alpha_i = align(h_Q, h_K^i) = \frac{exp(f(h_Q, h_K^i))}{\sum_i exp(f(h_Q, h_K^i))},$$

其中 f 一般被称为匹配函数，它可以有以下集中具体实现：

$$f(h_Q, h_K^i) = h_Q^T h_K^i,$$
$$f(h_Q, h_K^i) = h_Q^T W_a h_K^i$$
$$f(h_Q, h_K^i) = W_a[h_Q, h_K^i]$$
$$f(h_Q, h_K^i) = MLP(h_Q, h_K^i)$$

为了简洁以及计算的方便性，我们采用了点乘的方式来进行计算。

2.5. 阶段三：深度匹配

第三阶段的大体流程如图3所示。经过前两个阶段，我们得到了加权之后的知识表示 I_t，我们同样使用嵌入层和 GRU 对 Question 以及 Answer 进行表征，之后使用点乘计算 Question 与 Answer 之间的交互信息，同样对于加权后的知识也使用点乘计算它与 Answer 之间的交互信息，之后对于所得到的交互特征使用 CNN 进行特征提取，最后得到一个二维的向量，使用 softmax 函数对输出的概率进行归一化处理。数学上来说，softmax 的计算公式如下：

$$S_i = \frac{e^{y_i}}{\sum_{i \in 0,1} e^{y_i}}$$

，S 是给出的 softmax 之后的概率。

2.6. 损失函数

在这个问题中，我们采用的是交叉熵函数对整个训练过程进行优化，一般来说，损失函数使用下面的公式计算：

$$Loss = -\sum_{i \in 0,1} y_i log(p_i)$$

3. 实验

3.1. 超参数选取

在我们的模型实现中，我们对嵌入层的维度设置为 256 维；我们对所有的 GRU 都选用了 1024 维度的隐状态，并且 GRU 都是双向两层的架构；另外我们对于卷积采用了窗口大小为 3 的二维卷积及池化；我们使用 Adam 作为优化器来优化我们的损失函数，同时设定学习率为 0.001，同时我们划分了部分数据集作为线下测试集来使用 early stop 方法以避免出现过拟合的情况。我们使用了 MXNet 框架完成了我们的工作，MXNet 是一款高效，开源的深度学习框架；我们采用了服务器进行大规模的训练，服务器共拥有 4 块 titan xp 的显卡，256gb 的内存，以及 500gb 的 SSD，由于 titan xp 较大的显存，我们设定 batch size 的大小为 1000，这样网络收敛的会更加迅速。

问题	回复	知识库
中国最大的咸水湖是哪个湖”	”青海省阿坝州青海湖”	中国文化是东亚文化圈的文化宗主国
农行指的是什么银行	建设银行	银行，是依法成立的经营货币信贷业务的金融机构
武夷山是哪个省的	江西	武夷山位于江西与福建西北部两省交界处

表 1. 样例分析

3.2. 数据集增强与划分

3.2.1 训练数据增强与划分

我们仔细对参赛数据做出了分析，我们发现问题，回答的词长并不相同，为了方便起见，我们将所有的问题截取到 20 词的长度，将所有的回答补齐到 100 维的长度，我们共取了最后 1000 的问题以及所对应的答案作为了验证集。

3.2.2 知识图谱数据简介

我们主要采用了复旦大学的知识工厂作为知识库，知识工场源于复旦大学图数据管理实验室（GDM@FUDAN）。知识工场专注于各类大规模知识图谱构建、管理以及应用理论与方法研究。知识图谱表达了各类实体、概念及其之间的各种语义关系，成为了大数据时代知识表示的主要形态之一。知识图谱为语义理解提供了丰富的背景知识，为实现机器语言认知提供必需的知识支撑。知识工场以构建能够满足机器语言认知需要的大规模、高质量知识图谱为基本目标，并以推进知识图谱在文本理解、智慧搜索以及机器智脑等领域中的深入应用为主要使命。

3.3. 收敛曲线

为了验证我们的算法是高效而且可以收敛的，我们绘制了整个过程的收敛曲线，如图4所示，其中横轴是每个 epoch，纵轴是 loss 的大小，可以很清楚的观测到我们的算法是易于收敛的。

3.4. 结果呈现

如表所示，我们的算法在构造的线下测试集上 recall 可以达到 0.7，precision 可以达到 0.65，F1 可以达到 0.4+-0.07。这是因为我们的算法不仅建模了问题与回复在语义层级的匹配程度，这对于该赛题是至关重要的，良好的结果也从另外一个角度说明了我们算法的有效性。

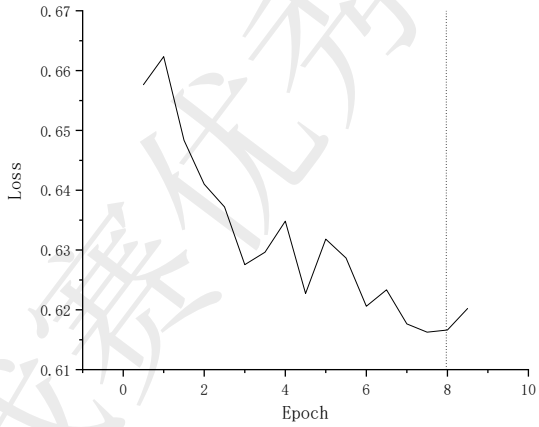

图 4. 收敛曲线

3.5. 样例分析

我们的算法在测试集上表现良好，但是我们同时也做了一些不好的样例的分析。知语出现错误主要有三种可能，在表1中我们共举了几个例子来说明这三种问题。

- 知识检索失败。因为知识库并未涵盖所有的知识，我们进行检索的时候往往可能检索不到，以前三个样例为例，它们都没有正确的检索到所需的知识。我们希望在下一步的算法中可以增大知识库的构成，甚至可以在线的去搜寻知识来解决这个问题。
- 知识过滤失败。因为知识过滤的时候并不是完全的可以正确选择知识，我们有部分样例虽然找到了知识，但是把有用的知识过滤出去了造成这个问题。下一步我们准备通过增加语义解析器来尝试通过语义解决这个问题。
- 深度匹配失败。存在这种三元组正确但是判别错误的例子，此现象我们暂时无法解释。猜测是与该

词条出现相近的样本出现的过少，造成了难以正确的去匹配。我们希望下一步的研究可以给出深度匹配在某些时刻出错的原因。

4. 总结与展望

综上所述，我们提出了知语模型-一款基于知识和语义双重匹配的深度文本匹配模型，它可以既从语义上又从事实上判别两段文本的匹配程度，我们的模型在线下测试集上取得了良好而鲁棒的特性，在实验模块我们对算法的收敛性，算法的错误分析，算法的结果做了详尽的分析。总体来看我们的贡献集中在知识与深度学习的融合上。在接下来的工作中，我们将主要关注在如何更好的解决知语所出现的三大问题，以及如何更加高效，迅速的训练，部署整个神经网络。

致谢. 十分感谢泰迪杯官方给出关于智能阅读模型的构建赛题，我们都知道，在当今信息爆炸的时代，信息无处不渗透进入人类的生活，如何有效选取信息是重要而深刻的课题，我们希望我们的初步尝试能够让民众更有利的获取自己所需要的信息。希望我们的工作可以成为人工智能黄金时代的一朵小小的浪花，为人类的事业添砖加瓦。

参考文献

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. A. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In OSDI, 2016. 3

[2] P. F. Brown, V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. Computational Linguistics, 18:467–479, 1992. 2

[3] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR, abs/1512.01274, 2015. 3

[4] J. Chung, Çaglar Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014. 4

[5] L. W. D'Avolio, T. M. Nguyen, and L. D. Fiore. The automated retrieval console (arc): open source software for streamlining the process of natural language processing. In IHI, 2010. 4

[6] H. Deubel and W. X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. Vision Research, 36:1827–1837, 1996. 2

[7] R. B. Girshick. Fast r-cnn. 2015 IEEE International Conference on Computer Vision (ICCV), pages 1440–1448, 2015. 3

[8] L. Gong and R. Ji. What does a textcnn learn? CoRR, abs/1801.06287, 2018. 2

[9] T. Joachims. Making large - scale svm learning practical. 1999. 2

[10] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In ACM Multimedia, 2008. 4

[11] L. Ma and Y. Zhang. Using word2vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data), pages 2895–2897, 2015. 2

[12] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, 2010. 2

[13] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, 2014. 2

[14] J. Ramos. Using tf-idf to determine word relevance in document queries. 2003. 2

[15] I. Rish. An empirical study of the naive bayes classifier. 2001. 2

[16] S. E. Robertson, H. Zaragoza, and M. J. Taylor. Simple bm25 extension to multiple weighted fields. In CIKM, 2004. 2

[17] L. Sha, B. Chang, Z. Sui, and S. Li. Reading and thinking: Re-read lstm unit for textual entailment recognition. In COLING, 2016. 3

[18] E. W. Steyerberg, F. E. Harrell, G. J. Borsboom, M. J. C. Eijkemans, Y. Vergouwe, and J. D. F. Habbema. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology, 54 8:774–81, 2001. 2

[19] R. Szeliski. Computer vision - algorithms and applications. In Texts in Computer Science, 2011. 4

[20] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In HLT-NAACL, 2016. 3

[21] H. Yu and J. Yang. A direct lda algorithm for high-dimensional data - with application to face recognition. Pattern Recognition, 34:2067–2070, 2001. 2

[22] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. CoRR, abs/1409.2329, 2014. 3