基于最大熵模型的中文阅读理解技术研究

李济洪，王凯华，王瑞波

山西大学语义网研究室，山西 太原 030006；

Email: {lijh, wangkaihua, wangruibo}@sxu.edu.cn

摘要：本文基于山西大学自主开发的中文阅读理解语料库 CRCC v1.1 版，根据问句和候选答案句的对应关系，构造了词层面以及句法层面共计 35 个特征，并使用最大熵模型对中文阅读理解问题回答进行了建模。考虑到语料库规模较小，我们以 1：1 的比例从 CRCC 语料库中随机选取了 5 组训练/测试集。在 5 组测试集上的平均 HumSent 准确率达到 75.46%。

关键词：阅读理解；问答系统；最大熵

Research on Chinese Reading Comprehension Based on Maximum Entropy Model

LI Ji-hong, WANG Kai-hua, WANG Rui-bo

Semantic Web Laboratory of Shanxi University, Taiyuan, Shanxi 030006

Email:{lijh, wangkaihua, wangruibo}@sxu.edu.cn

Abstract: In this paper, we construct 35 features (word-level feature and syntax-level feature) based on the CRRC v1.1 (Chinese reading comprehension corpus) built by Shanxi University. We built a model of Chinese reading comprehension by using the maximum entropy method. As the result of the corpus's scale is small, five groups of train/test sets are selected randomly by 1:1 scale from CRCC v1.1. The result shows that the average HumSent accuracy on the five testing sets has been 75.46%.

Key words: reading comprehension; question answering; maximum entropy methods

1 引言

首次尝试系统地考查自动系统阅读理解可行性的研究是 Hirschman(1999)[1]，Hirschman(1999)使用一个简单的词汇包（Bag-Of-Words）方法开发了第一个阅读理解问答系统 Deep read，在 Remedia 语料的测试集上得到了 36.3%的准确率。随后，Chaxniak(2000)[2]的研究工作将准确率提高为 41%。2000 年，Ellen Riloff 和 Michael Thelen[3]开发的一套基于规则的阅读理解系统 Quarc，Quarc 是一个启发式规则系统，它使用词汇和语义启发式规则来寻找问题的答案句，在 Remedia 语料上系统总体性能的 HumSent 准确率为 39.7%。2001 年，Hwee Ton Ng 等人[4]将机器学习方法用于阅读理解问答研究，并开发了系统 AQUAREAS，其系统总体性能在 Remedia 语料的测试集上得到了 39.3%HumSent 准确率。2005 年，杜永萍[5]采用扩展的 BOW（Bag-Of-Words）基本策略，并将表示问题的词集利用 WordNet 资源进行了同义词扩充，对问题中的不同元素（动词、实体名及基本名词短语）赋予不同的权值。并使用上下文辅助策略，系统在 Remedia 语料的测试集上的 HumSent 正确率为 42%。2006 年，Kui Xu[6]等利用了更深层次的自然语言处理技术。对问题句、候选答案句做句法分析，分析出词语的依赖关系和句法结构，再利

作者简介：李济洪（1964-），男，副教授，研究方向为自然语言处理.

用最大熵模型来选出答案句。该方法对 Remedia 语料的 HumSent 正确率达到 44.7%，对 ChungHua 双语语料库的 HumSent 正确率为 73.2%。

2007 年，王凯华、李济洪等[7]基于山西大学自主开发的中文阅读理解语料库 CRCC(v1.0)，使用 Hwee(2000)中的特征并加以改造为新的十个特征，采用最大熵方法[8,9]进行建模，在 CRCC 测试集上得到 61.5%的 HumSent 准确率。

2 特征表示

利用最大熵模型来解决阅读理解问题，首先需要设计一组特征，它尽可能包含语料库中能够将答案句与非答案句区分开的信息。ME 中的特征一般有以下形式：

$$f_i(x,y)=\begin{cases}C_i, & \text{如果特征 i 发生}\\ 0, & \text{其他}\end{cases}$$

其中 C_i 为实数。

我们实验中主要考虑了词层面和句法层面两大类特征，具体如下：

词层面特征保留了文章（王凯华、李济洪等，2007）中使用的 DMWM 等十个词层面特征。

句法层面的特征主要有以下三种特征：

1）**基本块特征**

我们采用清华大学周强老师提供的基本块标注工具[10]，对 CRCC 语料进行了自动基本块标注。基本块标记集有两层，一层是基本块名称标记，另一层是结构关系标记，标记集见文章[10]：

相应的，我们提取基本块特征也是根据基本块名标记和关系标记分别提取。

由于一个块内往往包含多个词，中心词起关键作用，因此，我们采用，同类型块内，块之间的匹配以块中心词匹配为准。这样可以突出句法层面语块匹配的信息。为此，我们还自行设计了基本块中心词提取工具，便于使用。

2）**结构关系特征**

在基本块标注后，同一类型基本块，再提取基本块的结构关系特征。

事实上，基本块特征与结构关系特征可以同时看其组合标记是否匹配，但考虑到可能带来数据稀疏，不便于数据处理，我们还是将其分别作为不同特征来处理。

3）**功能块特征**

我们使用清华大学周强老师提供的功能块标注工具[11]，将语料进行了标注，根据功能块标记的匹配情况，提取功能块特征。同样以功能块的中心词匹配作为功能块的匹配策略。

由于 S，P，O，D 功能块在语料库中占绝大多数（97%），故只考虑这四个功能标记所对应的特征，其它功能块不选为特征。

3 训练/测试集的选取

考虑到语料的规模不大，为了减少系统结果对训练集、测试集的较强依赖，我们在 CRCC 语

料库 v1.1 版本[2]中，选取 5 组训练/测试集，每组经随机选取 61 篇文章作为训练集，剩余 60 篇文章作为测试集。其中第一组中训练集共 844 个句子，291 个问句；测试集共 801 个句子，299 个问句，训练集与测试集中各问题类型分布如表 3-1：

表 3-1：训练集与测试集中各问题类型分布

问题类型	训练语料		测试语料	
	问题数	百分比	问题数	百分比
Q_DISCRIBE	147	50.52%	162	54.18%
Q_ENTITY	42	14.43%	35	11.71%
Q_HUMAN	12	4.12%	13	4.35%
Q_LOCATION	28	9.62%	23	7.69%
Q_NUMBER	37	12.71%	40	13.38%
Q_TIME	25	8.59%	26	8.70%
合计	291	100.00%	299	100.00%

其它组的训练测试问题类型分布与表 3-1 差别不大，不再详细列出，这里仅给出这 5 组随机测试集中问句个数分布：

表 3-2：各类型问句在 5 组测试集上的分布

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	合计
Q_DISCRIBE	162	149	148	164	161	784
Q_ENTITY	35	41	42	35	36	189
Q_HUMAN	13	13	13	13	13	65
Q_LOCATION	23	28	26	26	23	126
Q_NUMBER	40	41	37	37	38	193
Q_TIME	26	27	27	24	23	127
all	299	299	293	299	294	1484

4 系统评价方法

我们采用 HumSent 准确率（Hirschman *et al.*, 1999）[错误！未定义书签。]作为评价指标，HumSent 准确率是指能正确回答的问题占总测试问题的百分比。问题的正确答案为 CRCC 中文阅读理解语料中已经人工标出的句子。

在文章（王凯华、李济洪等，2007）中，也采用了 HumSent 准确率作为系统性能评价指标，在 ME 程序的自动输出标签（Y/N，是否是答案句）结果中，根据每个样本对应的文章编号、问句编号、句子编号，按问题确定测试集中每个问题对应的系统答案，如果系统自动标出的答案句有多个 Y（表示是答案的标签），则将最先出现的那句作为系统答案，然后将系统答案与 CRCC 测试集中人工标记的答案进行比较，若一致表示回答正确，否则表示回答错误，计算回答正确的问题个数占总问题个数的百分比，由此得到该系统的 HumSent 准确率。

但是，在大量试验中我们发现，直接采用 ME 程序的输出标签去判定是否是答案，得到的 HumSent 评价指标很不稳定。例如：问题 q 在 5 个句子构成的句子集中的标准答案标号为 002，即第二个句子为标准答案，ME 程序得到的各个句子为答案的概率分别为：0.6、0.9、0.4、0.1、0.2，系统根据是否大于 0.5 判断其是否答案，自动判定的答案标签分别为 Y、Y、N、N、N，由

[2] 由于多个答案的问题处理起来较为复杂，在实验中，我们已经预先将 CRCC 中多个答案的问题去掉了。

上面 HumSent 的判断原则（第一个 Y 标签为答案），答案为 001，与标准答案 002 不符，所以系统回答错误。而实际上，如果根据各个句子是答案句的概率值来判断，第二个句子对应的概率最大，为 0.9，答案为 002，系统答案与标准答案一致，所以该问题回答正确。因为在语料中，每个问题的答案多数只有一个句子，故这样判断比直接用第一个标签判断更为合理。

因此，在本文中，我们首先得到 ME 模块输出的各个句子为答案句的概率，然后将每个问题对应的各个句子的概率进行排序，概率最大的句子作为系统答案，由此计算 HumSent 准确率。为区别，我们将两个 HumSent 分别称为标签 HumSent 准确率和概率 HumSent 准确率。不特别声明的就是概率 HumSent 准确率。

5 实验结果及分析

如前所述，考虑到语料库的规模较小，系统的 HumSent 准确率比较容易受训练和测试集的影响，因此，我们随机从 CRCC 语料库中按训练/测试接近 1：1 的比例选取训练/测试集，基本上保持训练文章 61 篇，测试文章为 60 篇，同时相应各类型问题的训练测试比例也尽可能均衡。由于 CRCC 语料库中各篇文章中的各类型问句分布不相同，所以在训练/测试文章数的比例接近 1：1 的时候让各问句类型的训练/测试比例也尽可能均衡（即接近 1：1），人工选取这样的训练/测试语料难度较大，因此我们自行设计程序进行自动选择，在实验中，我们简单将均衡标准定为：训练/测试集中的问句数之差与之和的比值小于 0.1。例如：表 3-1 中的训练测试语料分布就符合这一标准。按这一标准，我们随机抽取 5 次，这样共得到 5 组训练/测试集。按这 5 组随机抽取的训练/测试集分别进行实验，以 5 组 HumSent 准确率的平均值作为评价指标。

在训练、测试模块，我们采用了张乐博士的最大熵工具包进行训练、测试。对于每类问题，我们使用相同的特征集训练一个模型。给定某一类型问题，训练样本由该类型问题中每个问题与相应文本中所有句子构成的“问题—句子”对的匹配情况而构建的特征信息构成。相应地，问题与标准答案句构成的“问题—句子”对为正例，其它“问题—句子”对为负例。

首先，我们分别使用原来的 10 个特征构成的特征集在这 5 组随机抽取的训练/测试集上进行实验。针对每一对训练/测试集，我们分别进行最大熵建模，然后在测试集上进行测试。这样，每种类型问句分别得到 5 个测试结果，取其 5 个测试结果的平均值作为该问句类型的测试结果.

使用文章（王凯华、李济洪，2007）中的 10 个特征，即词层面特征（DMWM、DMVM、DMWM_Prev、DMWM_Next、DMVM_Prev、DMVM_Next、PeO、DT、LO、VA），按照标签和概率两种评价标准进行实验，ME 训练时最大迭代次数统一设为 10000，5 组训练/测试集上的平均实验结果如表 5-1、表 5-2。

表 5-1：10 个特征在 5 组随机抽取的训练测试集上的标签 HumSent 评价结果

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	5 组平均
Q_DISCRIBE	67.90%	64.43%	68.92%	73.17%	70.19%	68.92%
Q_ENTITY	80.00%	70.73%	76.19%	71.43%	75.00%	74.67%
Q_HUMAN	23.08%	69.23%	69.23%	84.62%	53.85%	60.00%
Q_LOCATION	91.30%	85.71%	84.62%	88.46%	82.61%	86.54%
Q_NUMBER	75.00%	68.29%	83.78%	64.86%	76.32%	73.65%

Q_TIME	80.77%	81.48%	77.78%	87.50%	78.26%	81.16%
all	71.24%	69.57%	74.06%	74.92%	72.45%	72.45%

从表 5-1 可以看出，新的测试结果为 72.45%，比文章（王凯华、李济洪，2007）中测试结果 61.5%高出 10.95%，原因主要是文章（王凯华、李济洪，2007）中采用的是 v1.0 版的 CRCC 语料，而表 5-1 结果使用的是 v1.1 新版的 CRCC 语料，v1.1 新版的 CRCC 语料对 v1.0 版进行问题类型、问句编号、答案标记、标点、格式等的多次人工校对，因此测试结果有较大的上升。

表 5-2：10 个特征在 5 组随机抽取的训练测试集上的概率 HumSent 评价结果

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	5 组平均
Q_DISCRIBE	78.40%	77.18%	74.32%	78.05%	77.02%	76.99%
Q_ENTITY	85.71%	73.17%	80.95%	77.14%	77.78%	78.95%
Q_HUMAN	38.46%	76.92%	69.23%	76.92%	53.85%	63.08%
Q_LOCATION	91.30%	89.29%	92.31%	92.31%	86.96%	90.43%
Q_NUMBER	77.50%	82.93%	83.78%	78.38%	81.58%	80.83%
Q_TIME	88.46%	88.89%	81.48%	91.67%	78.26%	85.75%
all	79.26%	79.60%	78.50%	80.27%	77.55%	79.04%

表 5-2 中，采用概率评价得到的在 5 组随机测试集上的平均实验结果为 79.04%，比表 5-1 中 72.45%高出 6.59%，这主要是由于结果评价方式改变为更合理的概率评价所致，所以有较大幅度的上升，关于概率评价详见第 4 部分。

接下来，为了看基本块特征、结构关系特征、功能块特征对系统性能的影响，我们逐次加入这三类特征，分三组实验来进行：

1）18 个特征（即：10 个特征+8 个基本块特征[ap、dp、mbar、mp、np、sp、tp、vp]），我们使用这 18 个特征构成的特征集在 5 组随机抽取的训练/测试集上进行实验，平均结果如表 4-9：

表 5-3：18 个特征在 5 组随机抽取的训练测试集上的概率 HumSent 评价结果

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	5 组平均
Q_DISCRIBE	80.25%	78.52%	72.30%	76.22%	77.64%	76.99%
Q_ENTITY	88.57%	68.29%	78.57%	82.86%	72.22%	78.10%
Q_HUMAN	61.54%	76.92%	61.54%	76.92%	53.85%	66.15%
Q_LOCATION	91.30%	75.00%	88.46%	88.46%	82.61%	85.17%
Q_NUMBER	80.00%	78.05%	89.19%	75.68%	78.95%	80.37%
Q_TIME	80.77%	85.19%	66.67%	100.00%	78.26%	82.18%
all	81.27%	77.26%	75.77%	79.93%	76.53%	78.15%

2）31 个特征（即：10 个特征+8 个基本块特征+13 个结构关系特征[AD、AM、CD、LH、LN、NH、PO、RL、SB、SG、SX、XX、ZX]），我们使用这 31 个特征构成的特征集在 5 组随机抽取的训练/测试集上进行实验，平均结果如表 4-10：

表 5-4：31 个特征在 5 组随机抽取的训练测试集上的概率 HumSent 评价结果

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	5 组平均
Q_DISCRIBE	79.01%	76.51%	72.30%	78.66%	77.02%	76.70%
Q_ENTITY	82.86%	68.29%	78.57%	68.57%	75.00%	74.66%
Q_HUMAN	69.23%	53.85%	76.92%	61.54%	53.85%	63.08%
Q_LOCATION	91.30%	75.00%	92.31%	88.46%	78.26%	85.07%
Q_NUMBER	75.00%	75.61%	89.19%	75.68%	78.95%	78.88%

Q_TIME	61.54%	85.19%	66.67%	70.83%	56.52%	68.15%
all	77.93%	74.92%	76.79%	76.59%	74.49%	76.14%

3）35 个特征（即：10 个特征+8 个基本块特征+13 个结构关系特征+4 个功能块特征[D、O、P、S]），我们使用全部 35 个特征构成的特征集在 5 组随机抽取的训练/测试集上进行实验，平均结果如表 4-11：

表 5-5：35 个特征在 5 组随机抽取的训练测试集上的概率 HumSent 评价结果

问句类型	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组	5 组平均
Q_DISCRIBE	77.78%	78.52%	72.30%	78.05%	75.78%	76.48%
Q_ENTITY	85.71%	75.61%	76.19%	68.57%	72.22%	75.66%
Q_HUMAN	61.54%	46.15%	46.15%	53.85%	53.85%	52.31%
Q_LOCATION	91.30%	82.14%	92.31%	88.46%	82.61%	87.37%
Q_NUMBER	70.00%	78.05%	83.78%	75.68%	76.32%	76.76%
Q_TIME	61.54%	74.07%	77.78%	62.50%	56.52%	66.48%
all	76.59%	76.59%	75.43%	75.25%	73.47%	75.46%

将以上三组实验得到的平均测试结果与 10 个特征时得到的平均测试结果作比较，如图 5-1：

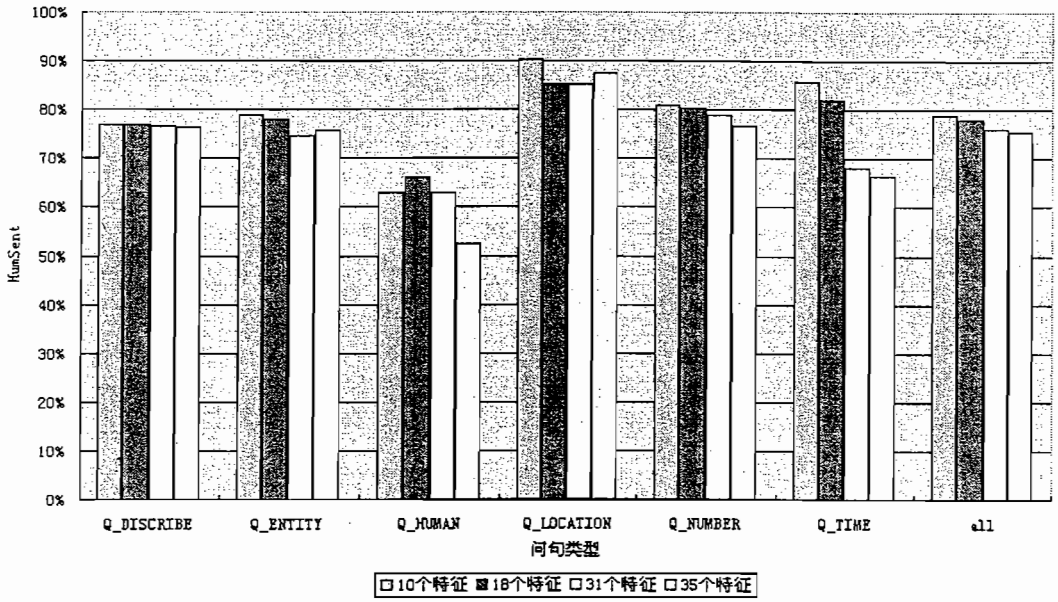

图 5-1： 10 个特征与 35 个特征分别对应的模型在 5 组随机测试集上的平均测试结果比较

从图 5-1 中，我们可以看出，逐次加入特征对 DISCRIBE 类型的问句影响不大，类型为 HUMAN 和 TIME 的波动较大，HUMAN 类型的问句在 18 个特征时达到最好的测试结果。总体而言，随着特征的逐渐增多，测试 HumSent 准确率并没有按我们原来设想的效果逐渐增加，而是随着特征的增多，测试准确率还有所下降。主要原因有以下几个方面：一是语料库规模比较小，随着特征的增多，特征矩阵会越来越稀疏，一些新的块特征信息在语料库中没有体现出来；二是基本块和功能块标注工具存在一定的错误率，实验用的测试集都是自动标注语料，所以句法层面的特征由于存在部分标注错误影响到系统的性能；三是 35 个特征之间可能存在相关性，影响了最终效果的发

挥。

结束语

本文基于山西大学自主开发的中文阅读理解语料库 CRCC v1.1 版，根据问句和候选答案句的对应关系，构建词层面特征、句法层面特征，采用最大熵模型对中文阅读理解问题回答进行建模。

下一步我们将继续研究各种特征如何有效组合，以及如何有效进行特征选择，使用多种建模方法，对阅读理解系统建模。在此基础上，再进一步研究如何加入更多的特征，比如语义特征，深入研究基于语义的中文阅读理解问答系统。

致谢

实验过程中分别使用了：山西大学 FC2000 分词软件、清华大学周强老师提供的基本块标注工具和功能块标注工具、张乐博士的最大熵工具包。在此，一并向他们表示感谢！

参考文献

[1] Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. Deep Read: a reading comprehension system[C]. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. College Park, Maryland, 1999:325-332.

[2] Eugene Charniak, Yasemin Altun, Rodrigo de Salvo Braz, et al. Reading comprehension programs in a statistical-language-processing class[C]. In Preceedings of the ANLP/NAACL 2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Seattle, Washington, 2000:1-5.

[3] Ellen Riloff and Michael Thelen. A Rule-based Question Answering System for Reading Comprehension Test[C]. ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems. Seattle, Washington. 2000:13-19.

[4] Hwee Tou Ng, Leong Hwee Teo, Jennifer Lai Pheng Kwan. A Machine Learning Approach to Answering Questions for Reading Comprehension Tests[C]. Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000.

[5] 杜永萍. 基于模式知识库的问题回答关键技术研究[D]. 博士论文. 复旦大学. 2005.

[6] Kui Xu, Helen Meng and Fuliang Weng. A Maximum Entropy Framework that IntegratesWord Dependencies and Grammatical Relations for Reading Comprehension[C], Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. 2006 : 185–188.

[7]王凯华, 李济洪, 张国华, 王瑞波. 基于最大熵模型的中文阅读理解问答系统技术研究[C] CNCCL-2007. 内容计算的研究与应用前沿. 北京:清华大学出版社, 2007:643-648.

[8] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing[J], Computational Linguistics, Philadelphia, USA, 1996 : 39-71.

[9] 周雅倩. 最大熵方法及其在自然语言处理中的应用[D]. 博士论文. 复旦大学. 2005.

[10] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007(3):21-27

[11] 赵颖泽. 汉语功能块的自动分析[D]. 硕士论文. 清华大学. 2006