·第 5 代指挥信息系统与关键技术专题· doi：10.15908/j.cnki.cist.2018.05.009

基于自交互注意力机制的文本摘要方法*

丰 冲[1] 潘志强[1] 撒 红[2] 陈洪辉[1]

（1 国防科技大学系统工程学院 长沙 410073）
（2 解放军 31010 部队 北京 100000）

摘 要：编码-解码框架及注意力机制已成功应用于自动文摘，但传统的自动文摘方法过于关注解码部分对显著性句子的抽取，且仅考虑了每个句子之前的历史信息，在文档编码过程中并未发掘句子间的联系及句子与整个文档的相关性。针对上述问题，提出了一种基于自交互注意力机制的、具有编码器-解码器结构的文本摘要模型（ESSA）来自动获取抽取式摘要。ESSA 先获取文档的整体信息，再计算不同句子间的关联信息，最后将二者结合得到丰富的文档向量表示。试验结果表明，ESSA 效果明显优于基准模型，该模型的 ROUGE-1、ROUGE-2、ROUGE-3、ROUGE-4 和 ROUGE-L 评分与较好的基准模型相比分别提高了 7.4%、24.3%、13.4%、7.1%和 7.6%。

关键词：文本摘要；注意力机制；编码器-解码器
中图分类号：TP391 **文献标识码**：A **文章编号**：1674-909X(2018)05-0057-05

Text Summarization Method Based on Self-Interactive Attention Mechanism

FENG Chong[1] PAN Zhiqiang[1] SA Hong[2] CHEN Honghui[1]
(1 College of System Engineering, National University of Defense Technology, Changsha 410073, China)
(2 Unit 31010 of PLA, Beijing 100000, China)

Abstract: The encoder-decoder architecture and the attention mechanism are successfully applied in automatic summarization. But the traditional automatic summarization methods put too much focus on the signal sentence extracting with the decoder. The methods only think about the history information of each sentence, without mining the relationship among the sentences and the association between the sentence and the whole document in the document encoding process. Aimed at the above problems, an encoder-decoder summarization model based on self-interactive attention (ESSA) is proposed to automatically get the extractive summary. In the ESSA, the global information of the document is obtained, and the relationship information among the different sentences is calculated. Finally, combining the two kinds of information, a rich document expression with a vector is got. The experimental results show that the effect of the ESSA is obviously better than the reference models. The model's ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 and ROUGE-L scorings are respectively increased about 7.4%, 24.3%, 13.4%, 7.1% and 7.6% more than the better reference model's.

Key words: text summarization; attention mechanism; encoder-decoder

*基金项目："十三五"预研课题和预研基金资助项目。
收稿日期：2018-06-27
引用格式：丰冲，潘志强，撒红，等.基于自交互注意力机制的文本摘要方法[J].指挥信息系统与技术，2018，9(5)：57-61.
FENG Chong, PAN Zhiqiang, SA Hong, et al. Text summarization method based on self-interactive attention mechanism[J]. Command Information System and Technology, 2018, 9(5): 57-61.

0 引 言

当今时代，互联网迅猛发展，各类文本数据急剧增长，发展自动文摘系统的需求愈发迫切。自动文摘系统指能够自主、快速地确定一篇文档的主要内容来生成文档摘要，为快速获取信息和情报提供了便利。抽取式摘要指通过从文档中抽取显著包含文档主要信息的句子来生成摘要[1]。在抽取式摘要方法中，编码器-解码器结构已得到有效应用[2]；此外，深度学习中的注意力机制能够帮助模型发掘包含重要文档信息的部分，对提升摘要效果有明显作用[3]。但是，现有基于编码器-解码器结构的自动文摘模型过多关注解码器部分[4-5]，关注与源文档更相关句子的获取，而未在编码器中发掘更丰富的文档信息，忽视了不同句子间的联系。由于句子间的关联信息对选取文档中的多样化信息具有重要作用，因此自动文摘时需先尽可能完整地获取文档信息。

基于以上考虑，本文针对单文档自动文摘问题提出了一种基于自交互注意力机制[6]、具有编码器-解码器结构的文本摘要模型(ESSA)。该模型由基于自交互注意力机制的语句编码器、文档编码器和语句抽取器构成。1) 语句编码器：利用词向量生成句子的向量表示；2) 文档编码器：以句子的向量表示为输入，先提取文档的整体信息，再利用自交互注意力机制得到不同句子间的关联信息，文档编码器再次读入文档语句，将语句信息与语句间的关联信息结合起来得到更丰富的文档表示；3) 语句抽取器：根据语句信息和包含文档特征信息的编码器输出确定应选取的句子。

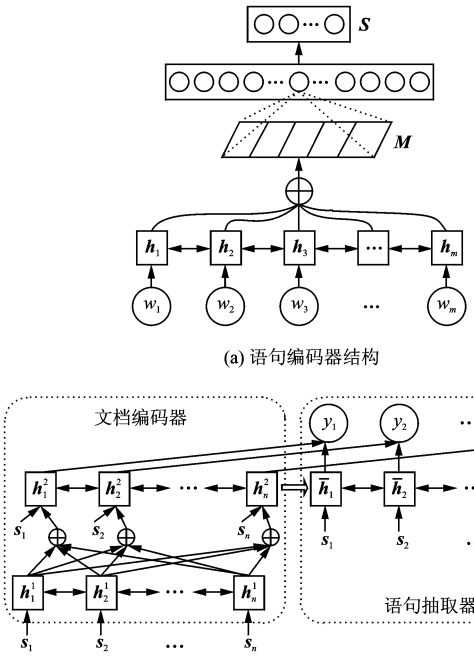
本文在美国有线电视新闻网(CNN)新闻数据集上使用基于召回率的要点评估方法(ROUGE)对 ESSA 进行自动评估，该数据集中每篇文档包含由作者书写的文章梗概作为标准摘要。试验结果表明，ESSA 得到的摘要具有更丰富的信息，能更好地替代原文档，提高了信息获取效率。

1 ESSA

首先，给出抽取式摘要任务的定义。给定一篇文档 d，由 n 个句子构成的序列 $(s_1, s_2, \cdots, s_n)$ 组成，从该序列中选取 $m(m<n)$ 个句子组成的子序列来构成文档 d 的摘要。为了实现该目标，对每个句子 s_i 进行打分并标注标签 $y_i \in (0,1)$，1 表明 s_i 应作为摘要的候选句子，否则不予考虑。在进行有监督的学习时，给定文档 d 及模型参数 θ。为将每个句子标签的概率 $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)$ 最大化，即：

$$p(\hat{\boldsymbol{y}} \mid \boldsymbol{d}; \theta) = \prod_{t=1}^{n} p(\hat{y}_t \mid \boldsymbol{d}; \theta) \tag{1}$$

提出了 ESSA，由语句编码器、文档编码器和语句抽取器 3 部分构成。ESSA 结构如图 1 所示。



(a) 语句编码器结构

(b) 文档编码器和语句提取器结构

图 1 ESSA 结构

语句编码器结构如图 1(a)所示。语句编码器主要以一个双向的长短时记忆网络(LSTM)为基础，通过加入自交互注意力机制[6]、利用词语信息及其之间的关联信息生成语句的向量表示。

文档编码器和语句抽取器结构如图 1(b)所示。文档编码器由 2 层双向 LSTM 构成，在第 1 层获取文档整体信息基础上，经过一个自交互注意力机制可提取不同语句间的关联信息，并将其输入第 2 层 LSTM，结合各语句信息得到最终的文档表示。

语句抽取器由一个单向 LSTM 构成，利用文档编码器生成的多样化文档信息进行句子筛选，得到最终摘要。

1.1 语句编码器

语句编码器作用是根据词向量得到包含丰富词语信息的句子向量。假设一个句子 s 由 m 个词 $(w_1, w_2, \cdots, w_m)$ 构成，将这些词输入语句编码器。为了获取相邻词汇间的依赖关系，使用一个双向 LSTM 网络来处理这些词汇：

$$\overrightarrow{\boldsymbol{h}}_t = \overrightarrow{\mathrm{LSTM}}(w_t, \overrightarrow{\boldsymbol{h}}_{t-1}) \tag{2}$$

$$\overleftarrow{\boldsymbol{h}}_t = \overleftarrow{\mathrm{LSTM}}(w_t, \overleftarrow{\boldsymbol{h}}_{t+1}) \tag{3}$$

将第 t 个词的隐状态$\vec{\boldsymbol{h}}_t$与$\overleftarrow{\boldsymbol{h}}_t$拼接起来便可得到$\boldsymbol{h}_t$，将得到的 m 个词的隐状态$\boldsymbol{h}_t$简化记为矩阵 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_m)$。

为了使长度不同的句子得到的句子向量长度相同，使用一种自交互注意力机制[6]对矩阵 $\boldsymbol{H}$ 进行线性连接处理，即将 $\boldsymbol{H}$ 作为输入，输出向量的权重为：

$$\boldsymbol{a}=\mathrm{softmax}(\boldsymbol{w}_{s2}\tanh(\boldsymbol{W}_{s1}\boldsymbol{H}^{\mathrm{T}})) \tag{4}$$

其中，$\boldsymbol{W}_{s1}$为权重矩阵，可通过训练得到；$\boldsymbol{w}_{s2}$为向量参数；softmax()为归一化指数函数。通过权重向量 $\boldsymbol{a}$ 对矩阵 $\boldsymbol{H}$ 各列进行求和可得一个句子的向量表示，但该向量表示仅包含句子某一特定位置的信息，而句子不同位置可能包含不同种类的重要信息，不可忽视。故需获得代表句子不同区域的信息，从而获得包含句子整体语义信息的向量表示。因此，将向量$\boldsymbol{w}_{s2}$扩展为矩阵$\boldsymbol{W}_{s2}$，可得输出矩阵的权重 $\boldsymbol{A}$ 为：

$$\boldsymbol{A}=\mathrm{softmax}(\boldsymbol{W}_{s2}\tanh(\boldsymbol{W}_{s1}\boldsymbol{H}^{\mathrm{T}})) \tag{5}$$

于是，可得表征句子不同区域信息的矩阵 $\boldsymbol{M}=\boldsymbol{A}\boldsymbol{H}$。因 $\boldsymbol{M}$ 的每一列包含的信息有较大相似性，故对 $\boldsymbol{M}$ 按列进行最大化采样，可得最终句子向量 $\boldsymbol{s}$。

1.2　文档编码器

文档编码器由一个 2 层 LSTM 构成。第 1 层先按顺序读入文档中的句子，获取其特征并得到初步的文档表示，再运用一种注意力机制获取包含文档多方面信息的向量表示，包括句子间的交互信息。第 2 层为了减少信息丢失会再次读入该句子序列，并将句子的向量表示与注意力机制层的输出结合起来，获取最终的文档表示。

给定文档 $\boldsymbol{d}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n)$，编码器的第 1 层在时刻 t 的隐状态更新方法为：

$$\vec{\boldsymbol{h}}_t^1=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{s}_t,\vec{\boldsymbol{h}}_{t-1}^1) \tag{6}$$

$$\overleftarrow{\boldsymbol{h}}_t^1=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{s}_t,\overleftarrow{\boldsymbol{h}}_{t+1}^1) \tag{7}$$

其中，$\boldsymbol{s}_t$为句子向量，初始时刻的隐状态 $\vec{\boldsymbol{h}}_0^1$和 $\overleftarrow{\boldsymbol{h}}_{n+1}^1$设置为零向量。将 $\vec{\boldsymbol{h}}_t^1$和$\overleftarrow{\boldsymbol{h}}_t^1$进行拼接，即可得到时刻 t 的状态向量$\boldsymbol{h}_t^1$。为便于计算，将 n 个隐状态合并记为 $\boldsymbol{H}=(\boldsymbol{h}_1^1,\boldsymbol{h}_2^1,\cdots,\boldsymbol{h}_n^1)$。

在时刻 t，隐状态$\boldsymbol{h}_t^1$不仅含有句子$\boldsymbol{s}_t$之前的历史信息，而且包含距离$\boldsymbol{s}_t$较远句子的信息。为了更好描述句子间的联系，得到更多有用信息，设计了一种自交互注意力机制来对编码器第 1 层的输出进行处理。

首先，对第 1 层的各隐状态赋予不同权重并求和：

$$\widetilde{\boldsymbol{h}}_t=\sum_{j=1}^{n}a_j^t\boldsymbol{h}_j^1 \tag{8}$$

其中，a_j^t为 t 时刻对第 j 个隐状态的归一化权重，有：

$$a_j^t=\exp(e_j^t)/\sum_{i=1}^{n}\exp(e_i^t) \tag{9}$$

e_j^t为仅利用 $\boldsymbol{H}$ 计算得到的初始权重值：

$$e_j^t=v_j^t\tanh(\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}) \tag{10}$$

v_j^t 和 $\boldsymbol{W}$ 为可训练的模型参数。于是，每个$\widetilde{\boldsymbol{h}}_t$均包含了$\boldsymbol{s}_t$和其他句子间的联系。

在时刻 t，将$\widetilde{\boldsymbol{h}}_t$输入编码器第 2 层。为了尽可能减少信息丢失，同时将句子向量输入并与$\widetilde{\boldsymbol{h}}_t$结合起来。时刻 t 的第 2 层隐状态的更新方式为：

$$\boldsymbol{h}_t^2=\mathrm{LSTM}([\boldsymbol{s}_t,\widetilde{\boldsymbol{h}}_t],\boldsymbol{h}_{t-1}^2) \tag{11}$$

其中，$[\boldsymbol{s}_t,\widetilde{\boldsymbol{h}}_t]$表明将$\boldsymbol{s}_t$与$\widetilde{\boldsymbol{h}}_t$连接起来。由此，可得文档的向量表示。

1.3　语句抽取器

语句抽取器由一个 LSTM 网络构成，可检测计算每个句子的显著性并进行标注。给定文档 $\boldsymbol{d}$ 及文档编码器的隐状态$(\boldsymbol{h}_1^2,\boldsymbol{h}_2^2,\cdots,\boldsymbol{h}_n^2)$，抽取器将结合当前解码的隐状态及对应位置编码的隐状态，对第 t 个句子的标签做出以下预测：

$$p(\boldsymbol{y}_t\mid\boldsymbol{s}_t,\boldsymbol{d})=\mathrm{softmax}(\sigma(\boldsymbol{h}_t^2,\bar{\boldsymbol{h}}_t)) \tag{12}$$

其中，$\sigma(\boldsymbol{h}_t^2,\bar{\boldsymbol{h}}_t)$表明一个多层网络，计算过程为：

$$\sigma(\boldsymbol{h}_t^2,\bar{\boldsymbol{h}}_t))=\boldsymbol{V}\tanh(\boldsymbol{U}_1\boldsymbol{h}_t^2+\boldsymbol{U}_2\bar{\boldsymbol{h}}_t) \tag{13}$$

其中，$\boldsymbol{U}_1$、$\boldsymbol{U}_2$ 和 $\boldsymbol{V}$ 为可训练的神经网络参数；$\bar{\boldsymbol{h}}_t$ 为语句抽取器的隐状态，计算方法为：

$$\bar{\boldsymbol{h}}_t=\mathrm{LSTM}(\boldsymbol{s}_t,\bar{\boldsymbol{h}}_{t-1}) \tag{14}$$

$\bar{\boldsymbol{h}}_0$ 即文档编码器最后输出的隐状态 $\boldsymbol{h}_n^2$。模型训练过程中使用的损失函数为：

$$\mathrm{loss}=-\frac{1}{n}\prod_{t=1}^{n}p(\boldsymbol{y}_t=\bar{\boldsymbol{y}}_t\mid\boldsymbol{s}_t,\boldsymbol{d}) \tag{15}$$

最终，对句子 $\boldsymbol{s}_t$ 是否应选作摘要的预测结果为：

$$\hat{y}_t=\arg\max_{y\in\{0,1\}}p(\boldsymbol{y}_t\mid\boldsymbol{s}_t,\boldsymbol{d}) \tag{16}$$

2　验证试验

本文需研究以下问题：1）与基准模型相比，ESSA 中注意力机制能否提升抽取式摘要的效果？2）生成摘要的长度，即 75 B、275 B 及完整长度（3 个句子）对结果有何影响？

基于编码器-解码器结构的模型通过从原始文档中选取显著的句子来生成摘要，故本文将其与抽取式摘要模型进行对比。考虑以下 2 个模型为基准模型：1）LEAD：一种选取文档前 3 句作为摘要的标准模型[1,4,7]；2）NN-SE：一种进行抽取式摘要的神经网络模型，包括分层文档编码器和一个基于注意

力机制的句子抽取器[4]。

数据集和试验细节：根据 CNN 新闻构造模型训练和测试的数据集[4]。CNN 新闻数据集广泛应用于自动问答系统研究，每篇文档包含新闻原文及新闻编辑人工写就的高亮文本，这些高亮文本可作为真正的生成式摘要，可将其作为标准摘要。Cheng 等[4]利用基于规则的方法给文档中每个句子均标注了标签 0 或 1(1 表明该句子与高亮文本相匹配；0 则相反)，经验证，句子标签正确率达 85%。数据集统计特征见表 1。由于数据集中超过 95%的句子不超过 50 个词，超过 95%的文档不超过 60 个句子，因此将句子长度设为 50，文档长度设为 60。针对文档编码器和句子抽取器，使用大小为 650 的长短时记忆 LSTM 单元。在 LSTM 输入进入隐层以及句子评分过程中使用的正则化丢弃率为 0.5。在训练过程中分批进行训练，1 个批次数据包含 20 篇文档，使用 Adam 优化器，其初始学习率为 0.001。

表 1 数据集统计特征

变量	训练集	测试集
文档数量	83 568	1 093
最长句子长度	1 341	1 426
最长文档长度	125	105
平均句子长度	23.6	23.1
平均文档长度	29.8	30.2

评估方法：在 CNN 的整个测试集上使用 ROUGE[8]对模型得出的摘要质量进行评估。ROUGE 是一种基于召回率的衡量方法。其中，ROUGE-N (N=1,2,3,4)能够衡量候选摘要与参考摘要间的 n 元词组的召回率，可用于衡量摘要包含的信息量；而 ROUGE-L 能够检测其最长公共子序列，可反映摘要的可读性和语言的流畅性。本文使用 ROUGE-1(R-1)、ROUGE-2(R-2)、ROUGE-3(R-3)和 ROUGE-4(R-4)反映摘要的信息量，ROUGE-L(R-L)反映摘要的流畅性，并给出生成完整长度和固定长度(前 75 B 和 275 B)的摘要。为了实现公平对比，选取得分最高的 3 个句子作为完整长度的摘要。

3 结果与分析

对 ESSA 和 2 个基准模型生成的 75 B、275 B 及完整长度(3 个句子)3 种长度摘要使用 ROUGE 评分验证模型。各模型不同长度摘要的 ROUGE 评分对比如表 2 所示。

表 2 不同长度摘要的 ROUGE 评分对比 %

摘要长度/B	模型	R-1	R-2	R-3	R-4	R-L
75	LEAD	12.2	2.6	1.1	0.6	11.1
	NN-SE	12.4	2.8	**1.3**	0.7	11.2
	ESSA	**13.1**	**3.0**	**1.3**	**0.8**	**11.6**
275	LEAD	34.6	11.0	5.1	3.2	32.3
	NN-SE	35.5	11.2	5.3	3.3	31.1
	ESSA	**38.8**↑	**12.6**	**6.7**↑	**4.1**	**33.8**↑
完整长度	LEAD	43.5	13.3	8.1	5.3	40.1
	NN-SE	44.5	13.6	8.2	5.6	39.4
	ESSA	**47.8**↑	**16.9**↑	**9.3**↑	**6.0**	**42.4**↑

注：下划线：不同长度摘要每一列最优的基准模型结果；

加粗：不同长度摘要每一列的最优结果；

↑(或↓)：经过 t 检验(p<0.05)ESSA 结果与最优基准模型结果相比有显著提升(或下降)。

1) 摘要模型效果

针对第 1 个研究问题，比较各模型完整长度摘要的 ROUGE 评分。由表 2 可见，对于 2 个基准模型，NN-SE 模型的 ROUGE-N 评分均比 LEAD 模型高，但其 R-L 评分略低于 LEAD 模型，表明 NN-SE 模型生成的摘要信息量更丰富，但 LEAD 模型得到的摘要更通顺流畅；ESSA 各项评分均超过了基准模型，取得了最优结果，且提升较明显，相较于 NN-SE 模型，ESSA 的 R-1、R-2、R-3、R-4 和 R-L 评分分别提升了 7.4%、24.3%、13.4%、7.1%和7.6%，其中有 4 项提升显著。可见，ESSA 在语句编码器和文档编码器中使用自交互注意力机制与双向 LSTM 结合后，有助于获取文档的主旨并选取有显著意义的句子作为摘要。

2) 不同长度摘要效果

针对第 2 个研究问题，试验还对比了 3 个模型生成的 75 B、275 B 和完整长度(3 个句子)3 种长度摘要的 ROUGE 评分。由表 2 可见，ESSA 在各种长度的摘要上均取得最好结果，而 2 种基准模型中 NN-SE 结果比 LEAD 稍好。在生成 75 B 长度摘要时，ESSA 各项 ROUGE 评分比 2 个基准模型均有较小提升；在生成 275 B 长度摘要时，ESSA 的 R-1、R-2、R-3 和 R-4 与 NN-SE 模型相比分别提升了 9.3%、12.5%、26.4%和 24.2%，ESSA 的 R-L 评分比 LEAD 模型提升了 4.6%，且 ESSA 的 R-1、R-3 和 R-L 提升显著。综上，ESSA 在生成长度较长的摘要时，效果更好。

4 结束语

本文提出了一种文本摘要模型来自动获取抽取

式摘要。该模型利用编码器-解码器结构，并使用自交互注意力机制，能够很好地发掘文本信息和结构特征，得到 ROUGE-N 和 ROUGE-L 评分更高的摘要，表明该模型生成摘要的信息量更多且更流畅。试验结果表明，ESSA 摘要效果显著优于 LEAD 和 NN-SE 2 个基准模型，特别是在生成长度较长的摘要时，效果更明显。后续将在不同数据集上测试 ESSA，检验该模型在不同领域文本上的有效性，也将利用文档的主题、标题及段落间相关性等其他特征，捕获文档更丰富、更显著的信息，改善自动摘要效果。

参考文献(References)：

[1] REN P, CHEN Z, REN Z, et al. Leveraging contextual sentence relations for extractive summarization using a neural attention model[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Tokyo: ACM, 2017.

[2] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th Advances in Neural Information Processing Systems. [S.l.]: NIPS, 2014: 3104-3112.

[3] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//ICLR 2015. San Diego: ICLR, 2015.

[4] CHENG J P, LAPATA M. Neural summarization by extracting sentences and words[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016.

[5] NALLAPATI R, ZHOU B, DOS SANTOS C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016: 484-494.

[6] LIN Z H, FENG M W, DOS SANTOS C N, et al. A structured self-attentive sentence embedding[C]//Proceedings of the 5th International Conference on Learning Representations. [S.l.]: ICLR, 2017: 1-15.

[7] NALLAPATI R, ZHAI F, ZHOU B. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents[C]//Proceedings of the 31st AAAI Conference on Artificial. San Francisco: AAAI, 2017: 3075-3081.

[8] LIN C Y, HOVY E. Automatic evaluation of summaries using N-gram co-occurrence statistics[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton: ACM, 2003: 71-78.

作者简介：

丰　冲，男(1993—)，硕士研究生，研究方向为指挥信息系统。

潘志强，男(1998—)，研究方向为指挥信息系统工程。

撒　红，女(1974—)，工程师，研究方向为信息技术。

陈洪辉，男(1969—)，教授，博导，研究方向为指挥信息系统、需求工程和体系结构。

(本文编辑：李素华)